36-467 Homework 3

Eu Jing Chua

September 17, 2018

Question 1

Q1 a)

The earliest date is -9600CE, while the most recent is 1900CE. The median date is 500CE.

Q1 b)

	PolPop	PolTerr	CapPop	levels	government	infrastr	writing	texts	money
Min.	1.417	-0.216	1.439	0.000	0.000	0.000	0.000	0.000	0.000
1st Qu.	4.158	3.650	3.477	1.762	0.241	0.342	0.259	0.100	1.800
Median	5.975	5.177	4.339	2.977	0.618	0.750	0.817	0.925	4.000
Mean	5.515	4.779	4.229	2.923	0.552	0.635	0.649	0.634	3.419
3rd Qu.	6.756	5.972	5.095	3.993	0.856	0.900	0.857	0.975	5.000
Max.	8.527	7.402	6.331	6.554	1.000	1.000	1.000	1.000	6.000

Q1 c)

These numbers are not of the actual population because it is highly improbable that the total populations and capital sizes were of sizes less than 10 each. Also, the mean and median of both variables are reasonably close to think that the distribution could be symmetric without much skew. However, this is unrealistic and population distributions tend to be skewed right.

Q1 d)

The transformation might be a log (base 10) of the original population sizes. After reversing the log transform, we get more sensible summaries:

	PolPop	CapPop
Min.	26.095	27.469
1st Qu.	14399.989	3000.000
Median	943248.639	21812.893
Mean	8759383.649	131894.979
3rd Qu.	5698220.601	124496.335
Max.	336419265.273	2142687.978

Q1 e)

Table 3: Covariances between the complexity measures

	PolPop	PolTerr	CapPop	levels	government	infrastr	writing	texts	money
PolPop	2.528	2.098	1.546	1.914	0.395	0.381	0.400	0.539	1.983
PolTerr	2.098	2.436	1.332	1.678	0.343	0.312	0.355	0.481	1.695
CapPop	1.546	1.332	1.236	1.332	0.265	0.265	0.261	0.360	1.254
levels	1.914	1.678	1.332	2.099	0.365	0.342	0.348	0.487	1.742
government	0.395	0.343	0.265	0.365	0.106	0.083	0.080	0.111	0.368
infrastr	0.381	0.312	0.265	0.342	0.083	0.098	0.077	0.110	0.368

	PolPop	PolTerr	CapPop	levels	government	infrastr	writing	texts	money
writing	0.400	0.355	0.261	0.348	0.080	0.077	0.117	0.132	0.426
texts	0.539	0.481	0.360	0.487	0.111	0.110	0.132	0.186	0.551
money	1.983	1.695	1.254	1.742	0.368	0.368	0.426	0.551	3.184

Q1 f) The correlations between the complexity measures are:

Table 4: Correlations between the complexity measures

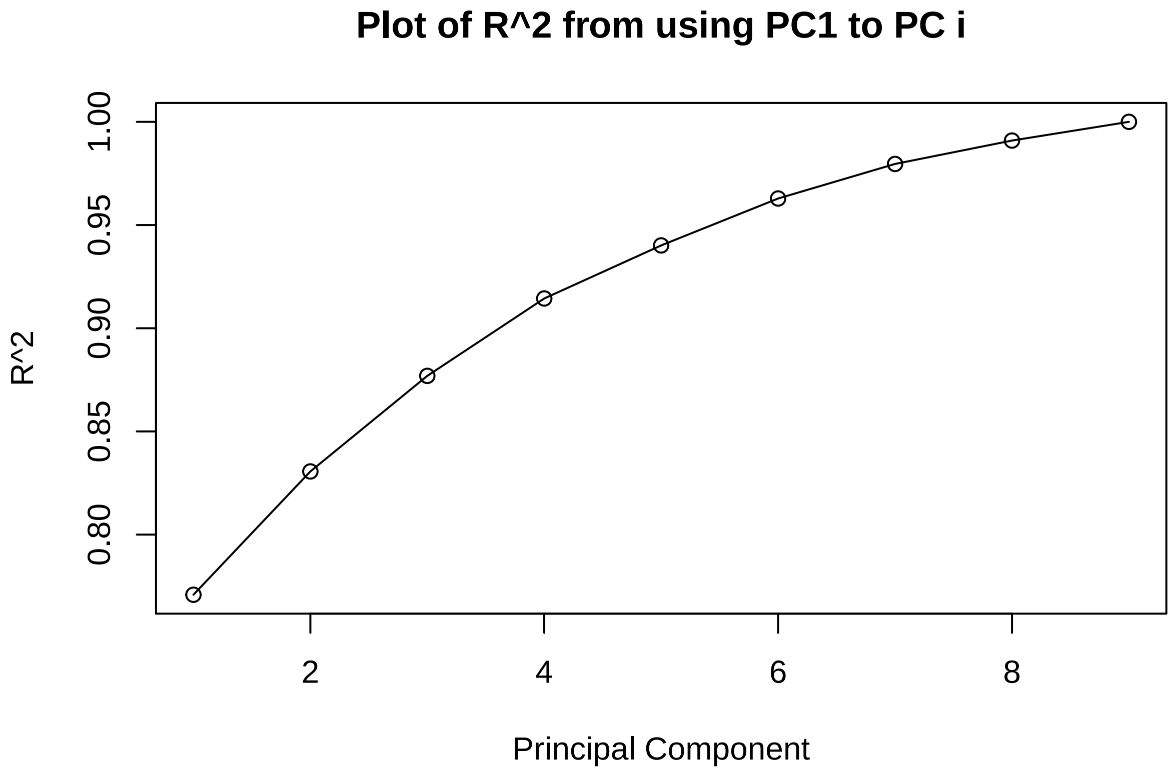
	PolPop	PolTerr	CapPop	levels	government	infrastr	writing	texts	money
PolPop	1.000	0.845	0.875	0.831	0.762	0.766	0.735	0.786	0.699
PolTerr	0.845	1.000	0.768	0.742	0.675	0.640	0.666	0.715	0.608
CapPop	0.875	0.768	1.000	0.827	0.734	0.764	0.686	0.751	0.632
levels	0.831	0.742	0.827	1.000	0.774	0.755	0.703	0.780	0.674
government	0.762	0.675	0.734	0.774	1.000	0.815	0.719	0.793	0.634
infrastr	0.766	0.640	0.764	0.755	0.815	1.000	0.721	0.820	0.661
writing	0.735	0.666	0.686	0.703	0.719	0.721	1.000	0.895	0.698
texts	0.786	0.715	0.751	0.780	0.793	0.820	0.895	1.000	0.716
money	0.699	0.608	0.632	0.674	0.634	0.661	0.698	0.716	1.000

Question 2

Q2 a)

It makes sense to scale the variables going in to PCA to all have variance 1 as they are all measures of different things and quantities with possibly different units.

Q2 b)



By using all 9 principal components, we do not lose any information and are simply projecting the existing data onto new orthogonal coordinates, so this projection should still fully capture all the variance of the original data.
To capture 75% of the variance, just using the 1^{st} PC is enough. For 90%, we would need to use at least 4 PC's.

Q2 c)

Table 5: First 3 Principal Component Vectors

	PC1	PC2	PC3
PolPop	0.351	-0.319	0.128
PolTerr	0.320	-0.476	0.319
CapPop	0.339	-0.377	-0.065
levels	0.341	-0.209	-0.072
government	0.332	0.097	-0.472
infrastr	0.334	0.174	-0.452
writing	0.328	0.437	0.107
texts	0.349	0.323	-0.072
money	0.302	0.388	0.655

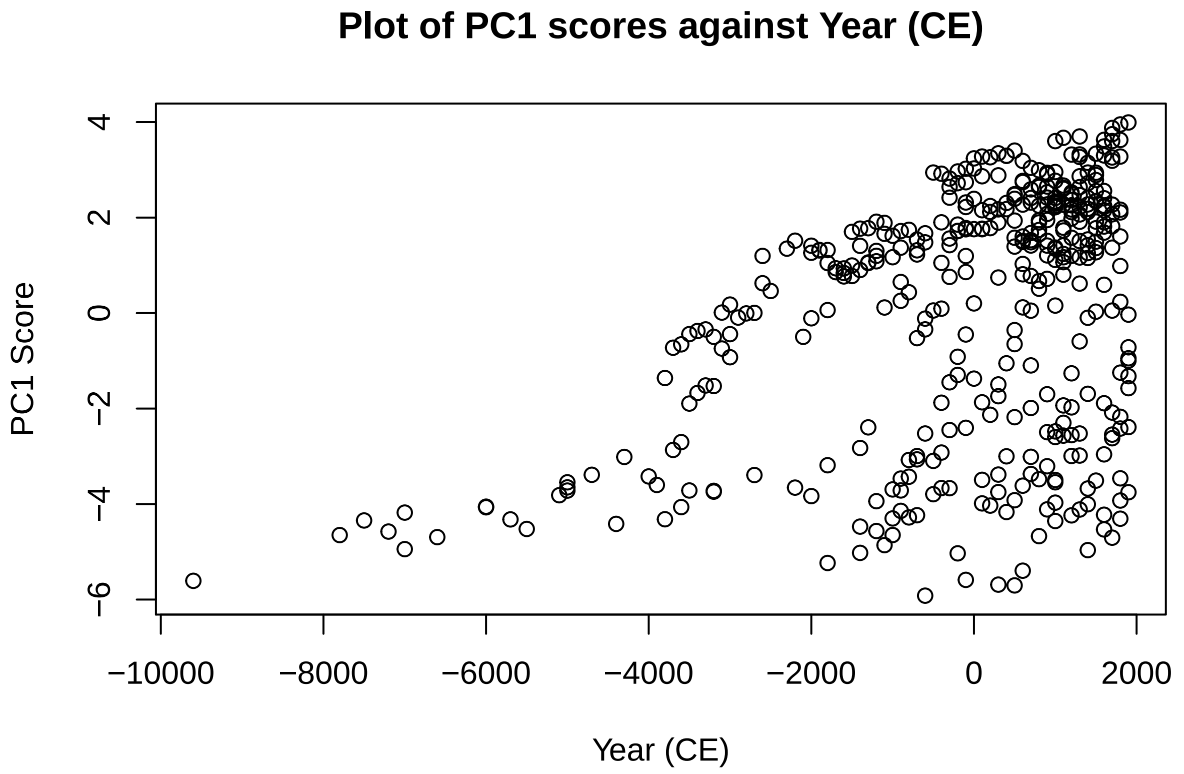
Q2 d)
As each variable is weighed roughly the same in PC1, a polity that has high complexity measures across the board will get a high score on PC1, while one with low scores across the board will get a low score on PC1.

Q2 e)

Polities that have small populations but have high measures of writing, texts, and money score higher scores on PC2, while those with larger populations but low measures of writing, texts and money score lower.

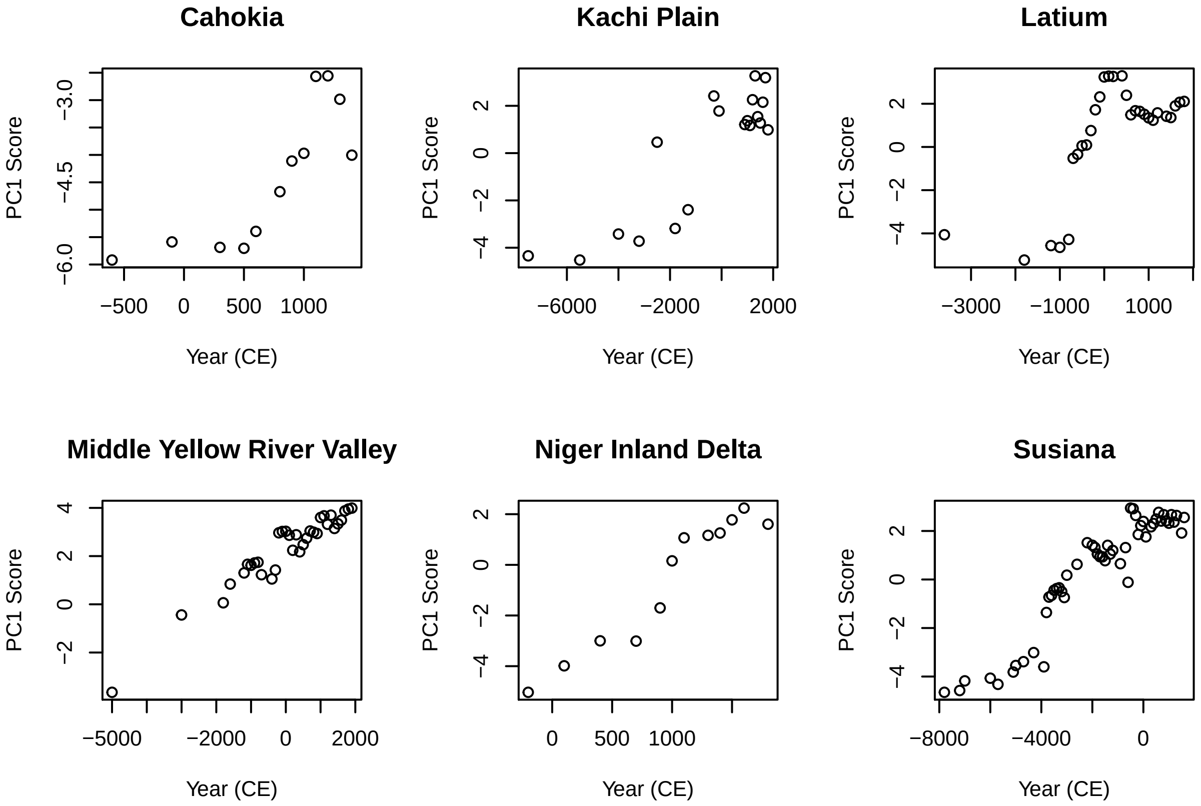
Question 3

Q3 a)



As time progressed, the PC1 score of polities tended to increase, but so did the spread of the PC1 score.

Q3 b)



- **Cahokia**: The few data points seem to indicate a rough non-linear positively increasing of PC1 score, up to the late 1000's where it might have dropped.

- **Kachi Plain**: There seems to be a rough positively increasing trend of PC1 score that seems linear, with much more data collected nearer to the 2000's.

- **Latium**: There seemed to be a rough positively increasing trend of PC1 score that seemed linear, up till around 500 CE, where the PC1 scores seemed to drop but steadily rise again.

- **Middle Yellow River Valley**: There seems to be a rough positively increasing trend of PC1 score that seems linear.

- **Niger Inland Delta**: There seems to be a rough positively increasing trend of PC1 score that seems linear.
- **Susuana**: There seems to be a rough positively increasing trend of PC1 score that seems linear.

Q3 c)
The common pattern seems to be that the PC1 score tends to increase over time, in general, for all 6 regions.

Q3 d)
This increasing of PC1 score over time for the polities indicates that all 9 of their complexity measures tended to increase over time.

Question 4

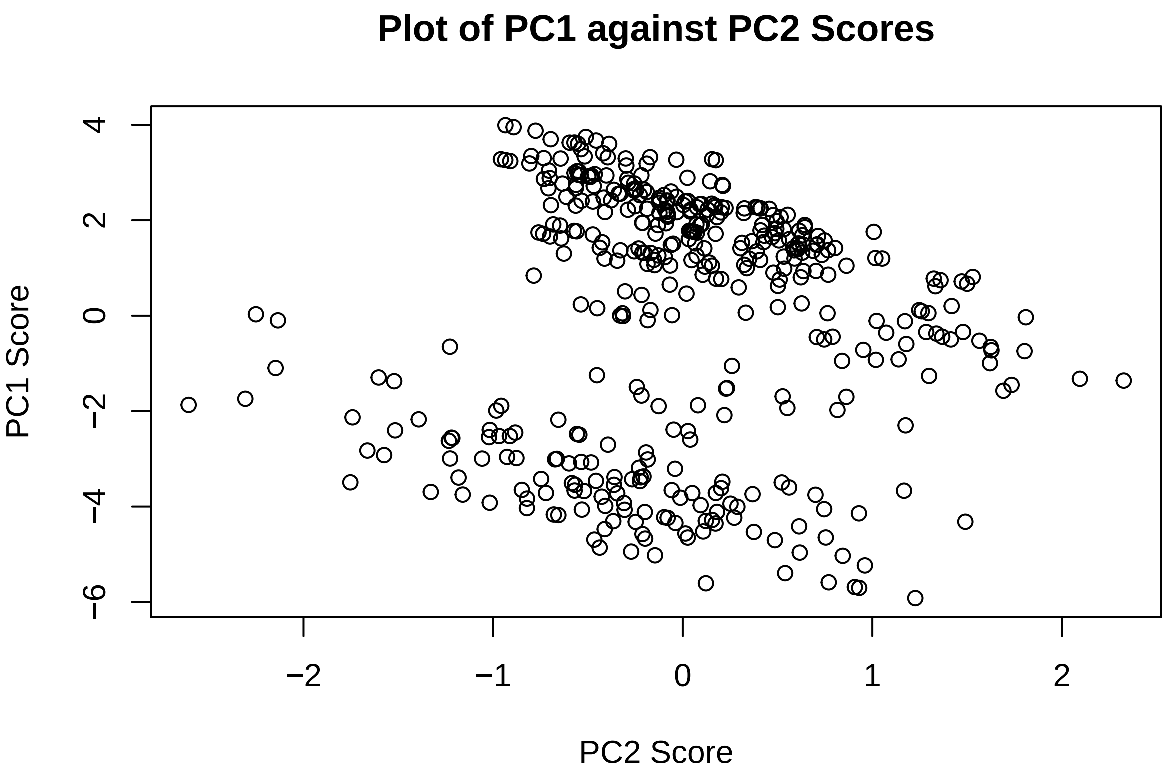
Q4 a)
The correlation between scores on PC1 and PC2 is 1.54×10^{-14}.

Q4 b)
The theoretical correlation should be 0. The difference between the theorical and calculated value is very small, and hence is not a cause for concern.

Q4 c)

```
data$PC2.score <- data.pca$x[, 2]

plot(
  PC1.score ~ PC2.score, data = data,
  xlab = "PC2 Score", ylab = "PC1 Score",
  main = "Plot of PC1 against PC2 Scores"
)
```



There seems to be two clusters in the plot, one with high PC1 and PC2 scores, and the other with low PC1 and PC2 scores.

Q4 d)
The existence of these two clusters from the plot indicates that the two PCs are not statistically independent of each other, as if PC1 is high, it is more likely for PC2 to be high too etc. Given that the correlations were very close to 0, we still cannot assume that the two PCs are independent soley from this fact; we would have to make an assumption that the variables were jointly distributed with a multivariate normal distribution. In this case, it appears such an assumption would be invalid.

Question 5

Q5 a)

Table 6: First 3 Principal Component Vectors

	PC1	PC2	PC3
PolPop	0.350	-0.361	0.113
PolTerr	0.322	-0.433	0.374
CapPop	0.338	-0.379	-0.151
levels	0.341	-0.237	-0.053
government	0.333	0.159	-0.464
infrastr	0.332	0.162	-0.490
writing	0.329	0.430	0.176
texts	0.350	0.323	-0.019
money	0.303	0.376	0.578

My interpretations would not differ much from those in Q2 d) and e).

Q5 b)

Table 7: Mean of First 3 PCs

	PC1	PC2	PC3
PolPop	0.352	-0.337	0.124
PolTerr	0.322	-0.473	0.316
CapPop	0.337	-0.373	-0.091
levels	0.341	-0.203	-0.057
government	0.333	0.128	-0.466
infrastr	0.333	0.168	-0.469
writing	0.328	0.422	0.128
texts	0.349	0.322	-0.052
money	0.302	0.387	0.637

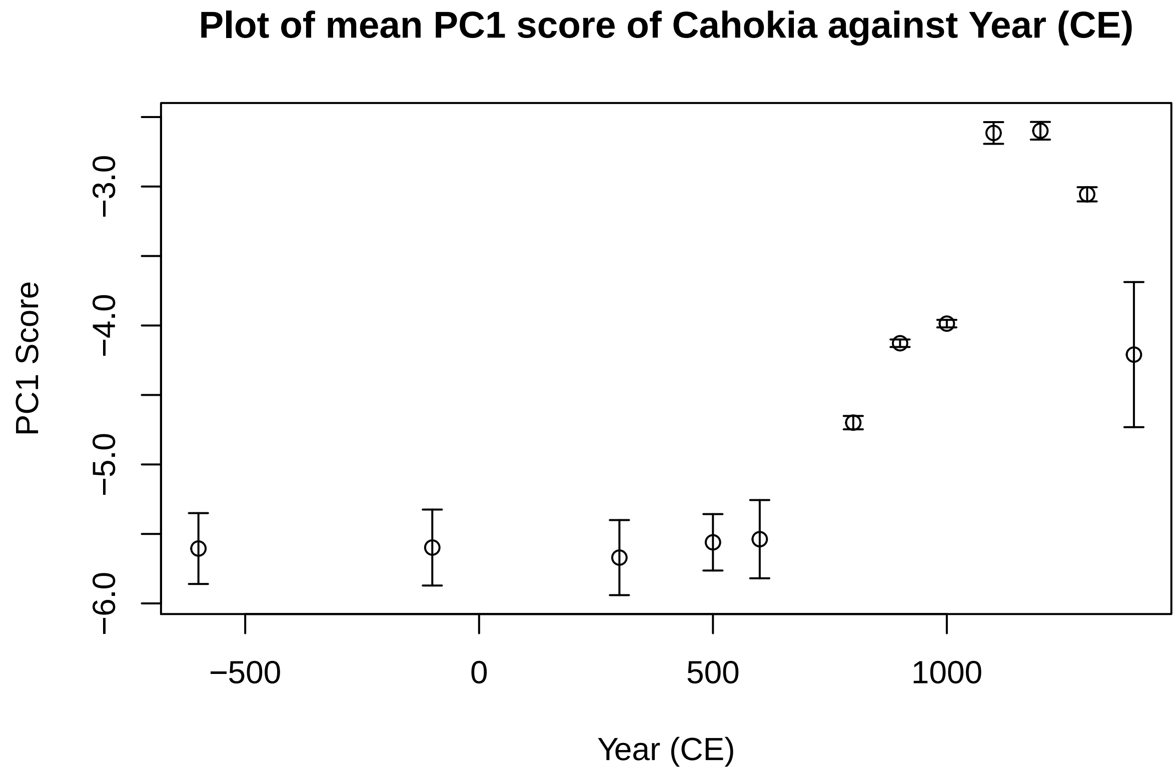
Table 8: Standard Errors of First 3 PCs

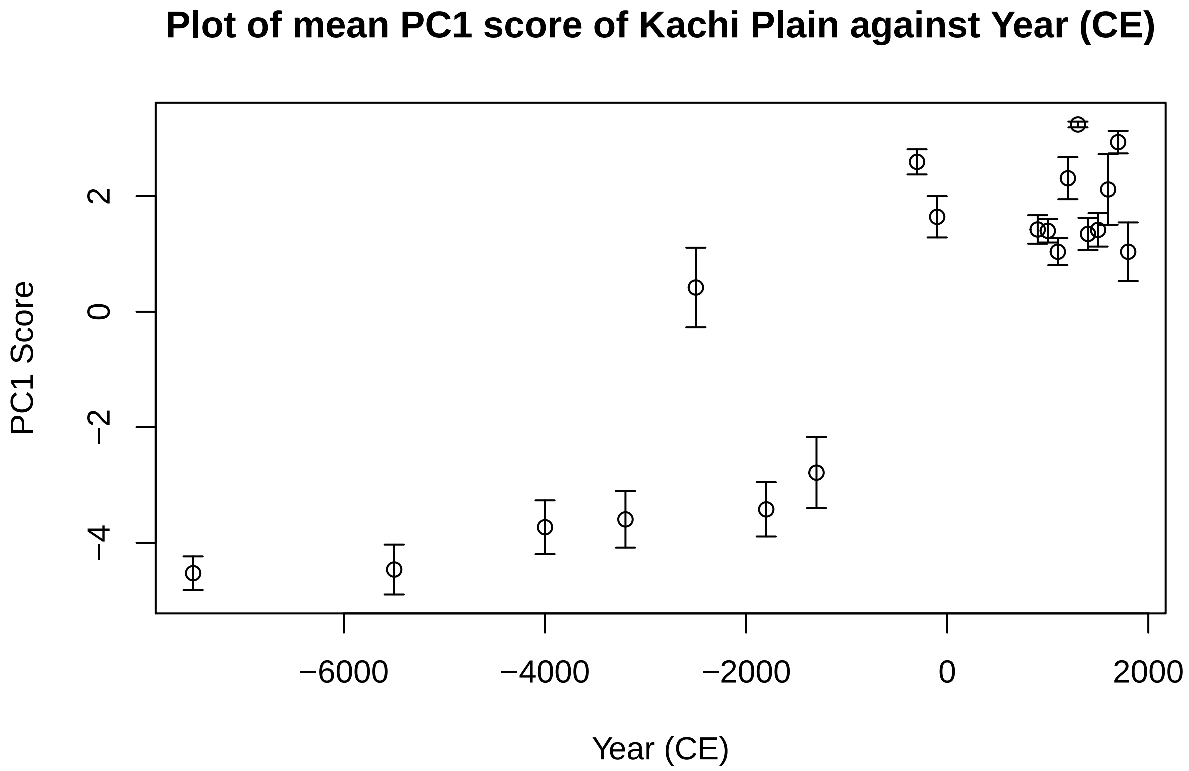
	PC1	PC2	PC3
PolPop	0.00034	0.00377	0.00841
PolTerr	0.00054	0.00787	0.01121
CapPop	0.00046	0.00535	0.00793
levels	0.00016	0.00427	0.00699
government	0.00024	0.00625	0.00379
infrastr	0.00025	0.00867	0.00474
writing	0.00019	0.00466	0.01083
texts	0.00013	0.00282	0.00828
money	0.00027	0.00912	0.00949

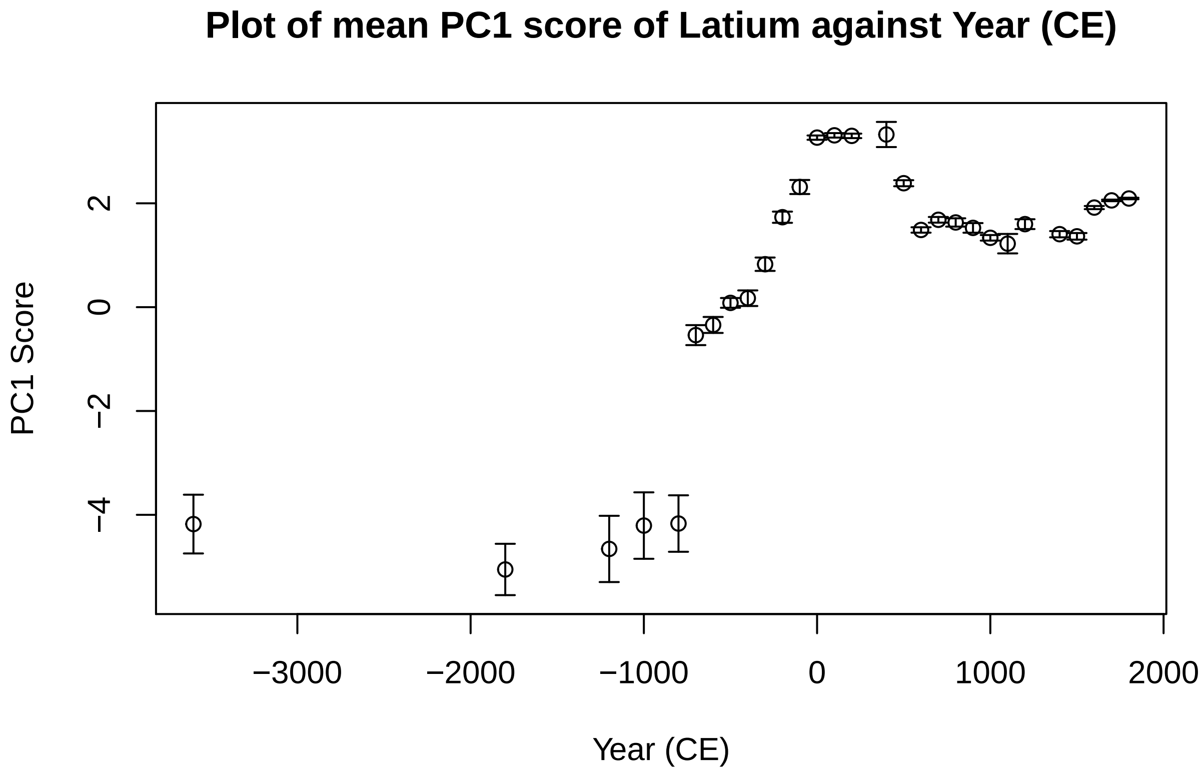
Q5 c)

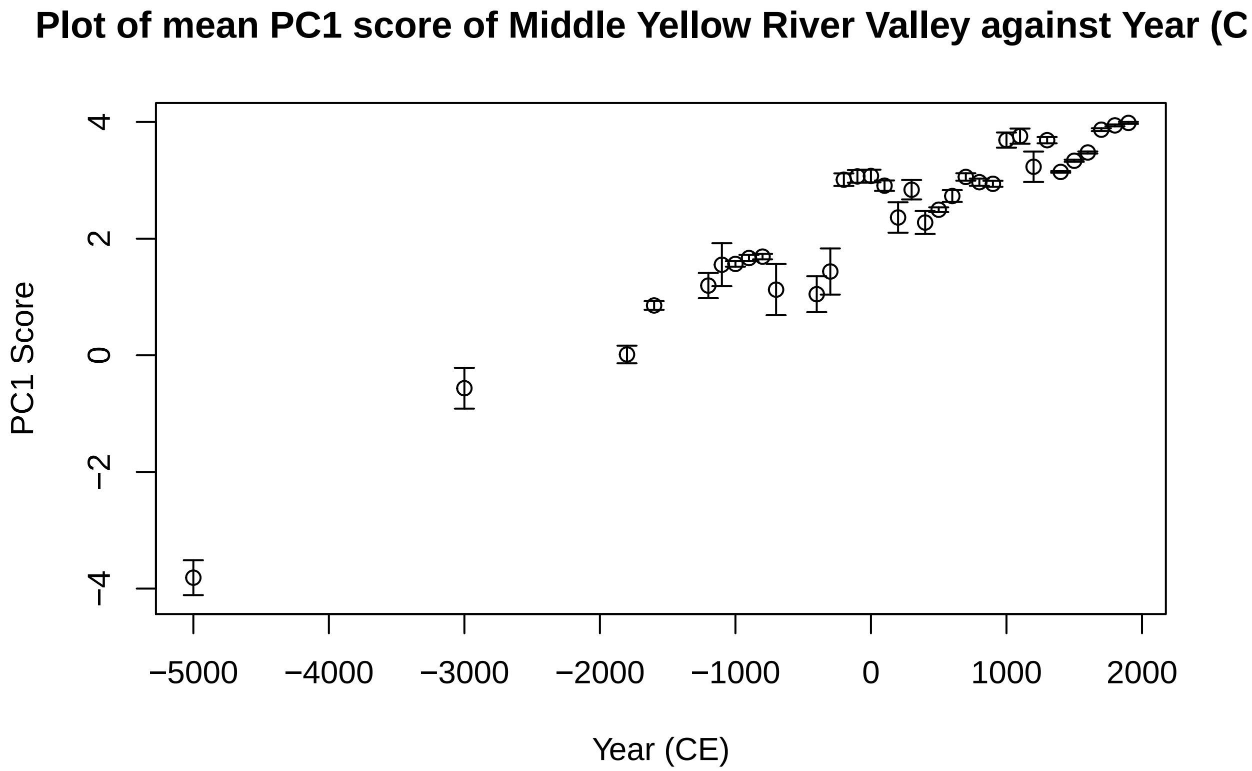
My interpretations would not differ much from those in Q2 d) and e).

Q5 d)

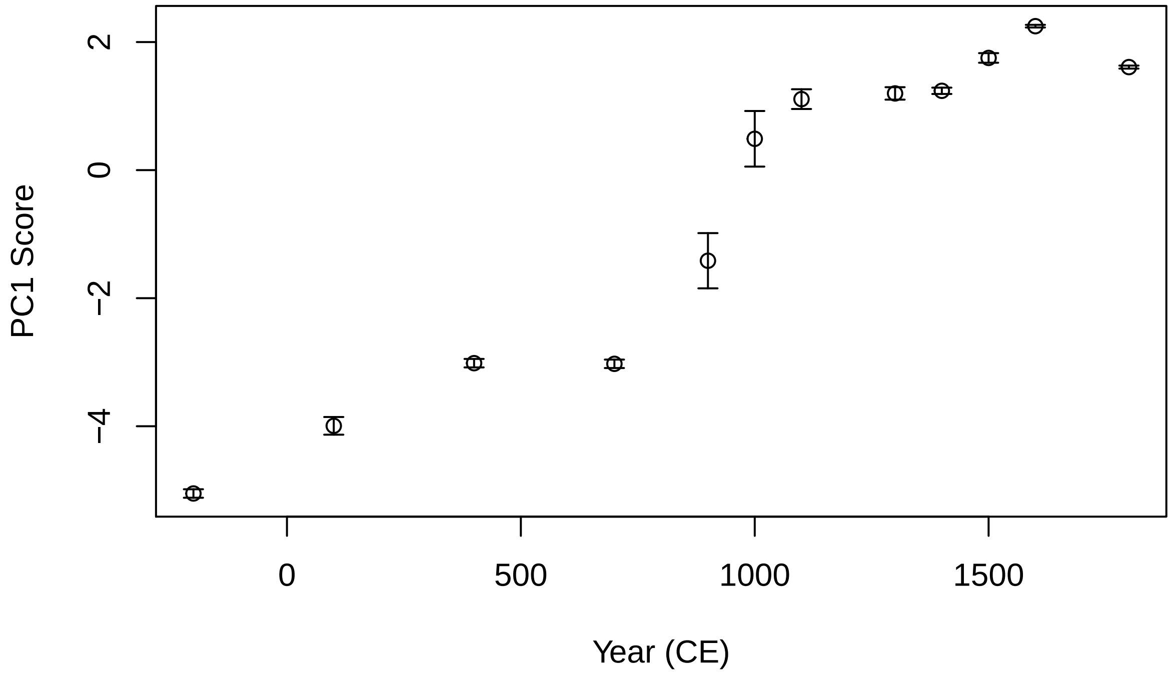

Plot of mean PC1 score of Cahokia against Year (CE)
PC1 Score
−3.0
−4.0
−5.0
−6.0
−500
0
500
1000
Year (CE)

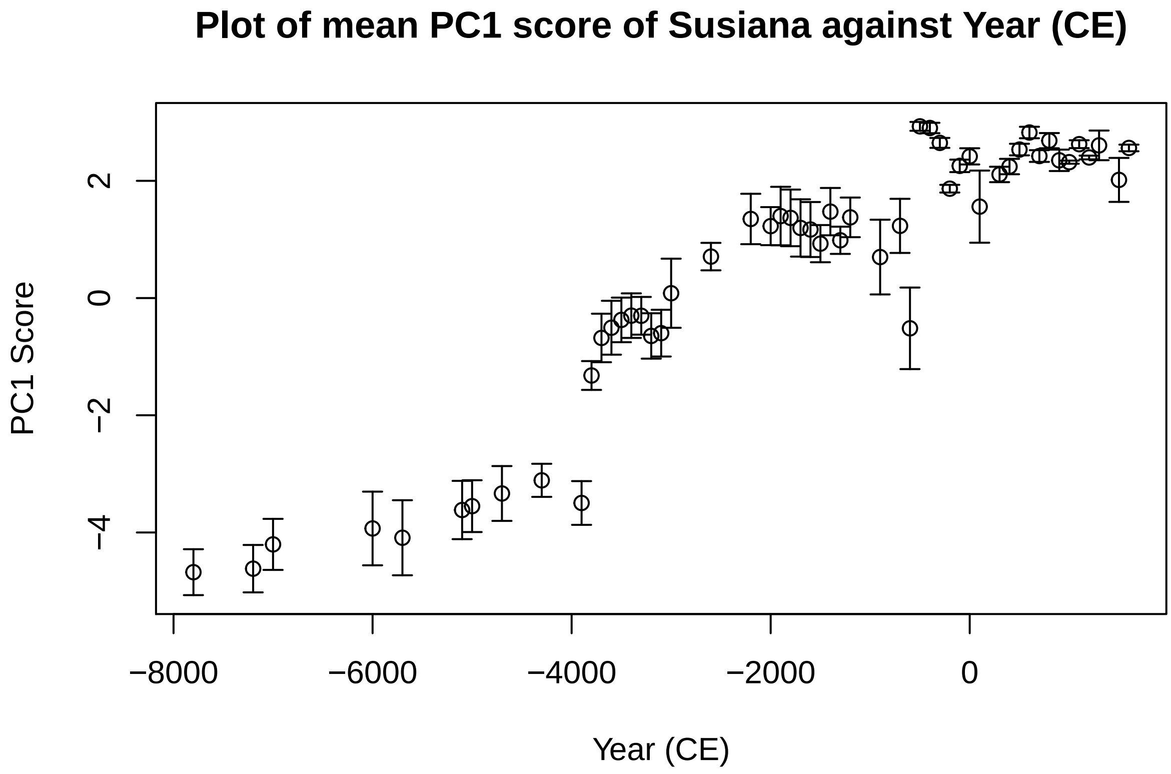

Plot of mean PC1 score of Kachi Plain against Year (CE)
PC1 Score
2
0
−2
−4
−6000
−4000
−2000
0
2000
Year (CE)


Plot of mean PC1 score of Latium against Year (CE)
PC1 Score
2
0
−2
−4
−3000
−2000
−1000
0
1000
2000
Year (CE)


Plot of mean PC1 score of Middle Yellow River Valley against Year (C
PC1 Score
4
2
0
−2
−4
−5000
−4000
−3000
−2000
−1000
0
1000
2000
Year (CE)

Plot of mean PC1 score of Niger Inland Delta against Year (CE)




Plot of mean PC1 score of Susiana against Year (CE)
PC1 Score
2
0
−2
−4
−8000
−6000
−4000
−2000
0
Year (CE)