Data Engineering in Soil Investigation

Jinling Li

Agenda

- Introduction
- Data Preparation
- Exploratory Data Analysis
- Learning Methods
- Results and Discussion
- Conclusion

Introduction

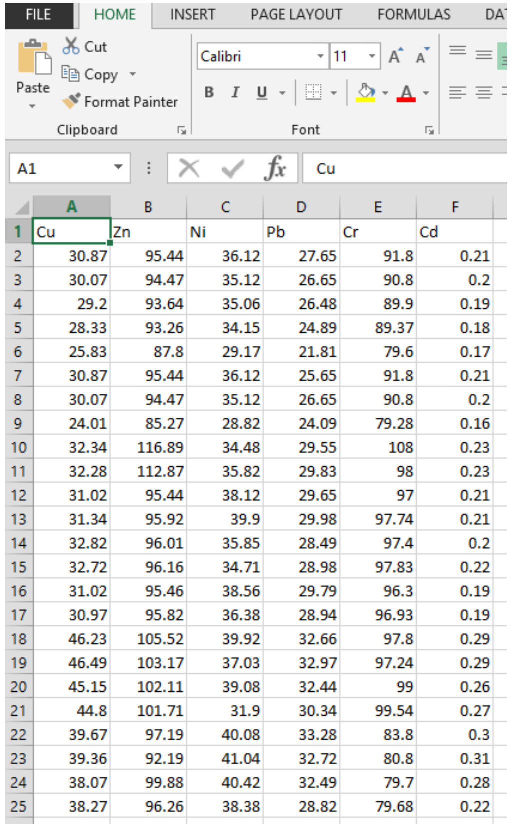
- Soil investigation is of great significance to agriculture and environment. Finding out the patterns between soil properties and attributes would help improve the understanding and controlling of soils.
- The objective of this study was: 1) to investigate soil C and N trends over depths, and 2) to analyze heavy metal trend (Cu, Zn, Ni, Pb, Cr and Cd).

Data Preparation

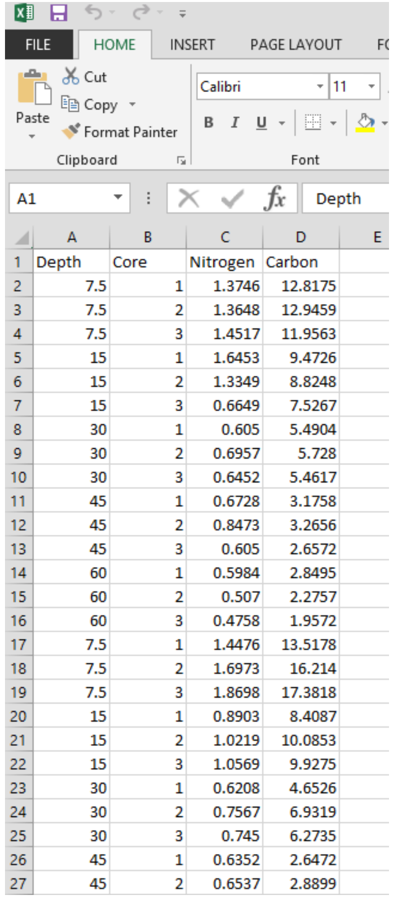
- In 2011, thousands of soil samples were collected randomly in Virginia, US. Each sample was measured for specific properties. Finally, a size of 2000+ data points were generated.
- Soil Carbon and Nitrogen determination

  Three soil cores were collected at each spot to a depth of 60 cm. Each sample core was then sectioned into 7.5, 15, 30, 45, and 60 cm increments. Soil carbon and nitrogen concentration was measured.
- Heavy metal determination

  A total of 24 subsamples in 0-15 cm depth were combined and selected for heavy metal investigation. The contents of total heavy metals were determined.



	A	B	C	D	E	F
1	Cu	Zn	Ni	Pb	Cr	Cd
2	30.87	95.44	36.12	27.65	91.8	0.21
3	30.07	94.47	35.12	26.65	90.8	0.2
4	29.2	93.64	35.06	26.48	89.9	0.19
5	28.33	93.26	34.15	24.89	89.37	0.18
6	25.83	87.8	29.17	21.81	79.6	0.17
7	30.87	95.44	36.12	25.65	91.8	0.21
8	30.07	94.47	35.12	26.65	90.8	0.2
9	24.01	85.27	28.82	24.09	79.28	0.16
10	32.34	116.89	34.48	29.55	108	0.23
11	32.28	112.87	35.82	29.83	98	0.23
12	31.02	95.44	38.12	29.65	97	0.21
13	31.34	95.92	39.9	29.98	97.74	0.21
14	32.82	96.01	35.85	28.49	97.4	0.2
15	32.72	96.16	34.71	28.98	97.83	0.22
16	31.02	95.46	38.56	29.79	96.3	0.19
17	30.97	95.82	36.38	28.94	96.93	0.19
18	46.23	105.52	39.92	32.66	97.8	0.29
19	46.49	103.17	37.03	32.97	97.24	0.29
20	45.15	102.11	39.08	32.44	99	0.26
21	44.8	101.71	31.9	30.34	99.54	0.27
22	39.67	97.19	40.08	33.28	83.8	0.3
23	39.36	92.19	41.04	32.72	80.8	0.31
24	38.07	99.88	40.42	32.49	79.7	0.28
25	38.27	96.26	38.38	28.82	79.68	0.22



	A	B	C	D	E
1	Depth	Core	Nitrogen	Carbon	
2	7.5	1	1.3746	12.8175	
3	7.5	2	1.3648	12.9459	
4	7.5	3	1.4517	11.9563	
5	15	1	1.6453	9.4726	
6	15	2	1.3349	8.8248	
7	15	3	0.6649	7.5267	
8	30	1	0.605	5.4904	
9	30	2	0.6957	5.728	
10	30	3	0.6452	5.4617	
11	45	1	0.6728	3.1758	
12	45	2	0.8473	3.2656	
13	45	3	0.605	2.6572	
14	60	1	0.5984	2.8495	
15	60	2	0.507	2.2757	
16	60	3	0.4758	1.9572	
17	7.5	1	1.4476	13.5178	
18	7.5	2	1.6973	16.214	
19	7.5	3	1.8698	17.3818	
20	15	1	0.8903	8.4087	
21	15	2	1.0219	10.0853	
22	15	3	1.0569	9.9275	
23	30	1	0.6208	4.6526	
24	30	2	0.7567	6.9319	
25	30	3	0.745	6.2735	
26	45	1	0.6352	2.6472	
27	45	2	0.6537	2.8899	

Exploratory Data Analysis

Soil Carbon and Nitrogen Content

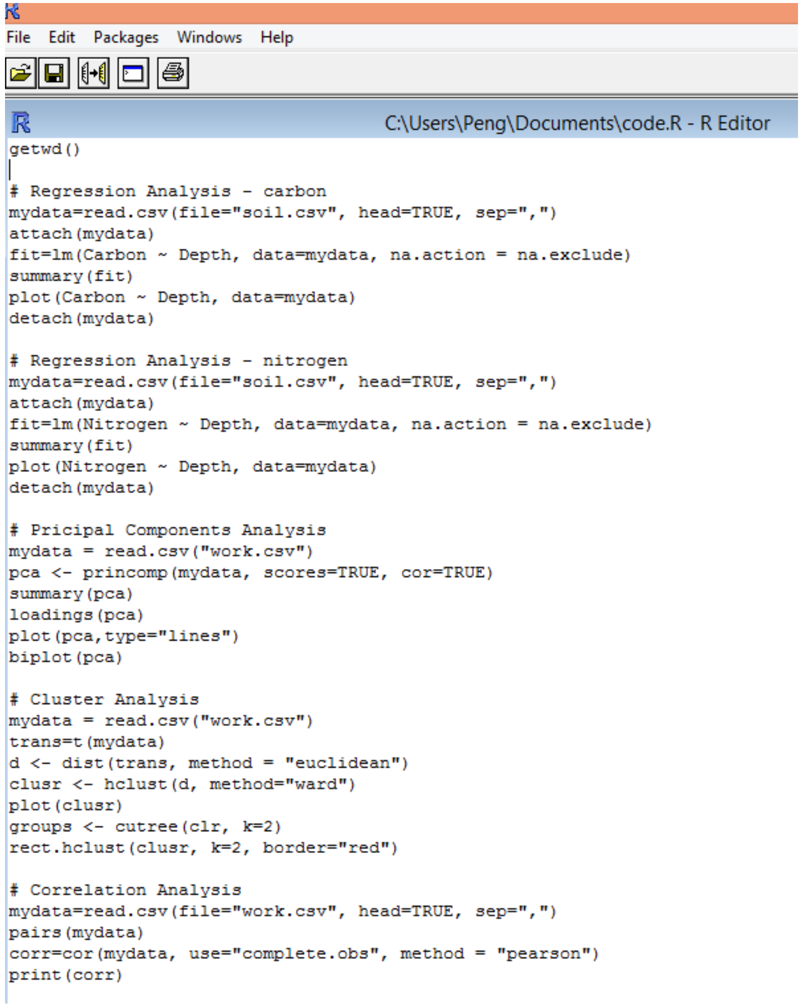
	Nitrogen	*Carbon*
	(g kg^{-1})	
Mean	1.06	8.14
Standard Error	0.04	0.28
Median	0.84	6.21
Standard Deviation	0.72	5.77
Sample Variance	0.52	33.25
Kurtosis	27.50	-0.25
Skewness	4.30	0.87
Range	6.63	25.58
Minimum	0.36	1.37
Maximum	6.99	26.95
Sum	443.55	3419.10
Count	420	420.00
Confidence Level(95%)	0.07	0.55

Heavy Metal Content

	Minimum	Maximum	Mean	Standard Deviation
	--------------------mg kg^{-1}------------------			
Cu	24.01	46.49	34.24	6.45
Zn	85.27	116.89	97.60	6.93
Ni	28.82	41.04	36.31	3.25
Pb	21.81	33.28	28.95	3.05
Cr	79.28	108.00	92.09	7.94
Cd	0.16	0.31	0.23	0.04

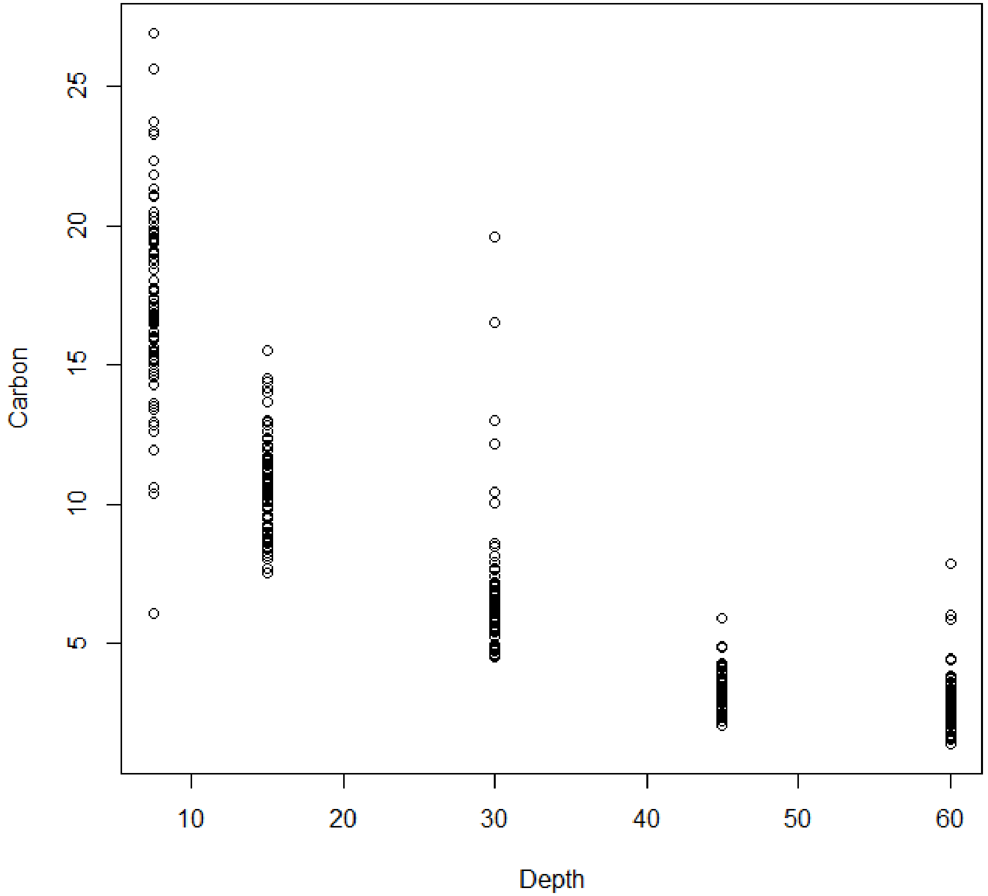
Learning Method

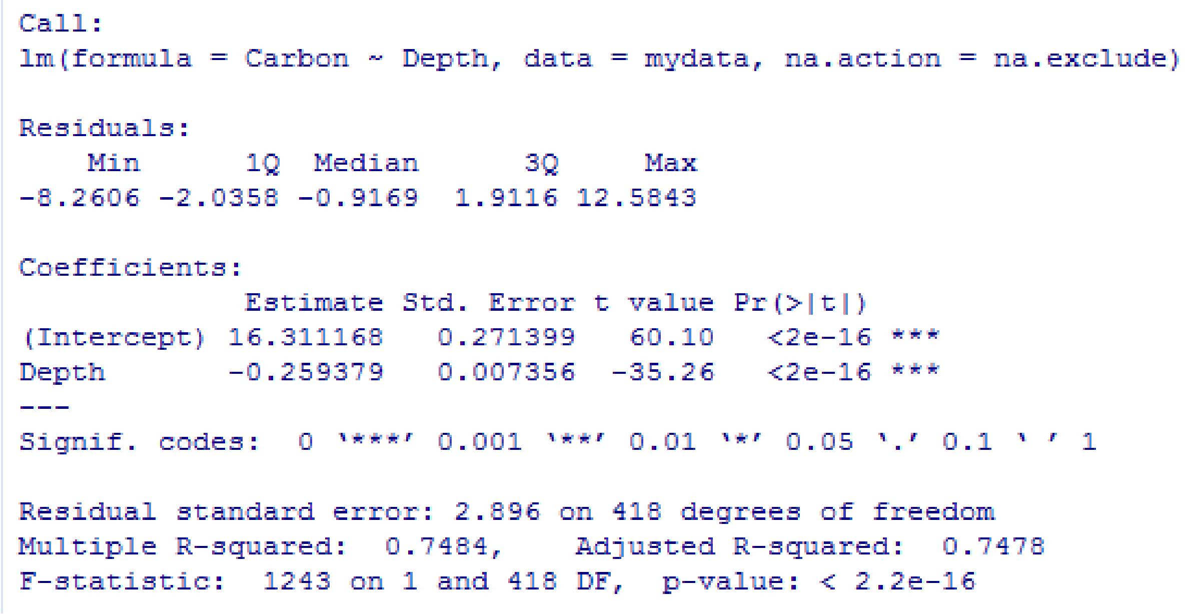
- The data obtained from analytical methods were treated statistically using R and Excel software.
- **Regression analysis** was conducted to detect the relationship between a dependent variable and independent variables.
- **Principal component analysis** was employed to infer the hypothetical components of heavy metals.
- **Cluster analysis** was applied to identify different geochemical groups within heavy metal content. The matrix was formulated according to the Ward algorithmic method.
- **Correlation analysis** was used to identify the relationship between the six elements. Pearson’s product moment correlation coefficient was calculated in the forms of matrix.


File Edit Packages Windows Help
C:\Users\Peng\Documents\code.R - R Editor

```
getwd()

# Regression Analysis - carbon
mydata=read.csv(file="soil.csv", head=TRUE, sep=",")
attach(mydata)
fit=lm(Carbon ~ Depth, data=mydata, na.action = na.exclude)
summary(fit)
plot(Carbon ~ Depth, data=mydata)
detach(mydata)

# Regression Analysis - nitrogen
mydata=read.csv(file="soil.csv", head=TRUE, sep=",")
attach(mydata)
fit=lm(Nitrogen ~ Depth, data=mydata, na.action = na.exclude)
summary(fit)
plot(Nitrogen ~ Depth, data=mydata)
detach(mydata)

# Pricipal Components Analysis
mydata = read.csv("work.csv")
pca <- princomp(mydata, scores=TRUE, cor=TRUE)
summary(pca)
loadings(pca)
plot(pca,type="lines")
biplot(pca)

# Cluster Analysis
mydata = read.csv("work.csv")
trans=t(mydata)
d <- dist(trans, method = "euclidean")
clusr <- hclust(d, method="ward")
plot(clusr)
groups <- cutree(clr, k=2)
rect.hclust(clusr, k=2, border="red")

# Correlation Analysis
mydata=read.csv(file="work.csv", head=TRUE, sep=",")
pairs(mydata)
corr=cor(mydata, use="complete.obs", method = "pearson")
print(corr)
```

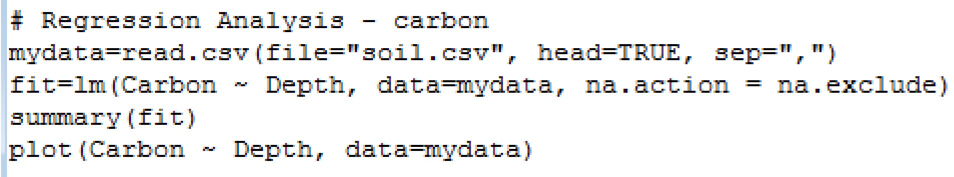
Regression Analysis - Carbon





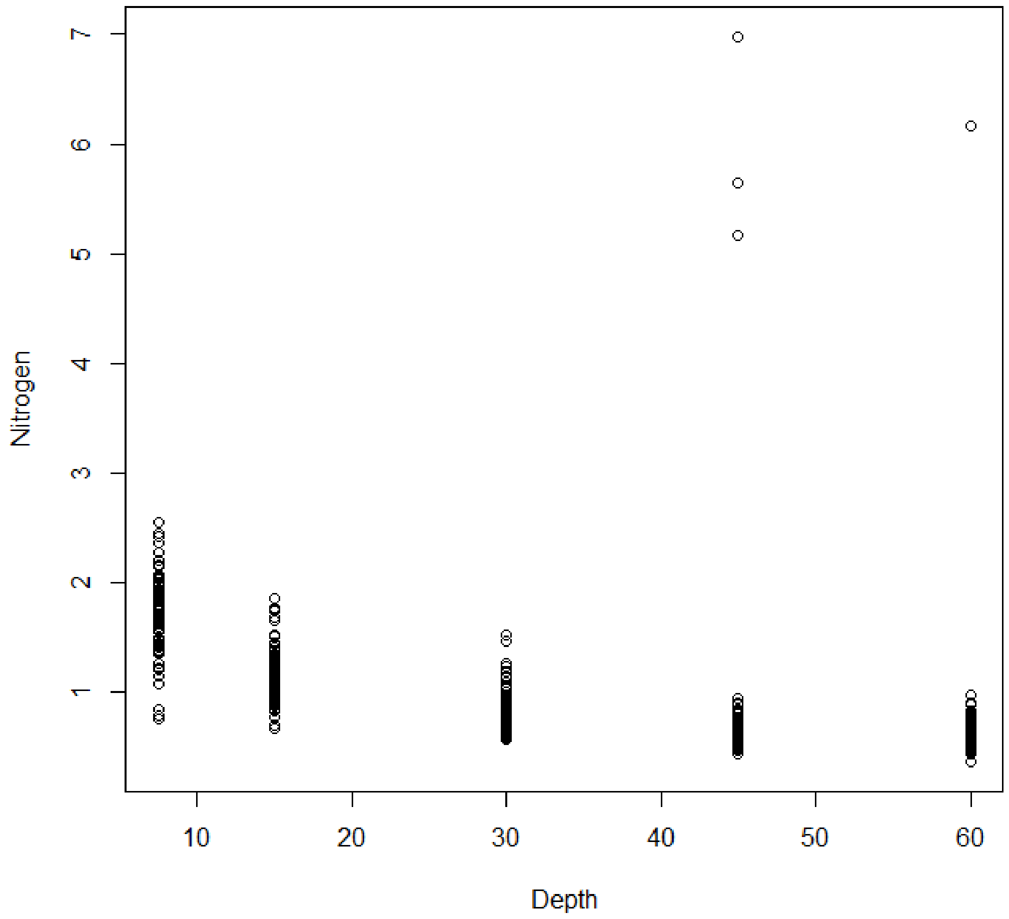
```
Call:
lm(formula = Carbon ~ Depth, data = mydata, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2606 -2.0358 -0.9169  1.9116 12.5843 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.311168   0.271399   60.10   <2e-16 ***
Depth       -0.259379   0.007356  -35.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.896 on 418 degrees of freedom
Multiple R-squared:  0.7484,	Adjusted R-squared:  0.7478 
F-statistic:  1243 on 1 and 418 DF,  p-value: < 2.2e-16
```

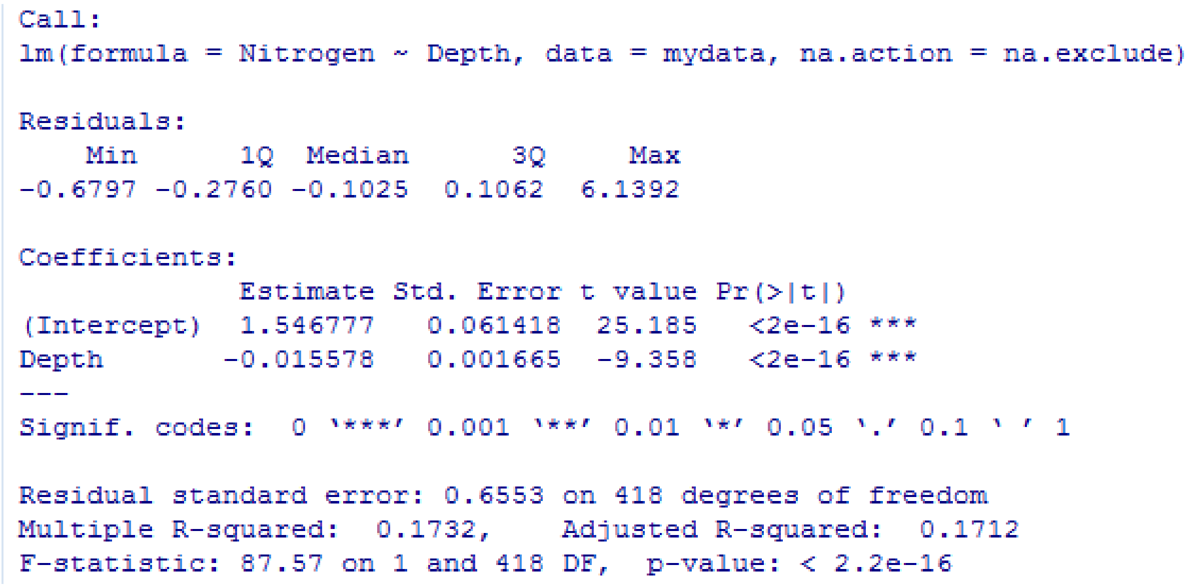
Results showed that a negative relationship between soil carbon content and sampling depth. The deeper in the earth, the lesser soil carbon presence.



```
# Regression Analysis - carbon
mydata=read.csv(file="soil.csv", head=TRUE, sep=",")
fit=lm(Carbon ~ Depth, data=mydata, na.action = na.exclude)
summary(fit)
plot(Carbon ~ Depth, data=mydata)
```

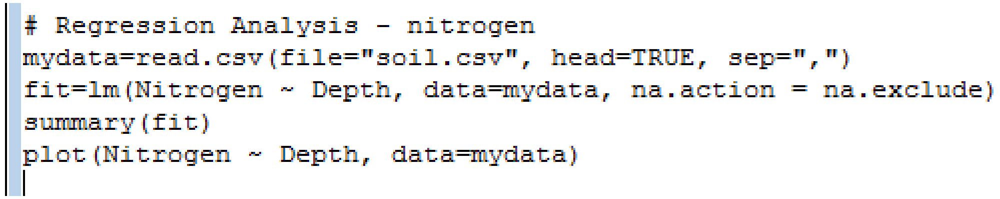
Regression Analysis - Nitrogen





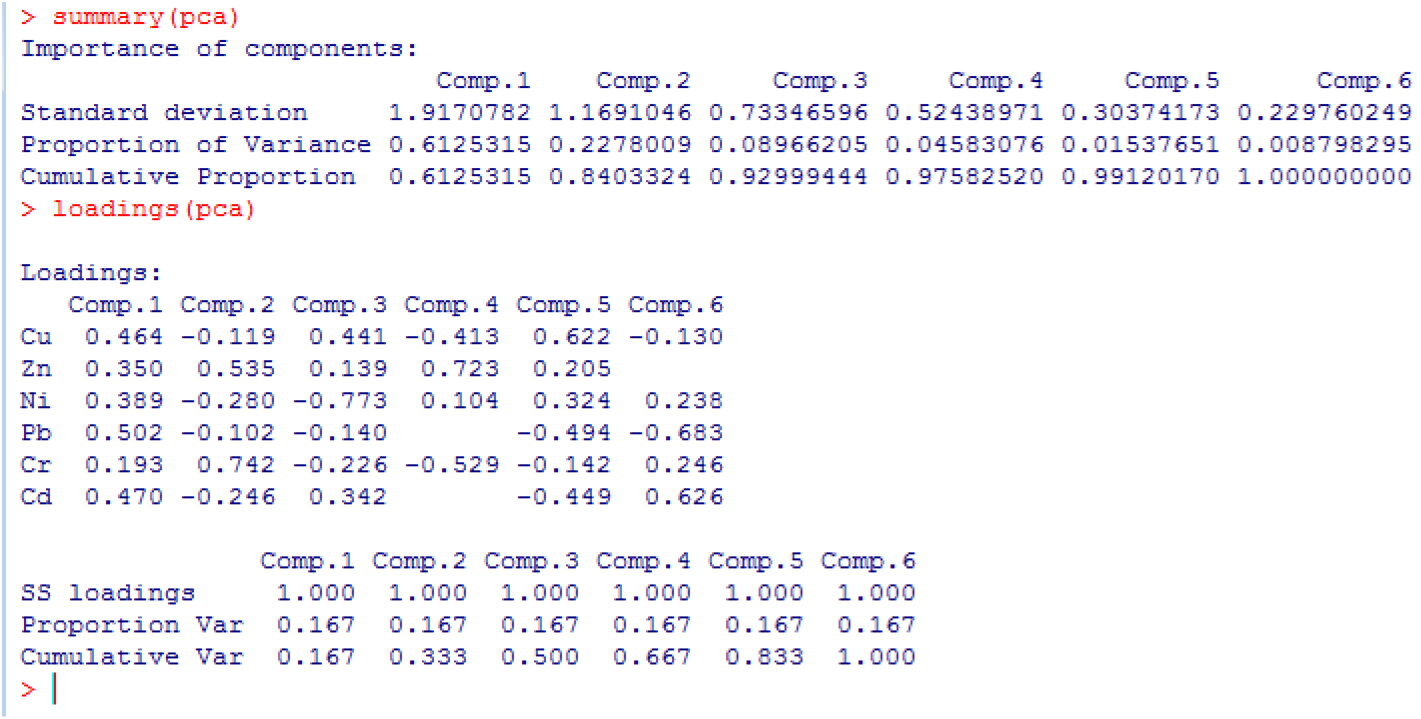
```
Call:
lm(formula = Nitrogen ~ Depth, data = mydata, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6797 -0.2760 -0.1025  0.1062  6.1392 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.546777   0.061418  25.185   <2e-16 ***
Depth       -0.015578   0.001665  -9.358   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6553 on 418 degrees of freedom
Multiple R-squared:  0.1732,	Adjusted R-squared:  0.1712 
F-statistic: 87.57 on 1 and 418 DF,  p-value: < 2.2e-16
```

Results showed that a negative relationship between soil nitrogen content and sampling depth.

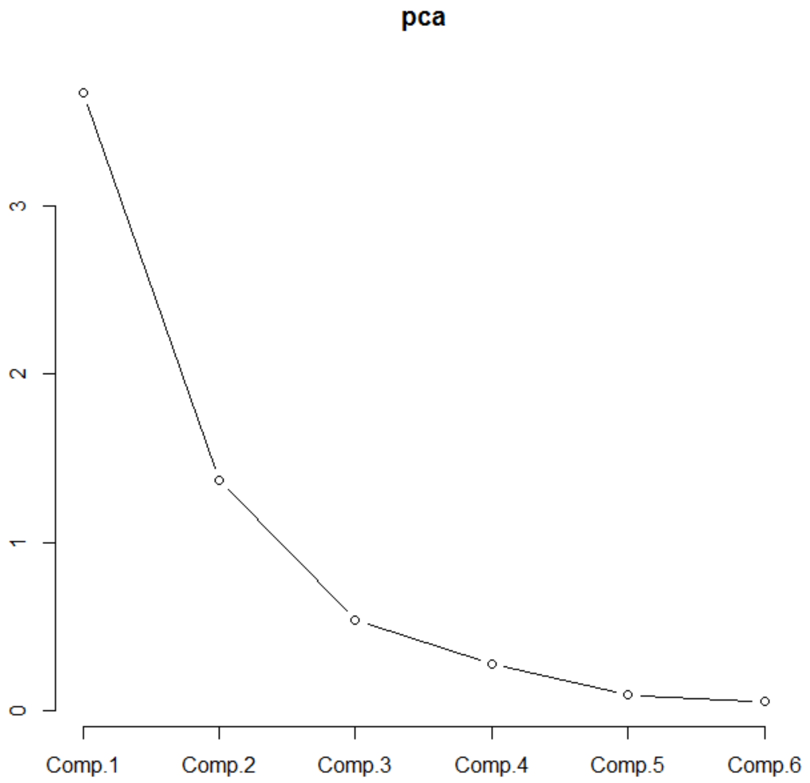


```
# Regression Analysis - nitrogen
mydata=read.csv(file="soil.csv", head=TRUE, sep=",")
fit=lm(Nitrogen ~ Depth, data=mydata, na.action = na.exclude)
summary(fit)
plot(Nitrogen ~ Depth, data=mydata)
```

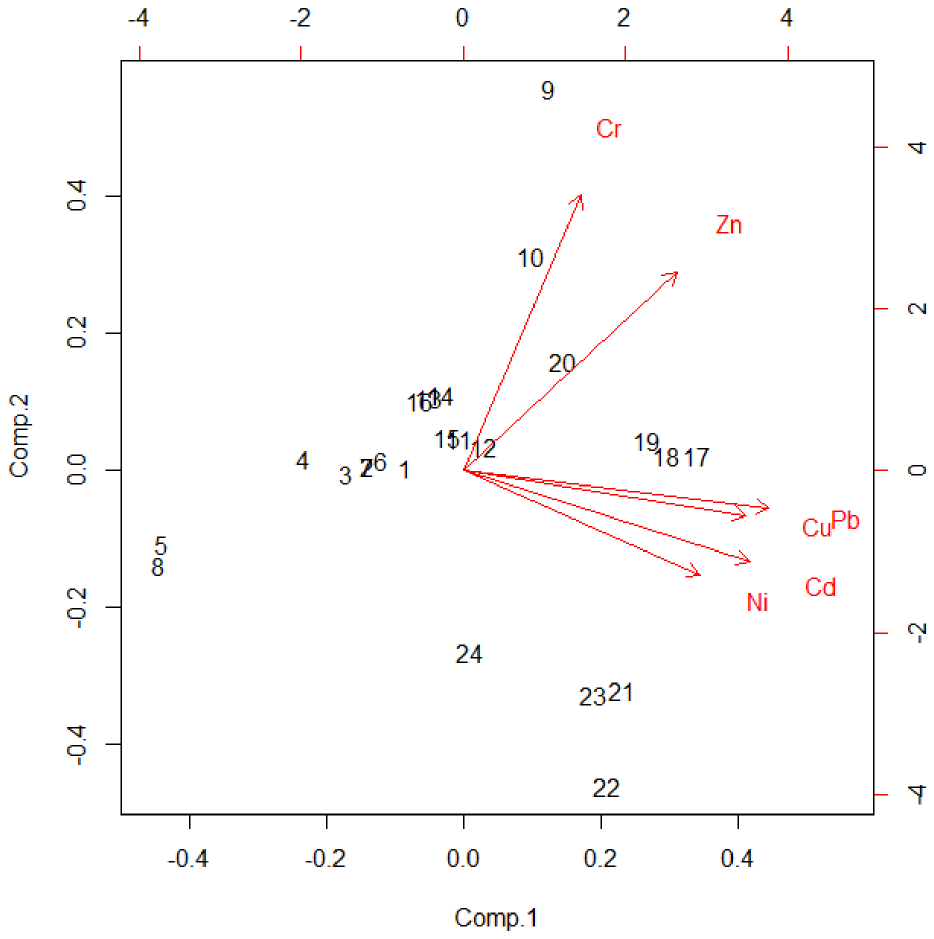
Principal Component Analysis

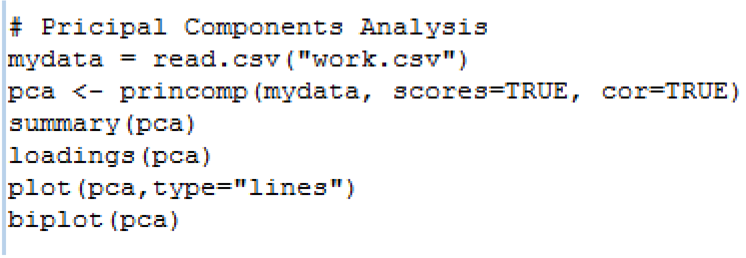


```
> summary(pca)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5      Comp.6
Standard deviation     1.9170782 1.1691046 0.73346596 0.52438971 0.30374173 0.229760249
Proportion of Variance 0.6125315 0.2278009 0.08966205 0.04583076 0.01537651 0.008798295
Cumulative Proportion  0.6125315 0.8403324 0.92999444 0.97582520 0.99120170 1.000000000
> loadings(pca)

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Cu  0.464 -0.119  0.441 -0.413  0.622 -0.130
Zn  0.350  0.535  0.139  0.723  0.205       
Ni  0.389 -0.280 -0.773  0.104  0.324  0.238
Pb  0.502 -0.102 -0.140        -0.494 -0.683
Cr  0.193  0.742 -0.226 -0.529 -0.142  0.246
Cd  0.470 -0.246  0.342        -0.449  0.626

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
> 
```



Principal Component Analysis



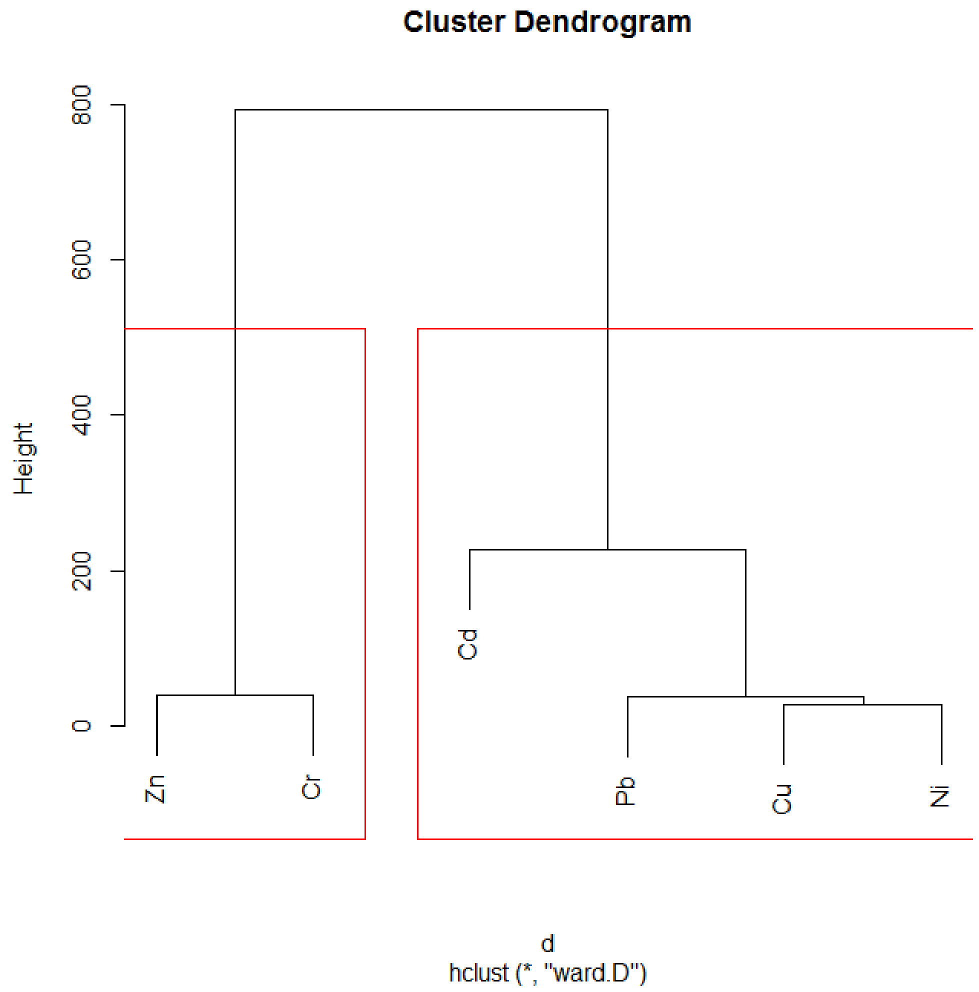


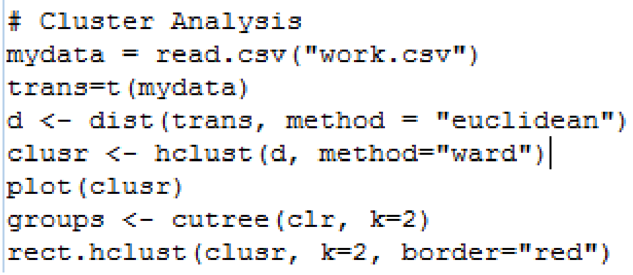
```
# Pricipal Components Analysis
mydata = read.csv("work.csv")
pca <- princomp(mydata, scores=TRUE, cor=TRUE)
summary(pca)
loadings(pca)
plot(pca,type="lines")
biplot(pca)
```

PC1, including Cu, Ni, Pb and Cd, can be defined as an anthropogenic component due to high-level presence of human activities.

PC2 could be considered as a natural component, because the variability of the heavy metals seems to be controlled by parent rocks, moreover, Zn and Cr contents were lower than the other elements.

Cluster Analysis

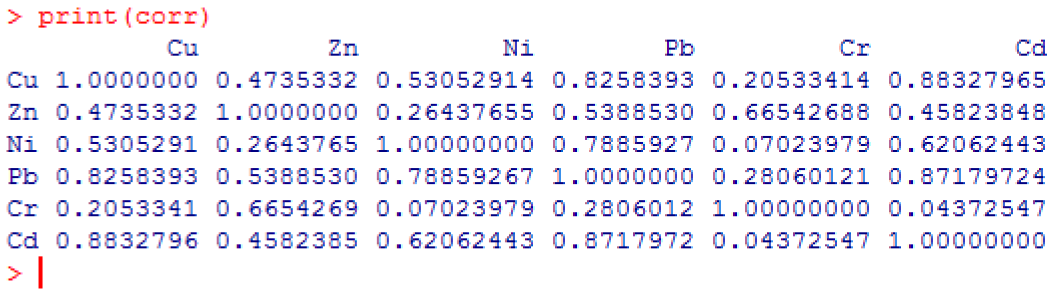




```
# Cluster Analysis
mydata = read.csv("work.csv")
trans=t(mydata)
d <- dist(trans, method = "euclidean")
clusr <- hclust(d, method="ward")
plot(clusr)
groups <- cutree(clr, k=2)
rect.hclust(clusr, k=2, border="red")
```

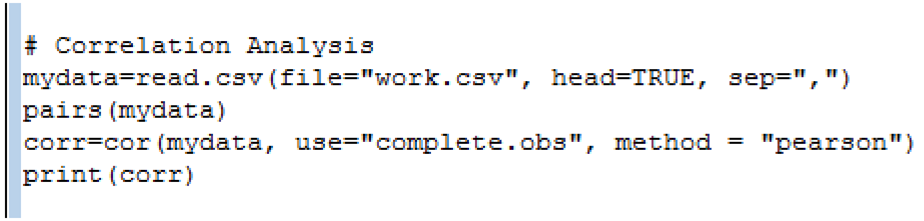
Clustering analysis re-confirmed the findings that the elements come from two different sources in the soils.

Correlation Analysis

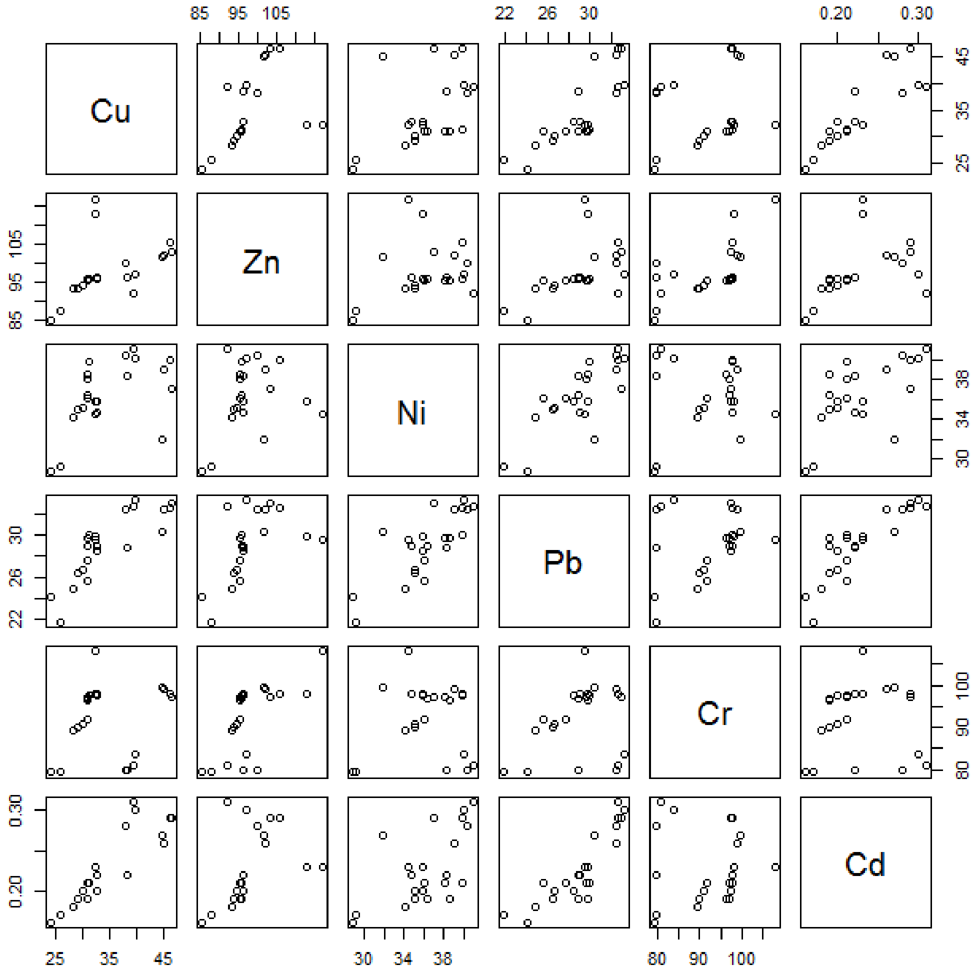


```
> print(corr)
          Cu        Zn         Ni        Pb         Cr         Cd
Cu 1.0000000 0.4735332 0.53052914 0.8258393 0.20533414 0.88327965
Zn 0.4735332 1.0000000 0.26437655 0.5388530 0.66542688 0.45823848
Ni 0.5305291 0.2643765 1.00000000 0.7885927 0.07023979 0.62062443
Pb 0.8258393 0.5388530 0.78859267 1.0000000 0.28060121 0.87179724
Cr 0.2053341 0.6654269 0.07023979 0.2806012 1.00000000 0.04372547
Cd 0.8832796 0.4582385 0.62062443 0.8717972 0.04372547 1.00000000
>
```

Anthropogenic metals, such as Cu, Ni, Pb and Cd, were significantly correlated. On the other hand, natural metals (Zn and Cr) were comparatively loosely correlated as a consequence of their external sources.



```
# Correlation Analysis
mydata=read.csv(file="work.csv", head=TRUE, sep=",")
pairs(mydata)
corr=cor(mydata, use="complete.obs", method = "pearson")
print(corr)
```



Conclusion

- This work uses high-level data engineering techniques to analyze soil carbon, nitrogen and heavy metals in the soils, disclosed relationships between carbon and nitrogen over depth, and identified heavy metal sources in the coastal soils of Mid-Atlantic region.
    - The regression analysis on soil carbon and sampling depth clearly showed a negative relationship between each other. Soil carbon and nitrogen content decreases with increasing sampling depth.
    - The PCA performed on six heavy metals identified two principal components. Cu, Ni, Pb and Cd (PC1) were related to the anthropogenic component. PC2, including Zn and Cr, could be considered as a natural component.
    - The same grouping was obtained from clustering analysis. Two main clusters of elements were extracted: the first one included natural elements (Zn and Cr) and the second cluster contained anthropogenic elements Cu, Ni, Pb and Cd.
    - In correlation analysis, anthropogenic metals, such as Cu, Ni, Pb and Cd, were significantly correlated, while natural metals (Zn and Cr) were comparatively loosely correlated.
- Data engineering plays an important role in decoding the soil conditions in the Mid-Atlantic region, and provides database for national soil investigation and survey systems.