Exercises unit 1.2: Principal Components Analysis (PCA)

The Principal Components procedure is designed to extract *k* principal components from a set of *p* quantitative variables *X*. The principal components are defined as the set of orthogonal linear combinations of *X* that have the greatest variance. Determining the principal components is often used to reduce the dimensionality of a set of predictor. **When the variables are highly correlated, the first few principal components may be sufficient to describe most of the variability present.**

Coefficient of variation is high for variable height and variable CoastDistance, so in this case, can be a solution to use PCA on these data.

Summary Statistics

	xcoord(east)	*ycoord(north)*	*factorR*	*height*	*Coastdistance*
Average	732037,	4,44837E6	188,814	696,329	26900,3
Standard deviation	21760,9	28975,9	32,9546	338,329	14025,4
Coeff. of variation	2,97265%	0,651382%	17,4535%	48,5875%	52,1386%

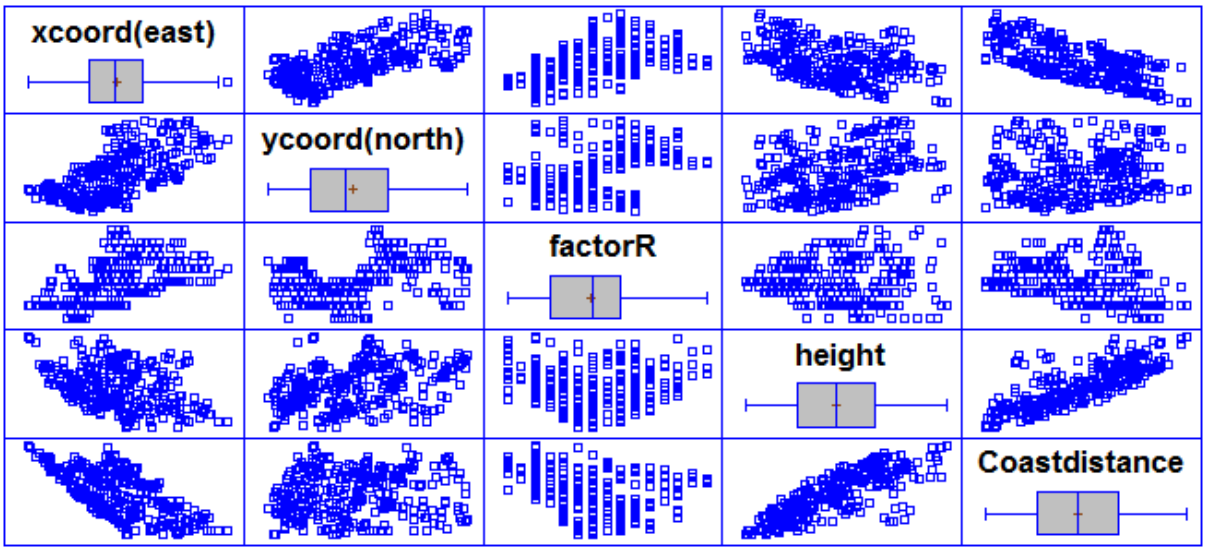
Correlations

	xcoord(east)	ycoord(north)	factorR	height	Coastdistance
xcoord(east)		0,6200	0,5291	-0,3955	-0,6759
		(295)	(295)	(295)	(295)
		0,0000	0,0000	0,0000	0,0000
ycoord(north)	0,6200		0,2886	0,3463	0,1031
	(295)		(295)	(295)	(295)
	0,0000		0,0000	0,0000	0,0769
factorR	0,5291	0,2886		-0,0605	-0,3506
	(295)	(295)		(295)	(295)
	0,0000	0,0000		0,3006	0,0000
height	-0,3955	0,3463	-0,0605		0,7994
	(295)	(295)	(295)		(295)
	0,0000	0,0000	0,3006		0,0000
Coastdistance	-0,6759	0,1031	-0,3506	0,7994	
	(295)	(295)	(295)	(295)	
	0,0000	0,0769	0,0000	0,0000	

Two variables, with strong positive correlation are CoastDistance and Height, the rest or relationships between variables is moderate or weak / positive or negative correlation. xcoord(east) and ycoord(north), xcoord(east) and factorR, xcoord(east) and height, xcoord(east) and Coastdistance, (north) and factorR, ycoord(north) and height, factorR and Coastdistance

These correlation coefficients range between -1 and +1 and measure the strength of the linear relationship between the variables. The third number (the red one) in each location of the table is a P-value which tests the statistical significance of the estimated correlations. P-values below 0,05 indicate statistically significant non-zero correlations at the 95,0% confidence level. **The following pairs of variables have P-values below 0,05 -> it means, that there is a linear relationship between these variables. xcoord(east) and ycoord(north), xcoord(east) and factorR, xcoord(east) and height, xcoord(east) and Coastdistance, (north) and factorR, ycoord(north) and height, factorR and Coastdistance, height and Coastdistance.**

Matrix plot for data is shown below:



The first principal component is that linear combination that has maximum variance. The second principal combination is that linear combination that has the next greatest variance, such to the same constraint on unit length and also to the constraint that it be uncorrelated with the first principal component. Subsequent components explain as much of the remaining variance as possible, while being uncorrelated with all of the other components.

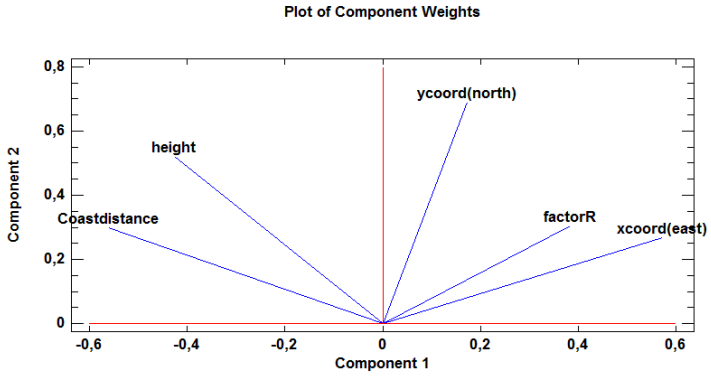
Table of Component Weights

	Component	*Component*	*Component*	*Component*	*Component*
	1	2	3	4	5
xcoord(east)	0,571681	0,267945	-0,270827	-0,0870997	0,721424
ycoord(north)	0,173131	0,689387	-0,420006	0,203197	-0,526381
factorR	0,383206	0,303767	0,846215	0,197955	-0,0749145
height	-0,425337	0,521787	0,181551	-0,705646	0,126214
Coastdistance	-0,561648	0,297346	0,0347413	0,64343	0,425356

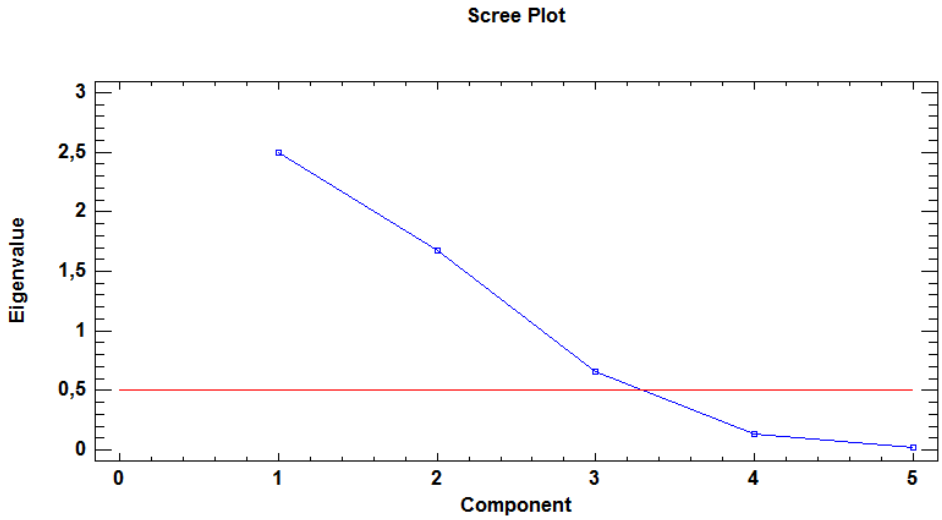
The goal of a principal components analysis is to construct *k* linear combinations of the *p* variables *X* that contain the greatest variance. The linear combinations take the form:
PC1 = 0,571681*xcoord(east) + 0,173131*ycoord(north) + 0,383206*factorR - 0,425337*height - 0,561648*Coastdistance
PC2 = 0,267945*xcoord(east) + 0,689387*ycoord(north) + 0,303767*factorR + 0,521787*height + 0,297346*Coastdistance
PC3 = -0,270827*xcoord(east) - 0,420006*ycoord(north) + 0,846215*factorR + 0,181551*height + 0,0347413*Coastdistance

Principal Components Analysis

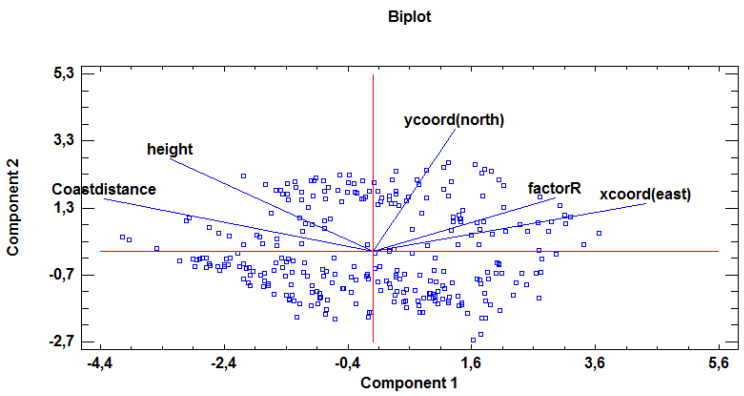
Component		*Percent of*	*Cumulative*
Number	*Eigenvalue*	*Variance*	*Percentage*
1	2,50076	50,015	50,015
2	1,67475	33,495	83,510
3	0,660021	13,200	96,711
4	0,139499	2,790	99,501
5	0,0249723	0,499	100,000

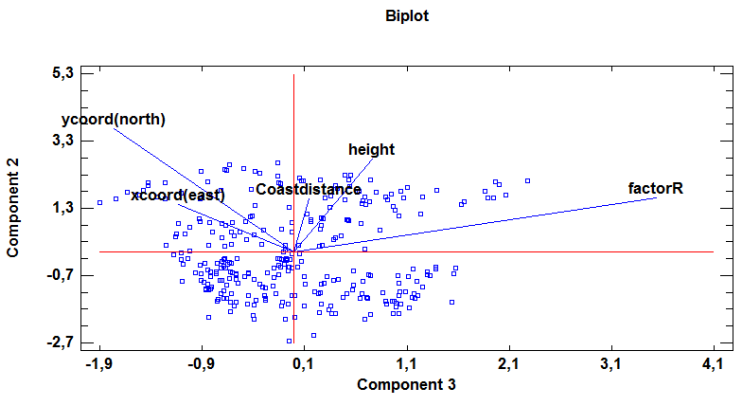


In the graph of Component Weights – every point belongs to one variable and its primary purpose is to compare the distance between clusters. Whilst is a short distance between the variables (characters), it means that there is a strong correlation between them. When the lines are perpendicular, it indicates zero correlation between variables. And finally, the longer the line is, the bigger impact on component it has. Moreover, it can be seen, there are two variables with negative values – CoastDistance and height.



In this Scree Plot can be seen, how was estimated eigenvalue (for the value 0,5) in order to have 3 components in final.





The first principal component is that linear combination that has maximum variance, subject to the constraint that the coefficient vector has unit length. In our case, first principal component can be identified with the two variables with a negative weight. It explains the interest in concentrating the displaced in specific places with areas of common recreation.

The second principal combination is that linear combination that has the next greatest variance, such to the same constraint on unit length and also to the constraint that it be uncorrelated with the first principal component. Subsequent components explain as much of the remaining variance as possible, while being uncorrelated with all of the other components.

Principal Components Analysis

Component		*Percent of*	*Cumulative*
Number	*Eigenvalue*	*Variance*	*Percentage*
1	2,50076	50,015	50,015
2	1,67475	33,495	83,510
3	0,660021	13,200	**96,711**
4	0,139499	2,790	99,501
5	0,0249723	0,499	100,000

$traceS$=$traceQ\ D\ QT$= λipi=1 and in addition the trace of S coincides with the total variation of the data.

The percentage of variability explained by the first "m" components is:

$$P_m = 100\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_m + \cdots + \lambda_p}$$

Percentage Criteria: We chose two principal components as the explained cumulative percentage of variance, Pm, is greater than 80%. The first component explains a percentage of the 50,015%. The first three principal components explain percentage of variability equal 96,711%

Kaiser Criterion: As we work with standardized variables, we look at eigenvalues greater than "0.7".
We have two principal components that meet this condition with a explained cumulative percentage of variance of 83.510%. That choice is also suitable for the percentage criterion.

In **analysis options** we marked "Extract by minimum eigenvalue" and write as eigenvalue minimum = 0.7, so that the red line appears. Component 3 is close to it.
From the second component the percentage of variance increases more slowly. We should draw the percentage of cumulative variance to see from the number of components that stabilizes.