Exercises unit 1.1: introduction to multivariate analysis

1/A

Summary Statistics

	factorR	*height*	*Coastdistance*
Count	295	295	295
Average	188,814	696,329	26900,3
Minimum	130,0	33,0	825,02
Maximum	270,0	1500,0	57371,7
Range	140,0	1467,0	56546,7

On the left can be seen the basic summary statistics. The number of amount of all variables is 295. There is a big Range of height and Coastdistance. Points with the lower hight are nearer to coast. As we can see, Coeff. of variation for factorR is 17,45%, so there is small dispersion of variables and the most of them are near the average. For height and Coastdistance there is bigger dispersion, about half of the points are near the average. Values of Stnd. kurtosis are outside the range of -2 to +2, so data don't follow normal distribution

Summary Statistics for Coastdistance

Count	295
Average	26900,3
Median	26957,1
Standard deviation	14025,4
Minimum	825,02
Maximum	57371,7
Range	56546,7
Lower quartile	15472,2
Upper quartile	38327,5
Interquartile range	22855,3

Summary Statistics for factorR

Count	295
Average	188,814
Median	190,0
Standard deviation	32,9546
Minimum	130,0
Maximum	270,0
Range	140,0
Lower quartile	160,0
Upper quartile	210,0
Interquartile range	50,0

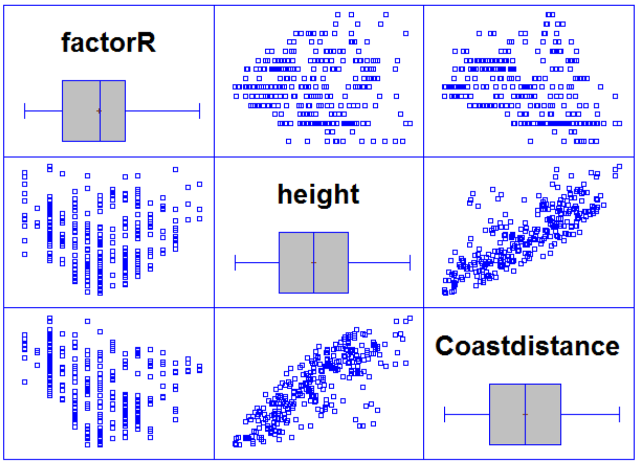
Outliers appear separated from the body of data on a box-plot.

As we can see, any of these variables does not have outliers.

Summary Statistics for height

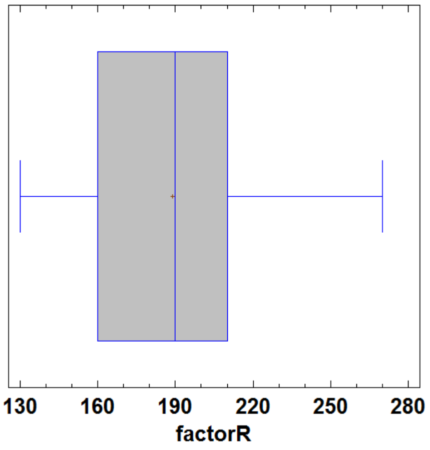
Count	295
Average	696,329
Median	690,0
Standard deviation	338,329
Minimum	33,0
Maximum	1500,0
Range	1467,0
Lower quartile	406,0
Upper quartile	980,0
Interquartile range	574,0

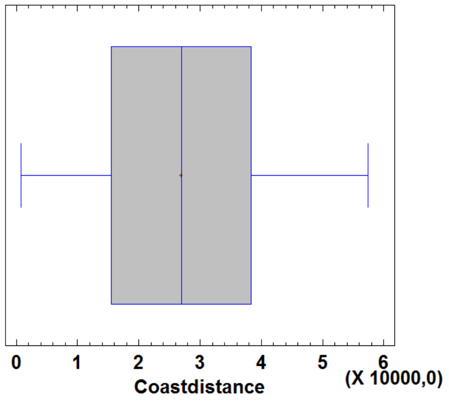
From Boxplot and Whisker plot, it can be seen, that those variables are positively skewed (have a positively skewed boxplot)

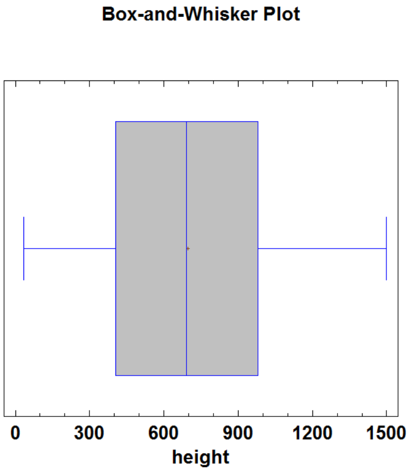


1/B

	factorR	height	coastDistance
Standard deviation	32,9546	338,329	14025,4
Coeff. of variation	17,4535%	48,5875%	52,1386%







Correlations

	factorR	height	Coastdistance
factorR		-0,0605	-0,3506
		(295)	(295)
		0,3006	0,0000
height	-0,0605		0,7994
	(295)		(295)
	0,3006		0,0000
Coastdistance	-0,3506	0,7994	
	(295)	(295)	
	0,0000	0,0000	

1/C

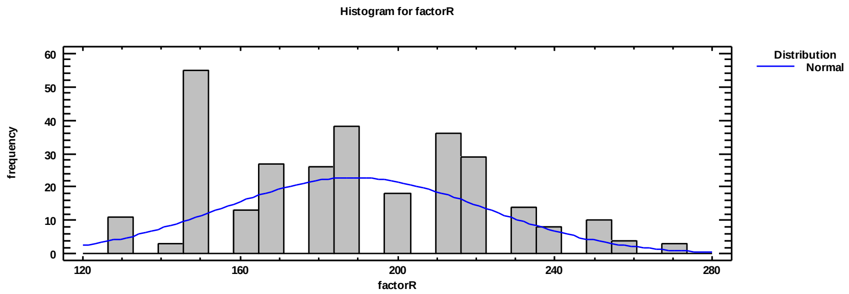
	factorR	height	coastdistance
Stnd. skewness	1,5606	0,860222	-0,284002
Stnd. kurtosis	-2,4803	-2,88126	-3,71268

2

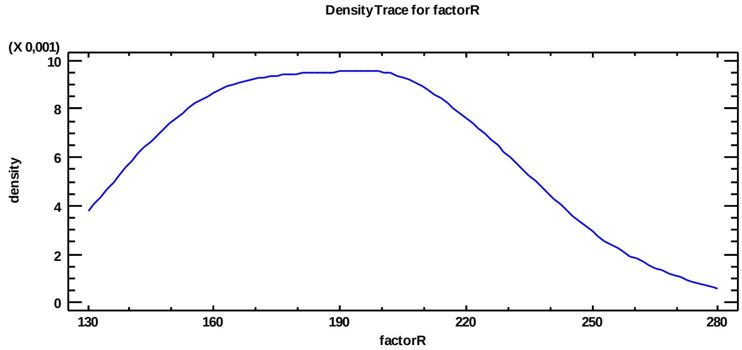
R-factor

Tests for Normality for factorR

Test	*Statistic*	*P-Value*
Chi-Square	839,207	0,0
Shapiro-Wilk W	0,941662	9,10383E-15
Skewness Z-score	1,11183	0,26621
Kurtosis Z-score	-3,87479	0,000106758



Since the smallest P-value amongst the tests performed is less than 0,05, we cen reject the idea that factorR comes from normal distribution with 95% confidence. The Shapiro-Wilk test is based upon comparing the quantiles of the fitted normal distribution to the quantiles of the data. Skewness Z-score uses standardized skewness to look for the lack of symmetry in the data. This test can not reject normality of factorR. This testifies to the value of distribution symmetric. Kurtosis Z-score show us that the shape of the distribution. P-value is a little more than 0, so distribution is leptocurtic



Kolmogorov-Smirnov Test

	Normal
DPLUS	0,114459
DMINUS	0,0923988
DN	0,114459
P-Value	0,00087936

Modified Kolmogorov-Smirnov D

	Normal
D	0,114459
Modified Form	1,98036
P-Value	<0.01

Kuiper V

	Normal
V	0,206857
Modified Form	3,58785
P-Value	<0.01

Cramer-Von Mises W^2

	Normal
W^2	0,512972
Modified Form	0,513357
P-Value	<0.05

Watson U^2

	Normal
U^2	0,503409
Modified Form	0,504436
P-Value	<0.01

Anderson-Darling A^2

	Normal
A^2	3,53617
Modified Form	3,53617
P-Value	<0.05

Goodness-of-Fit Tests for factorR

Chi-Square Test

	Lower Limit	*Upper Limit*	*Observed Frequency*	*Expected Frequency*	*Chi-Square*
at or below		125,33	0	7,97	7,97
	125,33	135,864	11	7,97	1,15
	135,864	142,748	3	7,97	3,10
	142,748	148,06	0	7,97	7,97
	148,06	152,483	55	7,97	277,38
	152,483	156,333	0	7,97	7,97
	156,333	159,784	0	7,97	7,97
	159,784	162,943	13	7,97	3,17
	162,943	165,88	0	7,97	7,97
	165,88	168,645	0	7,97	7,97
	168,645	171,275	27	7,97	45,41
	171,275	173,798	0	7,97	7,97
	173,798	176,236	0	7,97	7,97
	176,236	178,606	0	7,97	7,97
	178,606	180,925	26	7,97	40,76
	180,925	183,205	0	7,97	7,97
	183,205	185,459	0	7,97	7,97
	185,459	187,697	0	7,97	7,97
	187,697	189,93	0	7,97	7,97
	189,93	192,168	38	7,97	113,08
	192,168	194,422	0	7,97	7,97
	194,422	196,702	0	7,97	7,97
	196,702	199,021	0	7,97	7,97
	199,021	201,392	18	7,97	12,61
	201,392	203,829	0	7,97	7,97
	203,829	206,352	0	7,97	7,97
	206,352	208,982	0	7,97	7,97
	208,982	211,747	36	7,97	98,52
	211,747	214,684	0	7,97	7,97
	214,684	217,843	0	7,97	7,97
	217,843	221,294	29	7,97	55,45
	221,294	225,144	0	7,97	7,97
	225,144	229,567	0	7,97	7,97
	229,567	234,879	14	7,97	4,56
	234,879	241,764	8	7,97	0,00
	241,764	252,297	10	7,97	0,52
above	252,297		7	7,97	0,12

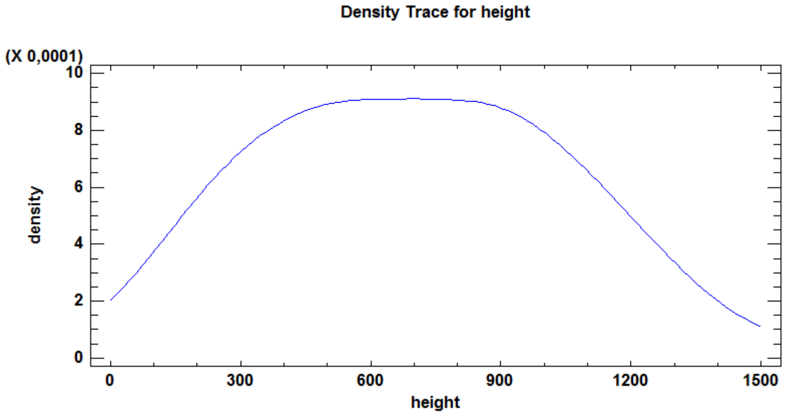
Chi-Square = 839,207 with 34 d.f.   P-Value = 0,0

Kurtosis is smaller, this graph is less pointed than a normal distribution.
Histogram for factorR shows us that it does not follow normal distribution. When the normal distribution is in the highest position, histogram is a little bit lower. It keeps symmetry except from values between 145-155. At the quantile-quantile plot we see that data does not follow normal distribution and there are fails in small values of variable height.

Height

Tests for Normality for height

Test	*Statistic*	*P-Value*
Chi-Square	40,7593	0,197492
Shapiro-Wilk W	0,958845	5,9155E-7
Skewness Z-score	0,617574	0,536853
Kurtosis Z-score	-5,05144	4,39226E-7



Kolmogorov-Smirnov Test

	Normal
DPLUS	0,0675349
DMINUS	0,056686
DN	0,0675349
P-Value	0,135651

Modified Kolmogorov-Smirnov D

	Normal
D	0,0675349
Modified Form	1,16849
P-Value	>=0.10

Kuiper V

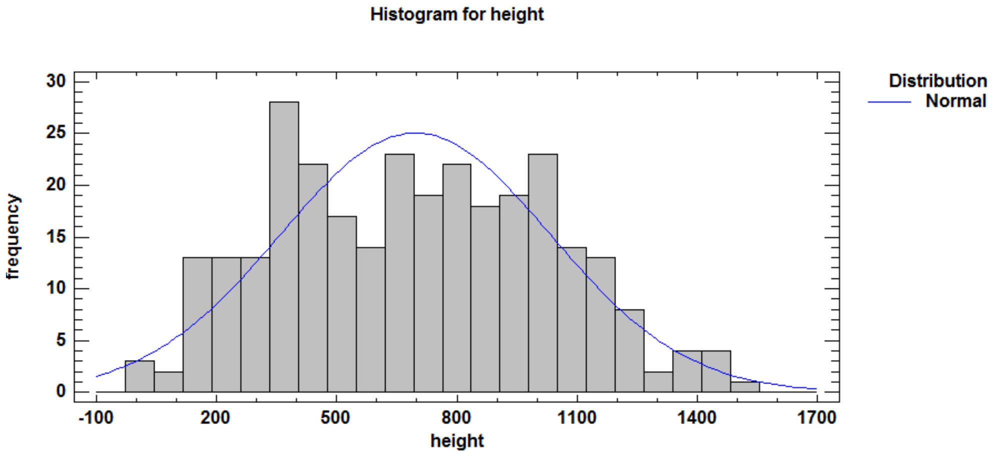
	Normal
V	0,124221
Modified Form	2,15456
P-Value	<0.01

Cramer-Von Mises W^2

	Normal
W^2	0,258781
Modified Form	0,258305
P-Value	>=0.10

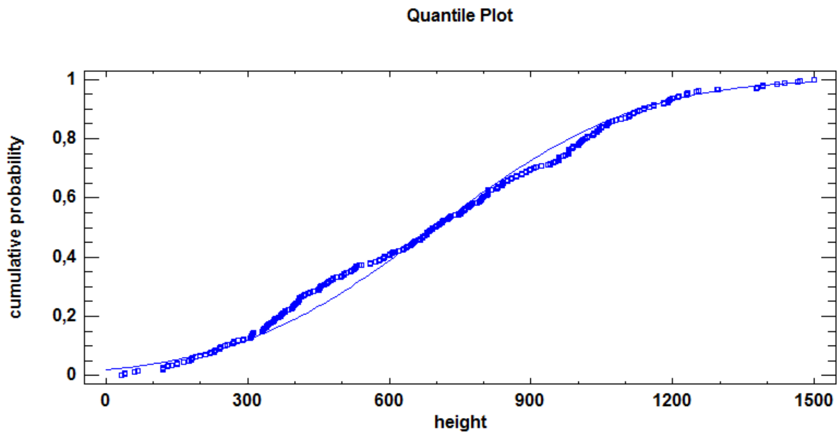
Watson U^2

	Normal
U^2	0,255923
Modified Form	0,256278
P-Value	<0.05



Anderson-Darling A^2

	Normal
A^2	1,60901
Modified Form	1,60901
P-Value	>=0.10



Since the smallest P-value amongst the tests performed is less than 0,05, we can reject the idea that height comes from normal distribution with 95% confidence. The Shapiro-Wilk test is based upon comparing the quantiles of the fitted normal distribution to the quantiles of the data. It reject null hipotesis. Skewness Z-score uses standardized skewness to look for the lack of symmetry in the data. This test cannot reject normality of height. This testifies to the value of distribution symmetric. Kurtosis Z-score show us that the shape of the distribution. Value equal almost 0 which is the value for normal distribution.

Goodness-of-Fit Tests for height

Chi-Square Test

	Lower Limit	*Upper Limit*	*Observed Frequency*	*Expected Frequency*	*Chi-Square*
at or below		44,5702	3	7,97	3,10
	44,5702	152,716	10	7,97	0,52
	152,716	223,4	10	7,97	0,52
	223,4	277,933	11	7,97	1,15
	277,933	323,341	9	7,97	0,13
	323,341	362,868	14	7,97	4,56
	362,868	398,298	13	7,97	3,17
	398,298	430,728	12	7,97	2,03
	430,728	460,883	8	7,97	0,00
	460,883	489,272	8	7,97	0,00
	489,272	516,273	5	7,97	1,11
	516,273	542,172	8	7,97	0,00
	542,172	567,197	2	7,97	4,47
	567,197	591,533	6	7,97	0,49
	591,533	615,339	5	7,97	1,11
	615,339	638,75	5	7,97	1,11
	638,75	661,888	6	7,97	0,49
	661,888	684,866	8	7,97	0,00
	684,866	707,792	8	7,97	0,00
	707,792	730,77	8	7,97	0,00
	730,77	753,908	4	7,97	1,98
	753,908	777,319	8	7,97	0,00
	777,319	801,124	8	7,97	0,00
	801,124	825,461	7	7,97	0,12
	825,461	850,486	8	7,97	0,00
	850,486	876,384	6	7,97	0,49
	876,384	903,385	6	7,97	0,49
	903,385	931,775	3	7,97	3,10
	931,775	961,93	9	7,97	0,13
	961,93	994,359	11	7,97	1,15
	994,359	1029,79	10	7,97	0,52
	1029,79	1069,32	14	7,97	4,56
	1069,32	1114,72	8	7,97	0,00
	1114,72	1169,26	9	7,97	0,13
	1169,26	1239,94	12	7,97	2,03
	1239,94	1348,09	4	7,97	1,98
above	1348,09		9	7,97	0,13

Chi-Square = 40,7593 with 34 d.f.   P-Value = 0,197492

Kurtosis is smaller, this graph is less pointed than a normal distribution. Histogram for height show us that it follows normal distribution. When the normal distribution is in the highest position, histogram is a little bit lower, and between 300-400 it is above the normal distribution. It keeps symmetry except from values between 300-400. At the quantile-quantile plot we see that data follow normal distribution and there are fails in small values of variable height.

If we only analyse symmetry, we can conclude that data follow the normal distribution, in general data don't follow normal distribution, as indicate some tests

Coast Distance

Tests for Normality for Coastdistance

Test	*Statistic*	*P-Value*
Chi-Square	75,3763	0,0000573
Shapiro-Wilk W	0,945549	8,69638E-13
Skewness Z-score	0,204515	0,837946
Kurtosis Z-score	-8,88095	0,0

Kolmogorov-Smirnov Test

	Normal
DPLUS	0,0586381
DMINUS	0,0737405
DN	0,0737405
P-Value	0,0808598

Modified Kolmogorov-Smirnov D

	Normal
D	0,0737405
Modified Form	1,27586
P-Value	<0.10

Kuiper V

	Normal
V	0,132379
Modified Form	2,29605
P-Value	<0.01

Cramer-Von Mises W^2

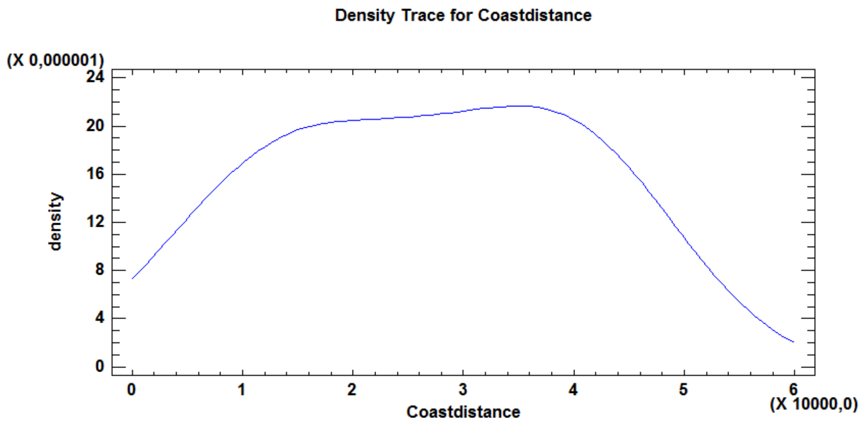
	Normal
W^2	0,416905
Modified Form	0,416965
P-Value	<0.10

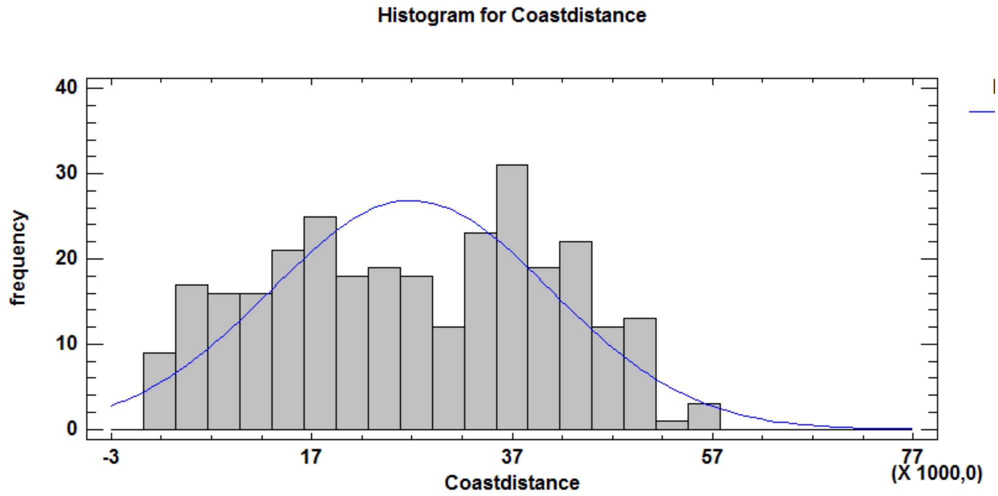
Watson U^2

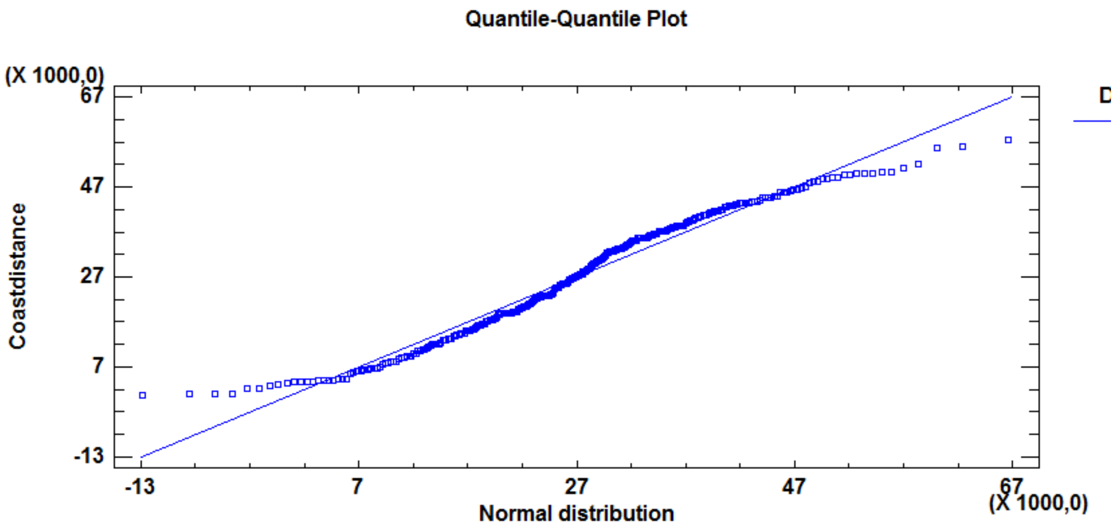
	Normal
U^2	0,415852
Modified Form	0,416641
P-Value	<0.01

Anderson-Darling A^2

	Normal
A^2	2,57794
Modified Form	2,57794
P-Value	<0.05







Since the smallest P-value amongst the tests performed is less than 0,05, we cen reject the idea that Coastdistance comes from normal distribution with 95% confidence. The Shapiro-Wilk test is based upon comparing the quantiles of the fitted normal distribution to the quantiles of the data. Skewness Z-score uses standardized skewness to look for the lack of symmetry in the data. This test can not reject normality of Coastdistance. This testifies to the value odf distribution symmetric. Kurtosis Z-score show us that the shape of the distribution. Value=0 is the value for normal distribution

Goodness-of-Fit Tests for Coastdistance

Chi-Square Test

	Lower Limit	*Upper Limit*	*Observed Frequency*	*Expected Frequency*	*Chi-Square*
at or below		-118,417	0	7,97	7,97
	-118,417	4364,79	20	7,97	18,14
	4364,79	7294,96	10	7,97	0,52
	7294,96	9555,64	12	7,97	2,03
	9555,64	11438,0	8	7,97	0,00
	11438,0	13076,6	9	7,97	0,13
	13076,6	14545,4	10	7,97	0,52
	14545,4	15889,8	7	7,97	0,12
	15889,8	17139,8	7	7,97	0,12
	17139,8	18316,7	7	7,97	0,12
	18316,7	19436,0	14	7,97	4,56
	19436,0	20509,7	6	7,97	0,49
	20509,7	21547,1	5	7,97	1,11
	21547,1	22556,0	7	7,97	0,12
	22556,0	23542,8	9	7,97	0,13
	23542,8	24513,3	2	7,97	4,47
	24513,3	25472,5	7	7,97	0,12
	25472,5	26425,1	2	7,97	4,47
	26425,1	27375,5	8	7,97	0,00
	27375,5	28328,0	6	7,97	0,49
	28328,0	29287,2	3	7,97	3,10
	29287,2	30257,7	5	7,97	1,11
	30257,7	31244,5	5	7,97	1,11
	31244,5	32253,4	2	7,97	4,47
	32253,4	33290,8	8	7,97	0,00
	33290,8	34364,5	9	7,97	0,13
	34364,5	35483,8	6	7,97	0,49
	35483,8	36660,7	14	7,97	4,56
	36660,7	37910,7	9	7,97	0,13
	37910,7	39255,1	10	7,97	0,52
	39255,1	40723,9	11	7,97	1,15
	40723,9	42362,5	8	7,97	0,00
	42362,5	44244,9	16	7,97	8,08
	44244,9	46505,5	10	7,97	0,52
	46505,5	49435,7	11	7,97	1,15
	49435,7	53918,9	9	7,97	0,13
above	53918,9		3	7,97	3,10

Chi-Square = 75,3763 with 34 d.f.   P-Value = 0,0000573

Kurtosis is smaller, this graph is less pointed than a normal distribution.

Histogram for Coastdistance show us that it does not follow normal distribution. When the normal distribution is in the highest position, histogram is low. It does not keeps entirely symmetry. At the quantile-quantile plot we see that data do not follow normal distribution and there are fails in small values of variable coastdistance.

3/A

Pearson's correlation coefficient gives a measure of the relationship between two variables on a scale from 1 to 1

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	

P-value in pairs of variable height/factorR and ycoord/Coastdistance indicates that there aren't any correlations.

This table shows Pearson product moment correlations between each pair of variables. These correlation coefficients range between -1 and +1 and measure the strength of the linear relationship between the variables. Also shown in parentheses is the number of pairs of data values used to compute each coefficient. The third number 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:

xcoord(east) and ycoord(north)
xcoord(east) and factorR
xcoord(east) and height
xcoord(east) and Coastdistance
ycoord(north) and factorR
ycoord(north) and height
factorR and Coastdistance
height and Coastdistance

3/B

Spearman Rank Correlations

	xcoord(east)	ycoord(north)	factorR	height	Coastdistance
xcoord(east)		0,6323	0,5773	-0,3635	-0,6467
		(295)	(295)	(295)	(295)
		0,0000	0,0000	0,0000	0,0000
ycoord(north)	0,6323		0,2853	0,3593	0,1304
	(295)		(295)	(295)	(295)
	0,0000		0,0000	0,0000	0,0254
factorR	0,5773	0,2853		-0,0800	-0,3928
	(295)	(295)		(295)	(295)
	0,0000	0,0000		0,1704	0,0000
height	-0,3635	0,3593	-0,0800		0,7964
	(295)	(295)	(295)		(295)
	0,0000	0,0000	0,1704		0,0000
Coastdistance	-0,6467	0,1304	-0,3928	0,7964	
	(295)	(295)	(295)	(295)	
	0,0000	0,0254	0,0000	0,0000	

The Partial Correlations indicates non-zero correlations between all the variables except from factorR/Coastdistance because P-value between this these variables is >0,05.

This table shows Spearman rank correlations between each pair of variables. These correlation coefficients range between -1 and +1 and measure the strength of the association between the variables. In contrast to the more common Pearson correlations, the Spearman coefficients are computed from the ranks of the data values rather than from the values themselves. Consequently, they are less sensitive to outliers than the Pearson coefficients. Also shown in parentheses is the number of pairs of data values used to compute each coefficient. The third number 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:

xcoord(east) and ycoord(north)
xcoord(east) and factorR
xcoord(east) and height
xcoord(east) and Coastdistance
ycoord(north) and factorR
ycoord(north) and height
ycoord(north) and Coastdistance
factorR and Coastdistance
height and Coastdistance

Partial Correlations

	xcoord(east)	ycoord(north)	factorR	height	Coastdistance
xcoord(east)		0,9434	0,4159	-0,4099	-0,7947
		(295)	(295)	(295)	(295)
		0,0000	0,0000	0,0000	0,0000
ycoord(north)	0,9434		-0,3280	0,4938	0,7148
	(295)		(295)	(295)	(295)
	0,0000		0,0000	0,0000	0,0000
factorR	0,4159	-0,3280		0,4039	0,0833
	(295)	(295)		(295)	(295)
	0,0000	0,0000		0,0000	0,1557
height	-0,4099	0,4938	0,4039		0,1328
	(295)	(295)	(295)		(295)
	0,0000	0,0000	0,0000		0,0233
Coastdistance	-0,7947	0,7148	0,0833	0,1328	
	(295)	(295)	(295)	(295)	
	0,0000	0,0000	0,1557	0,0233	

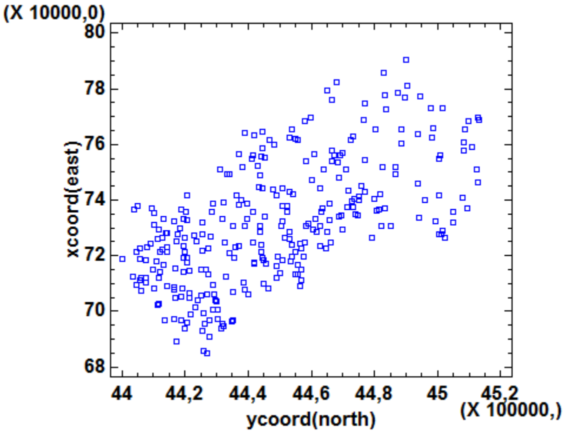
This test indicate that there is no relationship between variables height/factorR. P-value of the rest of the variables is <0,05.

This table shows partial correlation coefficients between each pair of variables. The partial correlations measure the strength of the linear relationship between the variables having first adjusted for their relationship to other variables in the table. They are helpful in judging how useful one variable would be in improving the prediction of the second variable given that information from all the other variables has already been taken into account. Also shown in parentheses is the number of pairs of data values used to compute each coefficient. The third number 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:

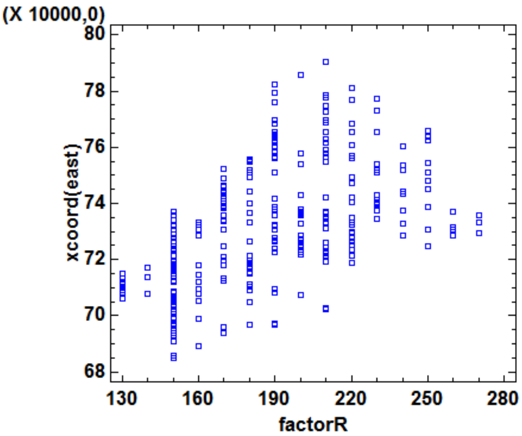
- xcoord(east) and ycoord(north)
- xcoord(east) and factorR
- xcoord(east) and height
- xcoord(east) and Coastdistance
- ycoord(north) and factorR
- ycoord(north) and height
- ycoord(north) and Coastdistance
- factorR and height
- height and Coastdistance

3/C

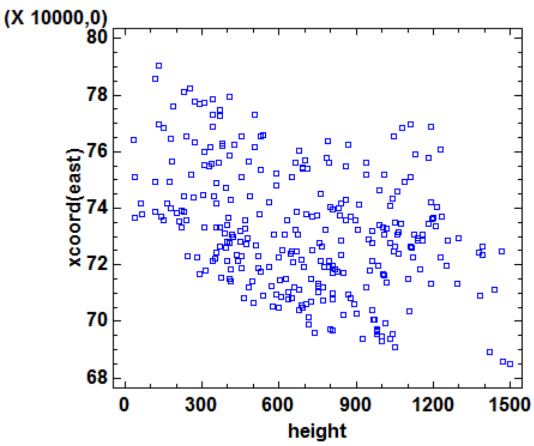
Plot of xcoord(east) vs ycoord(north)



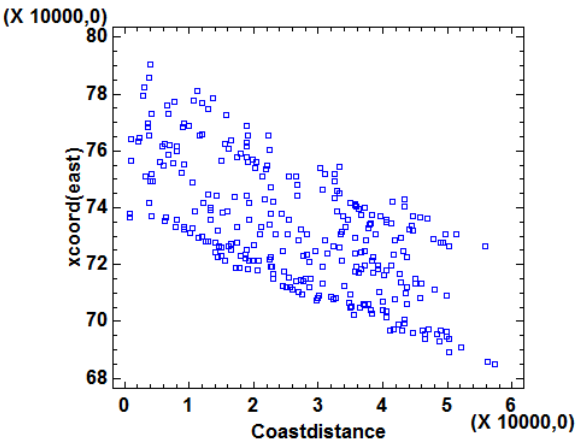
Plot of xcoord(east) vs factorR

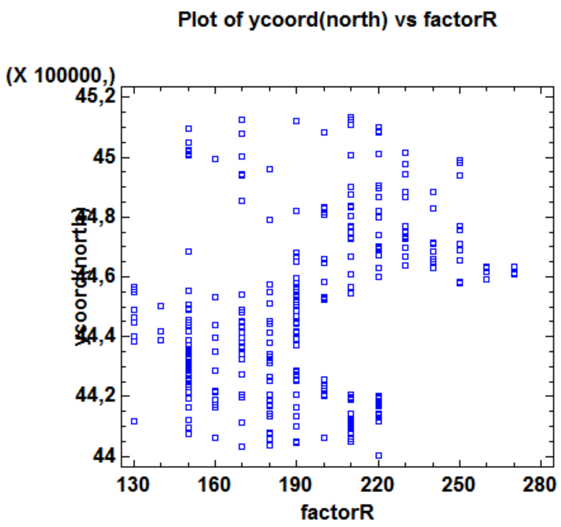


Plot of xcoord(east) vs height

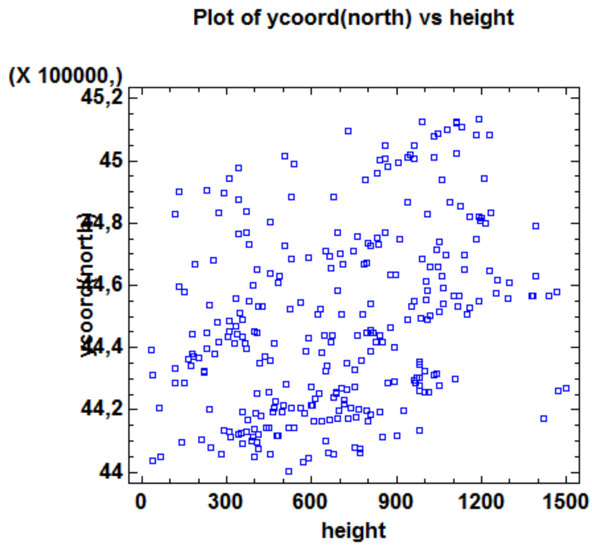


Plot of xcoord(east) vs Coastdistance

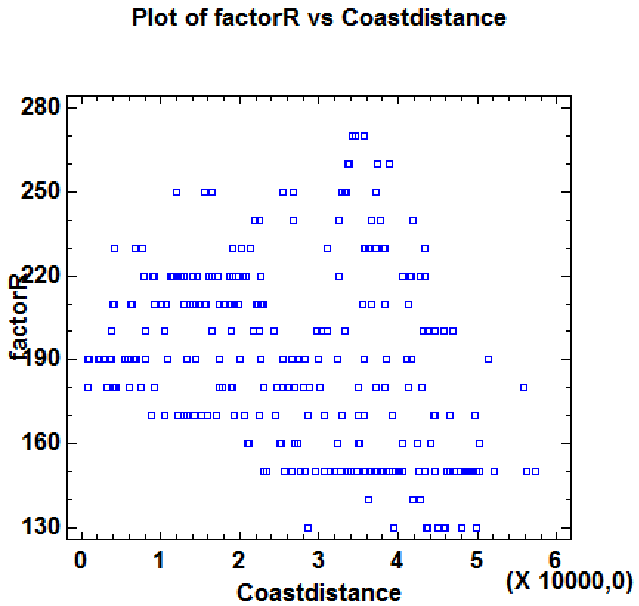




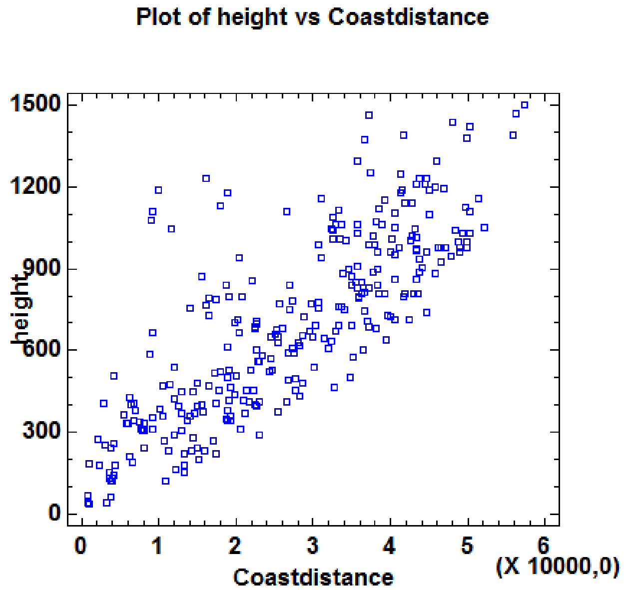
Plot of ycoord(north) vs factorR
(X 100000,)
ycoord(north)
45,2
45
44,8
44,6
44,4
44,2
44
130
160
190
220
250
280
factorR



Plot of ycoord(north) vs height
(X 100000,)
ycoord(north)
45,2
45
44,8
44,6
44,4
44,2
44
0
300
600
900
1200
1500
height



Plot of factorR vs Coastdistance
factorR
280
250
220
190
160
130
0
1
2
3
4
5
6
Coastdistance
(X 10000,0)



Plot of height vs Coastdistance
height
1500
1200
900
600
300
0
0
1
2
3
4
5
6
Coastdistance
(X 10000,0)