Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Uma Alternativa Estatística para a Análise Quantitativa de Dados Geológicos de Sistemas Granulométricos Fechados

P. S. Lucio, A. P. Serra e D. Souza

Relatório Técnico RTE-02/97

Relatório Técnico
Série Ensino

Uma alternativa estatística para a análise quantitativa de dados geológicos de sistemas granulométricos fechados

Paulo Sérgio Lucio,
Professor Adjunto do Departamento de Estatística da UFMG
Ana Paula Serra
e
Daniela de Souza
Bolsistas de Iniciação Científica do CNPq
Graduação do Departamento de Estatística da UFMG

02 de dezembro de 1997

Resumo

Este trabalho consiste na analise de dados reais em geologia quantitativa de sistemas fechados sob dois diferentes enfoques. O primeiro centrado nas técnicas estatísticas (Análise de Regressão) utilizadas na tese de mestrado da professora Cristina Augustin [Ausgustin, 1981] e o segundo considerando as técnicas estatísticas (Análise Fatorial, de Conglomerados e Modelos Lineares Generalizados) que melhor se adequaria à análise de sistemas fechados.

Nosso objetivo foi direcionado no sentido de priorizar a modelagem das variáveis sob estudo, uma vez que a textura tem um papel de fundamental importância na erodibilidade do solo e por conseqüência na vegetação que ali se instala, pois influencia diretamente na velocidade de infiltração, na resistência à dispersão, no deslocamento por salpico, na abrasão e nas formas de transporte do solo.

Algumas frações granulométricas são removidas mais facilmente que outras. A remoção de sedimentos é maior na fração areia media e diminui nas partículas maiores e menores, com as areias (Sand) apresentando os maiores índices de erodibilidade. A variação no teor de materiais como Silt (silte) e Clay (argilas) são fatores importantes, visto que solos mais siltosos são suceptíveis à erosão, e as argilas, se por um lado podem, às vezes, dificultar a infiltração das aguas, por outro lado são mais difíceis de serem removidas - devido à formação de agregados. Por outro lado, o teor de calcário (Gypsum) afeta de diversas maneiras a instalação da vegetação no solo.

Encontra-se uma elevada correlação entre Gypsum e Silt, em especial para horizontes com baixo teor de Sand e Clay. Contudo, outros componentes podem também afetar a estabilidade dos agregados, que foram aqui evidenciados por algumas características físicas e químicas (ph e condutividade elétrica) e o conteúdo de matéria orgânica. Solos com características latossólicas são quimicamente pobres em bases e ricos em sesquióxidos de Fe e Al, tendendo, em geral, a estruturar-se por meio de agregação ou em estruturas maciças. Estas formas de organização estrutural dão ao solo alta porosidade entre as partículas (*Augustin, 1997*; comunicação verbal).

Nota: Os autores deste trabalho, em breve, publicarão um estudo sobre: "*Fatores e Processos que influenciam Sistemas Fechados na Gênese e Evolução de Voçorocas*".

I. Introdução

A proposta de reavaliação das técnicas de análise estatística utilizadas na tese de mestrado da professora Cristina Augustin [*Augustin, 1981*] partiu da própria professora junto ao professor Paulo Sérgio Lucio que pretendia, com esta oportunidade, um

envolvimento maior de parte de suas orientandas em PBIC, Ana Paula Serra e Daniela de Souza, com a análise de dados reais em geologia quantitativa de sistemas fechados.

Inicialmente reproduzimos as mesmas etapas que foram apresentadas no capítulo "Statistical Analyses" da dissertação em estudo [*Augustin, 1981*] , sendo que alguns recursos foram adicionados e serão justificados na explanação dos resultados. Posteriormente, fizemos uma nova análise sob outro enfoque estatístico, sendo que em momento algum tivemos a pretensão de desmerecer todo o esforço da professora. Simplesmente, aproveitamos o conjunto de dados da tese para nos familiarizarmos com as técnicas estatísticas estudadas até o momento neste projeto de bolsa de iniciação científica.

A análise envolve, primeiramente, a determinação das relações entre as variáveis do solo (características físicas e quimicas) e o padrão desta relação. Estas análises foram realizadas com a utilização de dois "*softwares*" estatísticos: *MINITAB* e *GLIM*. No *MINITAB*, por ser um programa simples e de fácil acesso, foram feitas todas as análises de forma análoga às da tese. O uso do *GLIM*, no nosso trabalho, é justificado pelo fato dele tratar de Modelos Lineares Generalizados. Usaremos este conceito na análise feita adiante.

É importante ressaltar que a forma como nos explicitamos a respeito dos resultados são considerações puramente estatísticas. É essencial que isto seja colocado, uma vez que para esta revisão não tivemos o acompanhamento de um profissional da área geológica ou geomorfológica.

II. Um estudo de caso

Na tese [*Augustin, 1981*], a abordagem estatística procurou primordialmente:

1- associar os fatores ambientais a um padrão na ocorrência de espécies de plantas e
2- relacionar as variáveis do solo entre si.

Nosso trabalho concentrou-se no item **2** e a técnica usada foi a de análise de regressão. As variáveis do solo usadas no estudo foram consideradas as mais importantes para a caracterização da vegetação no que tange ao estudo de área. São elas:

$\mathbf{V_1}$- Clay [%] (Argila)
$\mathbf{V_2}$- Silt [%] (Silte)
$\mathbf{V_3}$- Sand [%] (Areia)
$\mathbf{V_4}$- PH
$\mathbf{V_5}$- Organic matter [%] (Material Orgânico)
$\mathbf{V_6}$- Electrical conductivity [ohms] (Eletro-condutividade)
$\mathbf{V_7}$- Gypsum [%] (Calcário)

Observe que $\mathbf{V_1}$, $\mathbf{V_2}$, $\mathbf{V_3}$ e $\mathbf{V_7}$ formam particularmente um sistema fechado, o que é freqüentemente encontrado em geologia, no que concerne a análise de dados granulométricos laboratoriais. Neste caso, a soma das variáveis é constante; assim temos:

$$\mathbf{V_1 + V_2 + V_3 + V_7 = 100\%.}$$

Desta forma, é evidente que se o valor de um grupo aumenta, aquele do complementar, diminui [*Guillaume, 1977*]. A correlação é necessariamente negativa entre pelo menos duas das combinações possíveis. Assim a técnica estatística de melhor adequação para a análise sobre este sistema é a análise fatorial e de conglomerados, que será comentada no final deste estudo (**Apêndice A**).

Com a meta de avaliar a medida de relação (dependência de $\mathbf{V_7}$ com relação aos outros fatores do solo $\mathbf{V_1}$ a $\mathbf{V_6}$), uma análise de regressão simples foi aplicada, em um primeiro estágio.

Análise descritiva geral

Variable	Gypsum	Clay	Sand	Silt	PH	EC	Organic
N	85	85	85	85	85	85	83
N*	65	65	65	65	65	65	67
Mean	10.862	21.02	14.29	64.76	7.6629	1.853	0.9610
Median	13.000	12.60	13.40	67.00	7.700	2.350	0.930
StDev	7.842	12.78	9.69	16.29	0.2248	0.931	0.4367
Min	0.058	0.60	0.40	31.00	7.1000	0.140	0.2200
Max	30.200	49.60	45.40	90.00	8.3000	2.930	2.3700
Q1	0.405	10.60	7.40	51.00	7.5000	0.805	0.7400
Q3	17.150	32.60	17.90	78.50	7.8000	2.450	1.1200

Observe que média das variáveis se apresenta em torno de vários valores. Isso se deve ao fato das diferentes escalas usadas e a natureza da flutuação de certas variáveis. O número de valores omissos (não observados) é bem significativo. A variável Silt tem maior variabilidade e o PH o menor desvio padrão, talvez devido ao efeito de escala e suporte, respectivamente.

III. Gráficos de Dispersão (Correlações e Similares)

Na dissertação [*Augustin, 1981*], realizou-se primeiramente, uma análise que buscou a identificação do tipo de relacionamento entre algumas características físico-química do solo. Neste passo, foram utilizados gráficos de dispersão com o objetivo de não só visualizar a relação existente entre as variáveis como também a forma deste relacionamento. Nesta etapa, a representação gráfica é muito importante porque, não apenas indica visivelmente se há relação como também a forma sistemática da dispersão dos dados.

Estes diagramas, mesmo sugerindo uma correlação entre as variáveis, não podem confirmar uma suspeita de relação do tipo causa e efeito, a menos que a relação seja sustentada por considerações teóricas ou conhecimento técnico do processo sob estudo.

Doravante, como em [*Augustin, 1981*], as análises serão feitas considerando os dados em três níveis diferentes de profundidade. Isto porque a concentração de sal pode variar de acordo com o nível de profundidade, uma informação muito importante na determinação do padrão de ocorrência de espécimes vegetais. Portanto a estratificação

dos dados por níveis de profundidade foi adotado por razões geológicas/geomorfológicas. Os três níveis considerados são os seguintes:

Nível 1	0 -15 cm profundidade
Nível 2	15-30 cm profundidade
Nível 3	30-45 cm profundidade

Iniciaremos a análise nível a nível.

Nível 1

Estatísticas Descritivas

Variable	Gypsum	Clay	Sand	Silt	PH	EC	Organic
N	50	50	50	50	50	50	50
Mean	11.30	19.90	15.08	65.28	7.6900	1.842	1.0388
Median	13.00	12.10	13.40	72.50	7.7000	2.350	0.9800
StDev	8.08	11.89	10.08	16.57	0.2299	0.950	0.4078
Min	0.06	9.60	0.40	31.00	7.1000	0.150	0.3400
Max	30.20	49.60	45.40	90.00	8.3000	2.930	2.3700
Q1	0.41	10.60	9.40	51.75	7.6000	0.675	0.8200
Q3	17.88	29.85	18.40	79.2	7.8000	2.450	1.1900

As médias das variáveis neste nível são maiores que as médias dos outros dois níveis, isto devido ao efeito do número de amostras.

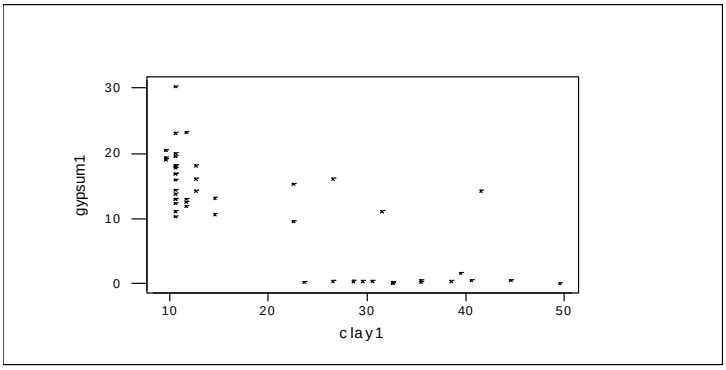


Gráfico de dispersão de Gypsum *versus* Clay

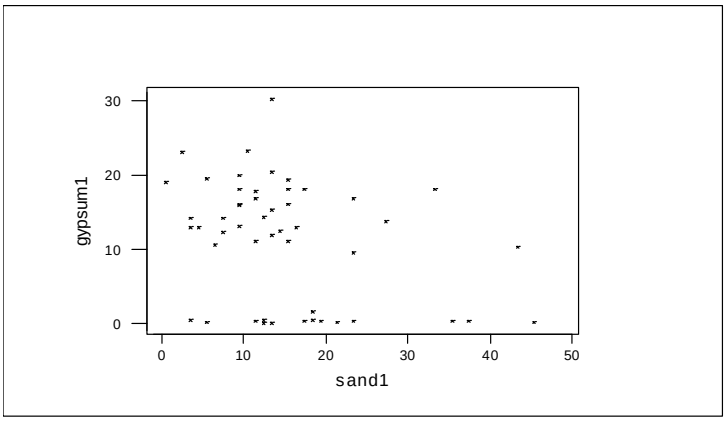


Gráfico de dispersão de Gypsum *versus* Sand

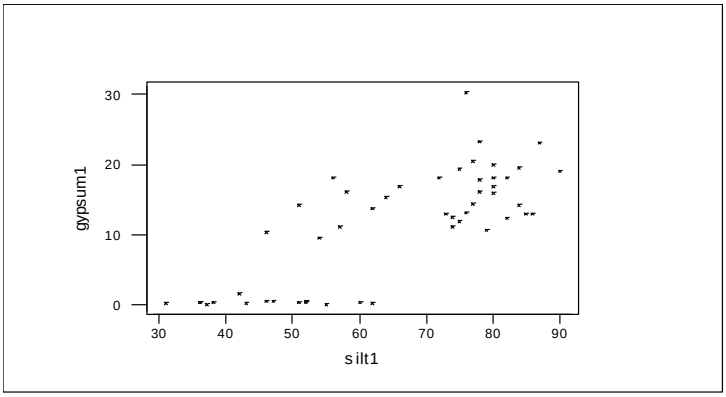


Gráfico de dispersão de Gypsum *versus* Silt

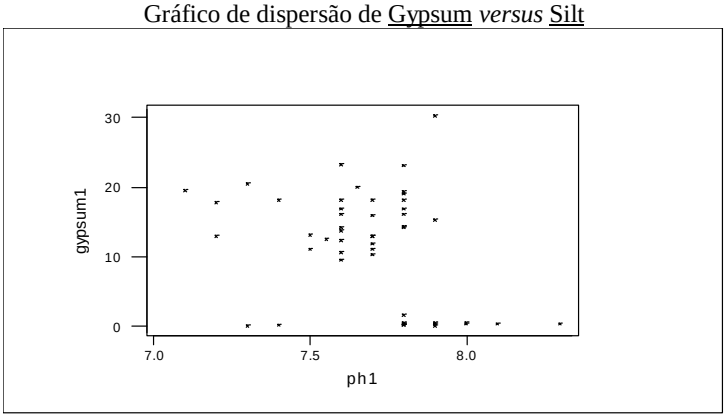


Gráfico de dispersão de Gypsum *versus* PH

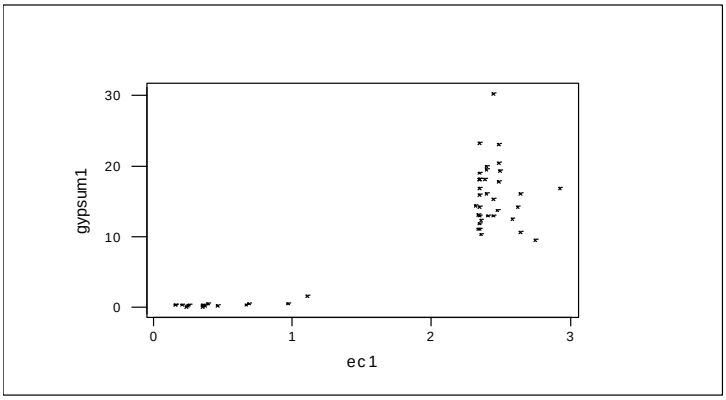


Gráfico de dispersão de Gypsum *versus* EC

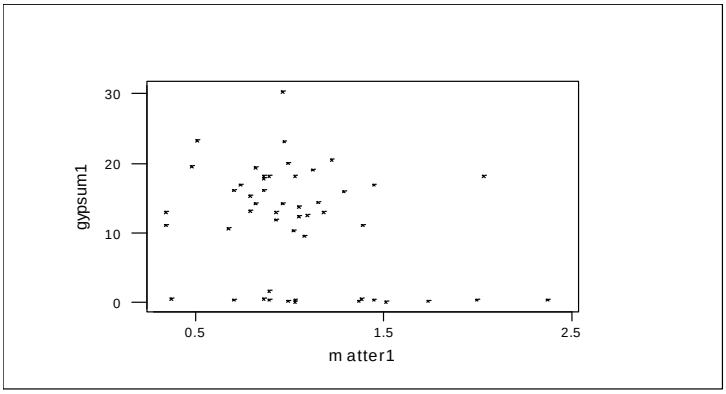


Gráfico de dispersão de Gypsum *versus* Orgmat

Neste nível, todos os requisitos foram amostrados, pois era o nível de profundidade de solo dominante na área estudada, isto devido a facilidade de acesso. O gráfico Gypsum *versus* Clay sugere que porcentagens pequenas de Clay produzem porcentagens maiores de Gypsum, isto já poderia ser esperado pois o sistema é fechado. Esse comportamento ainda pode ser observado para as variáveis Sand e Organic Matter (*Orgamat*) mesmo que simultaneamente também se produza porcentagens menores de Gypsum .

A variável Silt apresenta um comportamento diferente das demais variáveis mencionadas anteriormente, ou seja, altas porcentagens de Silt produzem porcentagens maiores de Gypsum. Desta forma há um forte indício que este **sistema fechado** possa se subdividir em dois subconjuntos, ($\mathbf{V_7}$ e $\mathbf{V_2}$) e ($\mathbf{V_1}$ e $\mathbf{V_3}$), no espaço das variáveis [*Guillaume, 1977*], o que será evidenciado pelas análises feitas no **Apêndice A**.

Até então, os gráficos indicavam uma fraca tendência de formação de dois grupos no espaço de dados, com efeitos diferenciados no Gypsum. Para o PH pode-se notar a tendência de formação de três grupos. Mas coube à variável Electrical conductivity (EC) o gráfico mais significativo da tendência de formação de grupos. Inclusive, posteriormente, esta indicação será incorporada à análise.

Na maioria dos gráficos também observa-se uma faixa de valores na parte inferior dos mesmos. Uma "anomalia" que dificilmente se ajustaria de forma satisfatória a uma regressão linear simples talvez devido a algum procedimento inadequado de

amostragem ou à própria característica do solo na região (heterogeneidade), ou mesmo devido à natureza das variáveis e ao processo de mensuração (sensibilidade na quantificação laboratorial).

Notemos, que em [*Augustin, 1981*], a descrição destes gráficos é feita por comportamentos locais como se as unidades fossem estratificadas em regiões amostrais.

Nível 2

Estatísticas Descritivas

Variable	Gypsum	Clay	Sand	Silt	PH	EC	Organic
N	31	31	31	31	31	31	29
N*	19	19	19	19	19	19	21
Mean	10.28	22.08	13.14	64.58	7.6306	1.873	0.8641
Median	12.60	12.60	12.40	67.00	7.7000	2.360	0.8200
StDev	7.55	14.05	9.58	15.89	0.2227	0.923	0.4761
Min	0.09	0.60	1.40	36.00	7.1000	0.140	0.2200
Max	21.00	44.60	37.40	88.00	8.1000	2.850	2.2500
Q1	0.38	11.60	5.40	50.00	7.5000	0.920	0.5100
Q3	15.60	36.60	17.40	77.00	7.8000	2.450	1.0850

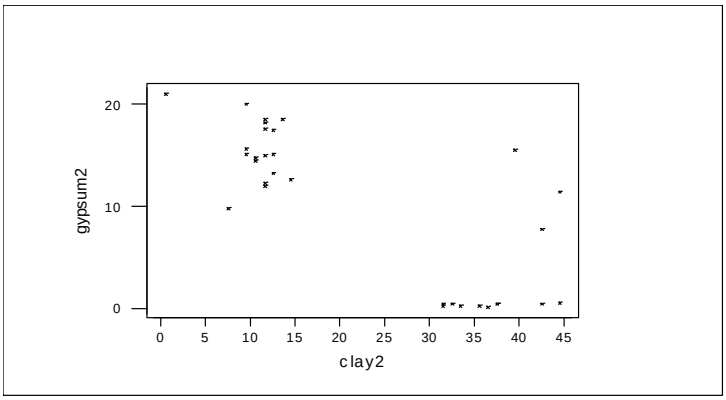


Gráfico de dispersão de Gypsum *versus* Clay

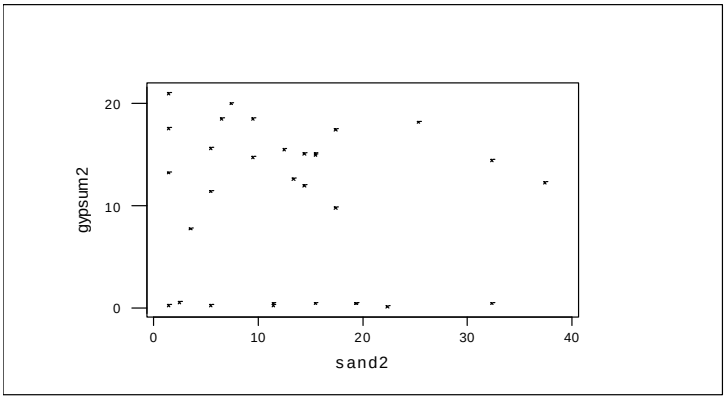


Gráfico de dispersão de Gypsum *versus* Sand

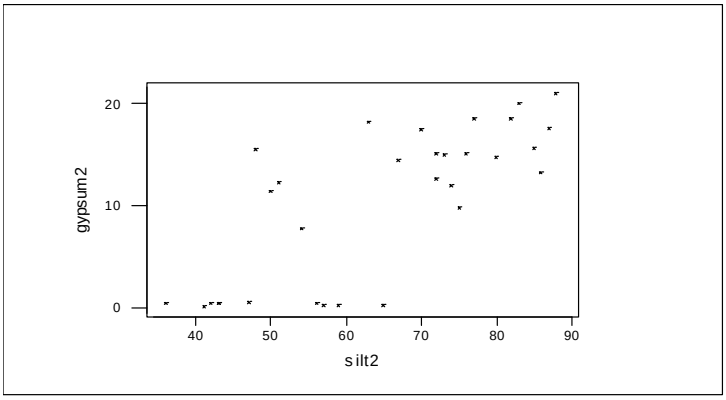


Gráfico de dispersão de Gypsum *versus* Silt

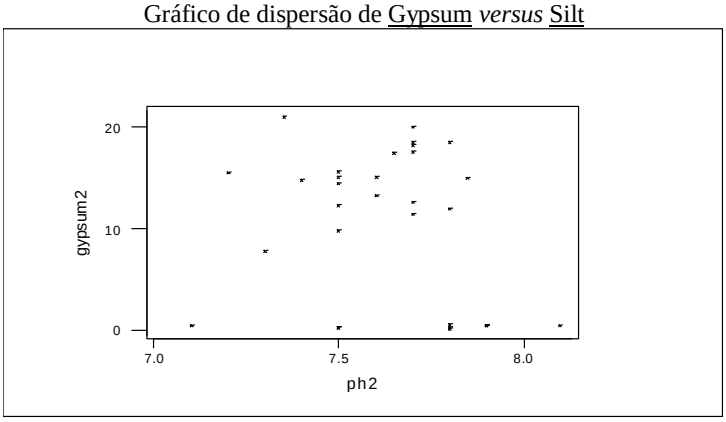


Gráfico de dispersão de Gypsum *versus* PH

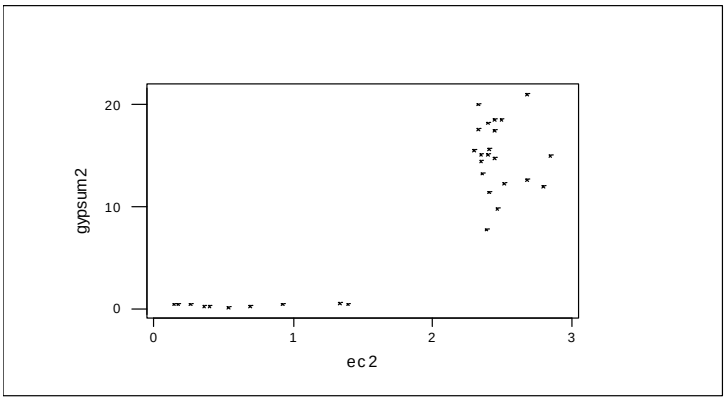


Gráfico de dispersão de Gypsum *versus* EC

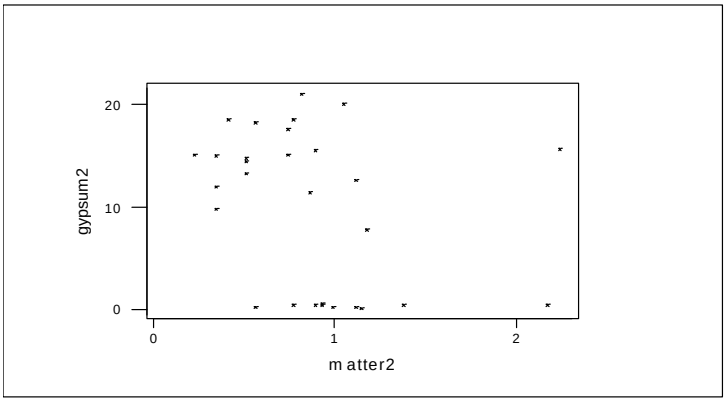


Gráfico de dispersão de Gypsum *versus* Orgmat

A tendência à formação de grupos continua prevalecendo neste nível, exceto pelas variáveis Sand e Orgmat que apresentam um certo padrão de dispersão dos dados com relação ao Gypsum. Neste nível não foram avaliadas todas as unidades amostrais - restrição estabelecida pela própria gênese do problema.

O gráfico Gypsum *versus* Clay sugere mais claramente que porcentagens pequenas de Clay produzem porcentagens maiores de Gypsum. Como no primeiro nível, o Silt continua apresentando a melhor relação com o Gypsum. Resultado que corrobora nossas expectativas baseadas no **nível 1** (sistema fechado). Para a variável PH predomina os valores altos de Gypsum e com a variável EC repetiu-se o comportamento observado no **nível 1**, ou seja, uma forte tendência que nos direciona à formação de grupos.

Nível 3

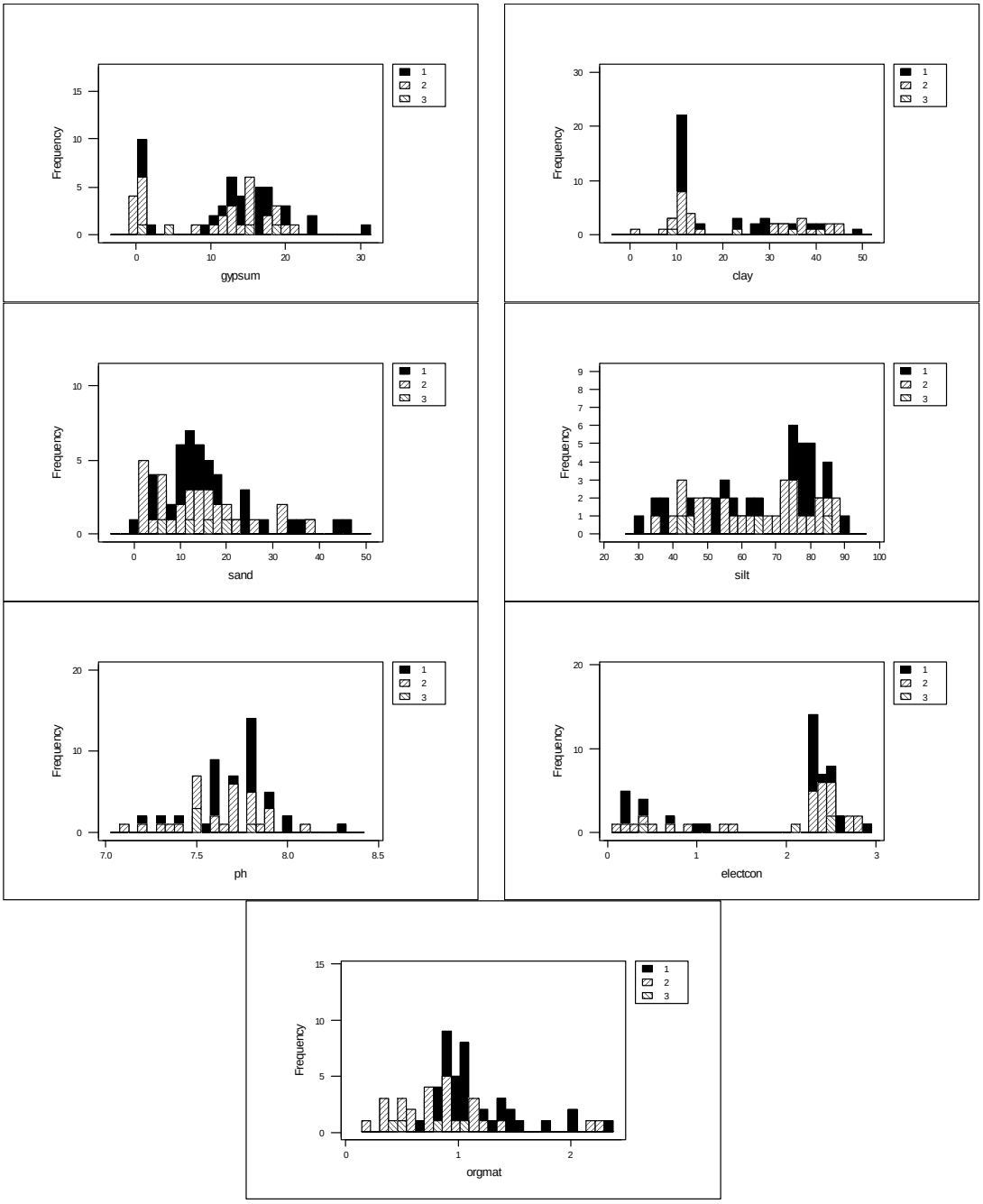
Estatísticas Descritivas

Variable	Gypsum	Clay	Sand	Silt	PH	EC	Organic
N	4	4	4	4	4	4	4
N*	46	46	46	46	46	46	46
Mean	9.89	26.85	13.40	59.75	7.5750	1.837	0.690
Median	9.90	28.60	13.90	55.50	7.5000	2.270	0.650
StDev	8.86	14.48	5.48	19.45	0.1500	1.006	0.299
Min	0.35	8.60	6.40	43.00	7.5000	0.350	0.410
Max	19.40	41.60	19.40	85.00	7.8000	2.460	1.050
Q1	1.44	12.10	7.90	43.75	7.5000	0.785	0.428
Q3	18.33	39.85	18.40	80.00	7.7250	2.457	0.992

Neste nível foram avaliadas somente quatro unidades amostrais. Portanto qualquer análise estatística torna-se não representativa.

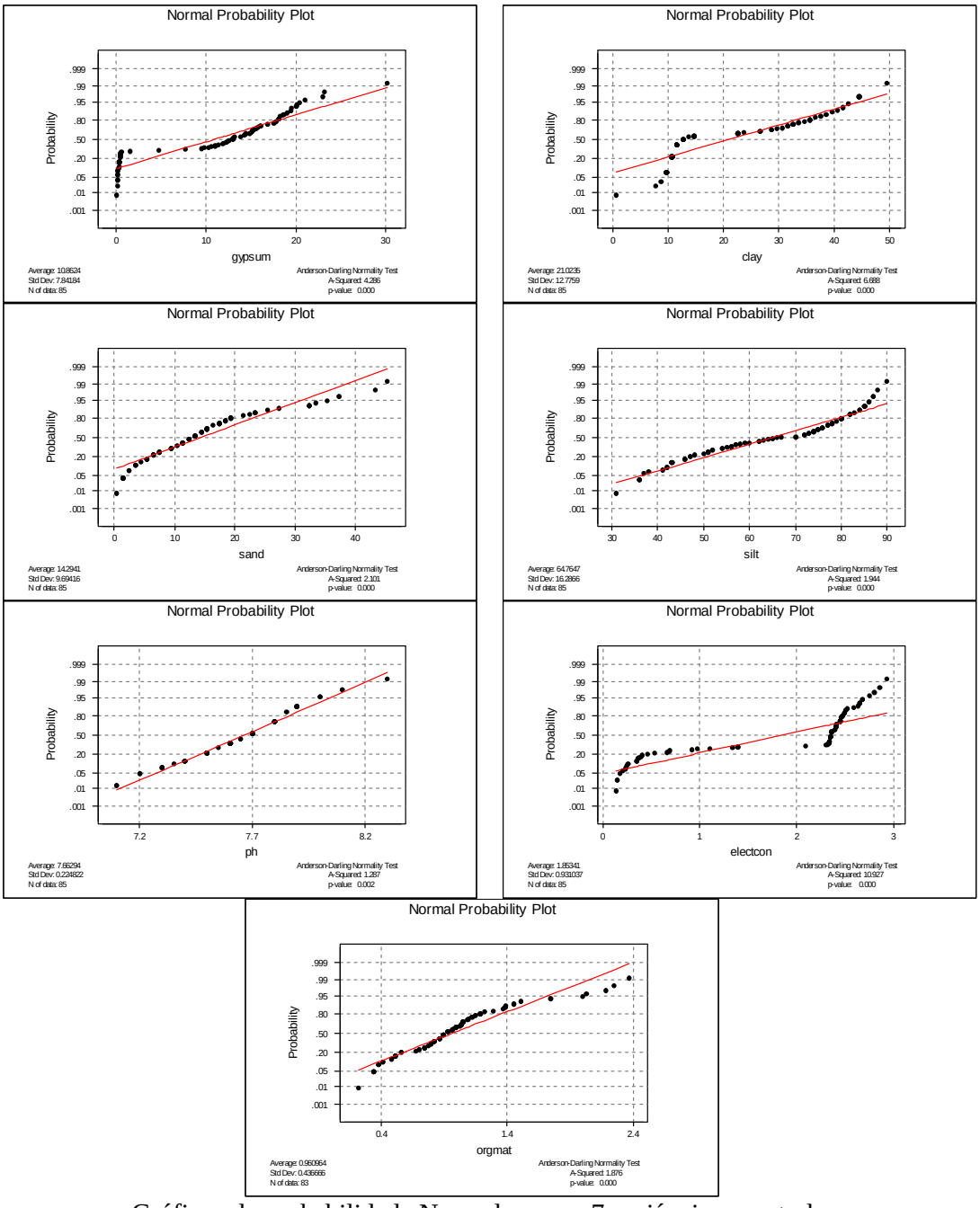
IV. Análise global do padrão descrito pelos dados granulométricos

Nos histogramas abaixo, fez-se uma diferenciação das variáveis em cada nível de profundidade. Notamos então, que não há uma estratificação dos valores para cada nível, ou seja, os dados apresentam um comportamento aleatório (espalhamento uniforme) ao longo do eixo horizontal em todos os níveis de profundidade. Mas estes gráficos servem para se ter uma idéia bem superficial do comportamento dos dados associados a cada uma das variáveis. Para que as suposições feitas nesta análise sejam confirmadas, deveríamos testá-las, mas como os tamanhos de amostras são diferentes, fica mais complexa esta comparação. Por isso não vamos testar, mas analisaremos os dados nos três níveis em que foram coletados (na verdade dois níveis, pois para o terceiro nível tem-se apenas quatro observações) e faremos comentários associando os resultados.



Histogramas das 7 variáveis em estudo identificados por níveis.

Pelos gráficos de probabilidade normal relativo a cada uma das variáveis, que se seguem abaixo, podemos notar que apenas o PH, que é uma variável de característica química do solo, apresenta os pontos mais bem distribuídos ao longo da reta, evidenciado pelo teste de Shapiro-Wilk que não se rejeita a hipótese de normalidade a um nível de significância de até 10%.



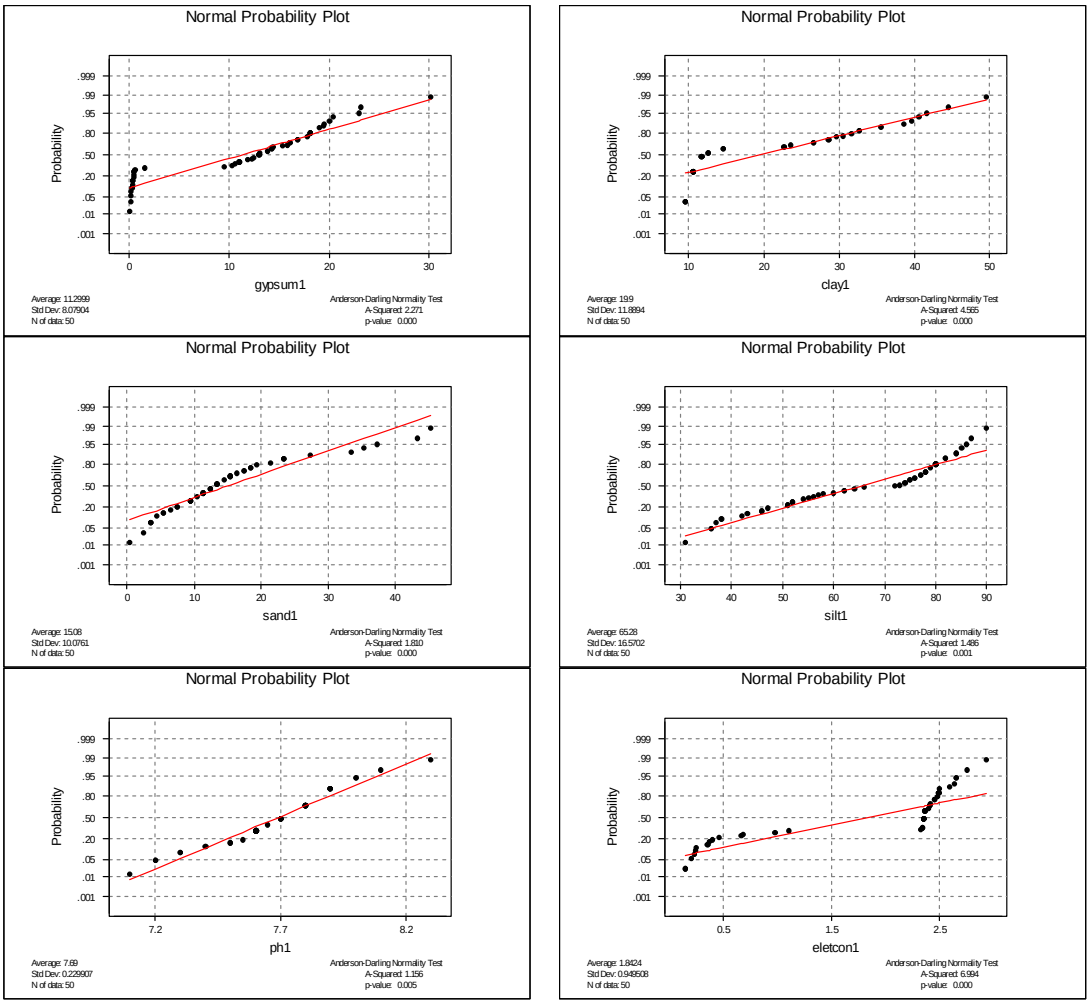
Gráficos de probabilidade Normal para as 7 variáveis em estudo.

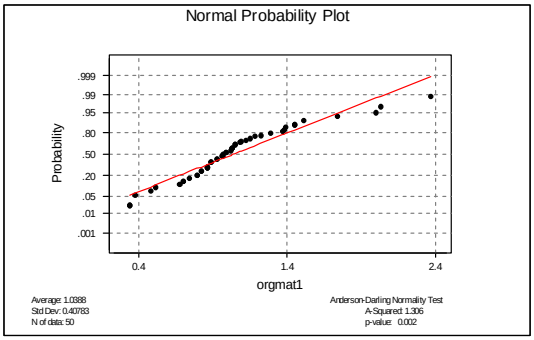
Os outros gráficos não apresentam nenhuma configuração que possa ser considerada de

uma distribuição normal. E relativo aos testes, os p-valores são extremamente baixos, ou seja, há uma forte evidência de que rejeitamos a hipótese de normalidade. Mas isto não impede que tratemos as variáveis em média como normalmente distribuídas, pois sabemos que o **Teorema Central do Limite** nos permite aproximar a média amostral por uma normal, dado que temos uma amostra que pode ser considerada grande (85 dados).

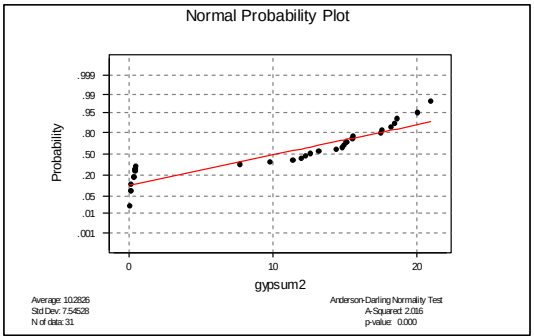
Podemos observar, que mesmo construindo os gráficos de probabilidade normal para os dados estratificados em níveis, em que foram coletados, continuamos com as mesmas características da análise global.

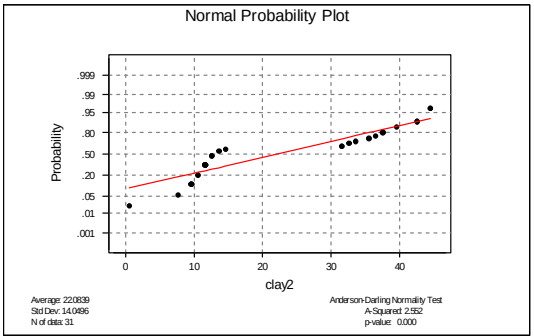
Nível 1

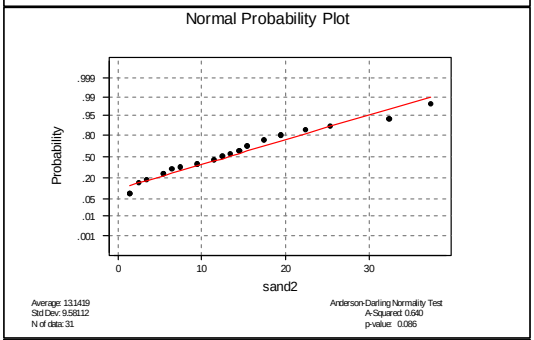


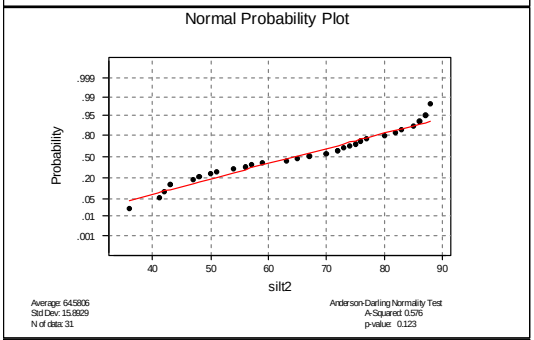

Normal Probability Plot
Probability
.999
.99
.95
.80
.50
.20
.05
.01
.001
0.4
1.4
2.4
orgmat1
Average: 1.0388
Std Dev: 0.40783
N of data: 50
Anderson-Darling Normality Test
A-Squared: 1.306
p-value: 0.002

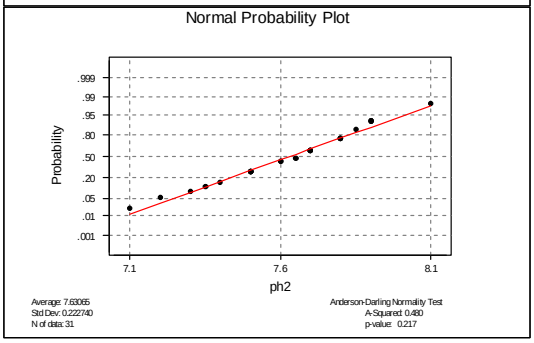
Nível 2

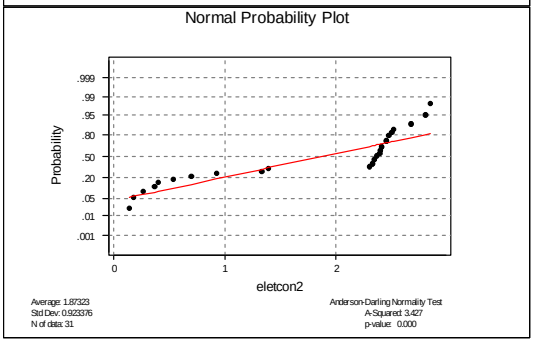

Normal Probability Plot
Probability
gypsum2
Average: 10.2826
Std Dev: 7.54528
N of data: 31
Anderson-Darling Normality Test
A-Squared: 2.016
p-value: 0.000

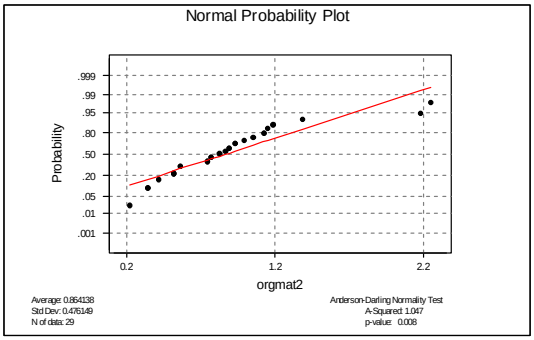

Normal Probability Plot
Probability
clay2
Average: 22.0839
Std Dev: 14.0496
N of data: 31
Anderson-Darling Normality Test
A-Squared: 2.552
p-value: 0.000


Normal Probability Plot
Probability
sand2
Average: 13.1419
Std Dev: 9.58112
N of data: 31
Anderson-Darling Normality Test
A-Squared: 0.640
p-value: 0.086


Normal Probability Plot
Probability
silt2
Average: 64.5806
Std Dev: 15.8629
N of data: 31
Anderson-Darling Normality Test
A-Squared: 0.576
p-value: 0.123


Normal Probability Plot
Probability
ph2
Average: 7.63065
Std Dev: 0.222740
N of data: 31
Anderson-Darling Normality Test
A-Squared: 0.480
p-value: 0.217


Normal Probability Plot
Probability
eletcon2
Average: 1.87323
Std Dev: 0.923376
N of data: 31
Anderson-Darling Normality Test
A-Squared: 3.427
p-value: 0.000



Para sabermos qual o grau de associação entre as variáveis, calculamos o coeficiente de correlação. O coeficiente de correlação é uma medida do grau de linearidade entre quaisquer duas variáveis. Valores próximos de +1 ou -1 indicam um alto grau de linearidade, enquanto valores próximos de zero indicam ausência de colinearidade. Valores positivos mostram que o valor de uma variável tende a crescer com o crescimento do valor da outra, enquanto valores negativos mostram que uma tende a decrescer com valores crescentes da outra.

Primeiramente fizemos as análises de correlação entre as variáveis sem a diferenciação por níveis de profundidade.

Correlations (Pearson)

	Clay	Sand	Silt	PH	EC	Orgmat
Gypsum	-0.782	-0.254	0.772	-0.269	0.872	-0.289
Clay		0.037	-0.817	0.220	-0.716	0.240
Sand			-0.595	0.193	-0.251	0.191
Silt				-0.286	0.717	-0.283
PH					-0.329	0.151
EC						-0.328

A variável Gypsum apresenta maior índice de correlação linear com Clay (negativamente), Silt (positivamente) e Electrical Condutivity (positivamente). A variável Clay está correlacionada negativamente com a variável Silt e também com Eletctrical condutivity. A variável Sand está correlacionada negativamente com Silt, mas numa ordem de grandeza inferior. E a variável Silt está correlacionada positivamente com a variável Electrical condutivity. Isto evidencia a formação de dois grupos com referência ao **sistema fechado**.

Nível 1

Correlations (Pearson)

	Clay1	Sand1	Silt1	PH1	EC1	Orgmat1
Gypsum1	-0.785	-0.331	0.777	-0.352	0.874	-0.288
Clay1		0.104	-0.793	0.369	-0.762	0.161
Sand1			-0.682	0.211	-0.333	0.318
Silt1				-0.397	0.760	-0.276
PH1					-0.425	0.283
EC1						-0.355

Neste nível as variáveis se comportam aproximadamente seguindo um mesmo padrão do caráter global, apenas a correlação entre a variável Sand e Silt apresenta um ligeiro aumento.

Nível 2

Correlations (Pearson)

	Clay2	Sand2	Silt2	PH2	EC2	Orgmat2
Gypsum2	-0.787	-0.113	0.761	-0.242	0.884	-0.319
Clay2		-0.081	-0.840	0.102	-0.669	0.401
Sand2			-0.449	0.141	-0.100	-0.108
Silt2				-0.170	0.652	-0.312
PH2					-0.227	-0.089
EC2						-0.279

As observações são análogas às mencionadas anteriormente, ressaltando que a correlação entre as variáveis Sand e Silt está menos evidenciada que anteriormente. Podemos assim assinalar que o **nível 1** contribui significativamente para tal relação global.

V. Modelagem

A partir de agora estaremos à procura de um modelo para os dados, tomando o Gypsum como variável resposta. Os modelos serão analisados nos níveis 1 e 2 separadamente. Não modelaremos os dados no nível 3 por não haver um número suficiente de observações que possa permitir sua análise através de métodos da estatística clássica.

Nível 1

Primeiramente ajustamos um modelo de regressão simples, onde consideramos os dados como normalmente distribuídos, da mesma forma como sugerido na dissertação [*Augustin, 1981*].

Construímos abaixo uma tabela com os modelos de regressão simples para cada variável, e seus respectivos valores de R^2 [*Dobson, 1990*]. Estes valores de R^2, podem

ser interpretados como a proporção da variação total dos dados que é explicada tomando como base o modelo.

Modelo	R^2
21,9 - 0,534 Clay	61,7%
15,3 - 0,265 Sand	10,9%
-13,4 + 0,379 Silt	60,4%
106,0 – 12,4 PH	12,4%
-2,4 + 7,43 EC	76,3%
17,2 – 5,71 Orgmat	8,3%

Observando os valores de R^2 para os modelos ajustados, percebemos que não há um bom ajuste da variável resposta Gypsum com as variáveis Sand, PH e Orgmat. Os melhores ajustes são com relação às variáveis EC, Clay e Silt, respectivamente.

Agora podemos confirmar através da Análise de Variância, se a variabilidade dos dados que os modelos explicam, são significativas.

Análise de variância

Modelo (simples)	G.L	SQ	SQM	F
+EC	1	2441.7	2441.7	154.92
Resíduos	48	756.5	15.8	
+Clay	1	1973.1	1973.1	77.30
Resíduos	48	1225.2	25.5	
+Silt	1	1931.3	1931.3	73.17
Resíduos	48	1267.0	26.4	
+PH	1	395.88	395.88	6.78
Resíduos	48	2802.39	2802.39	
+Sand	1	350.17	350.17	5.90
Resíduos	48	2848.10	59.34	
+Orgmat	1	265.71	265.71	4.35
Resíduos	48	2932.57	61.10	

Comparando com uma $F_{0,05}(1,48) \approx 4{,}00$, concluímos que em todos os modelos simples rejeitamos a hipótese de que o coeficiente de uma das variáveis explicativas seja igual a zero.

No quadro abaixo, analisaremos o processo de exclusão-inclusão de variáveis, técnica do “stepwise”, onde verificaremos quais variáveis compõem o melhor ajuste associado a estes dados.

Stepwise Regression

F-to-Enter: 4.00 F-to-Remove: 4.00

Response is gypsum1 on 6 predictors, with N = 50

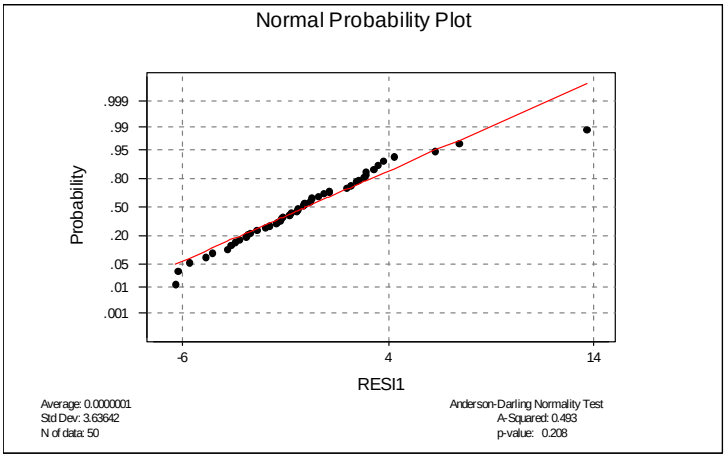
Step	1(UM)	2(DOIS)
Constant	-2.397	4.858
EC1 T-Ratio	7.43 12.45	5.59 6.47
Clay1 T-Ratio		-0.194 -2.81
S	3.97	3.71
R-Sq	76.35	79.74

Pelos resultados do "*stepwise*", concluímos que o melhor modelo para os dados considerados como normalmente distribuídos é :

Gypsum = 4,858 + 5,59EC - 0,194Clay

Podemos observar que a variável Electrical condutivity é a que mais influencia na variável resposta (*Gypsum*), o aumento de uma unidade nesta variável acarreta num aumento de 5,59 unidades na variável resposta. A variável Clay se comporta no sentido inverso, um aumento de uma unidade nesta variável acarreta numa diminuição de 0,194 unidades na variável resposta.

Gráfico de Probabilidade Normal para os resíduos do modelo final



No gráfico acima, podemos perceber que o resíduo do modelo pode ser considerado normal, o que nos leva a concluir que a suposição feita de que o componente aleatório do modelo final tenha distribuição normal, não foi violada.

Dando margens ao nosso caráter intuitivo e levando-se em consideração as informações sobre a natureza física das variáveis, vamos nos direcionar sob a tentativa da determinação de um modelo sem o intercepto:

Stepwise Regression

F-to-Enter: 4.00 F-to-Remove: 4.00

Response is gypsum1 on 6 predictors, with N = 50

Step **No constant**	**1(UM)**	**2(DOIS)**	**3(TRÊS)**
EC1 T-Ratio	6.40 22.95	6.97 21.71	5.04 5.58
Clay1 T-Ratio		-0.086 -2.98	-0.145 -3.82
Silt1 T-Ratio			0.075 2.27
S	4.08	3.79	3.63

Assim, o melhor modelo para os dados sem ajuste do intercepto é :

Gypsum = 5,04EC - 0,145Clay + 0,075Silt

Note que entrou mais uma variável no modelo, porém com um coeficiente muito baixo.

Nível 2

Uma análise análoga àquela feita para o **nível 1,** faremos para o **nível 2**, continuando a considerar as variáveis como normalmente distribuídas. Abaixo apresentamos os modelos obtidos com a regressão simples.

Modelo	**R_2**
19,6 - 0,423 Clay	61,9%
11,4 - 0,089 Sand	1,3%

-13,1 + 0,362 Silt	58,0%
72,8 - 8,19 PH	5,8%
-3,24 + 7,22 EC	78,1%
14,4 - 5,15 Orgmat	10,2%

Na tabela acima podemos observar que os piores ajustes da variável resposta Gypsum são com relação às variáveis Sand, PH e Orgmat. Já os melhores ajustes são com as variáveis Electrical condutivity, Clay e Silt. O mesmo foi observado no **nível 1**, sendo que o R^2 do ajuste com a variável Sand é bastante inferior neste nível.

Análise de variância

Modelo (simples)	**G.L**	**SQ**	**SQM**	**F**
+EC	1	1333.2	1333.2	103.19
Resíduos	29	374.7	12.9	
+Clay	1	1057.36	1057.3	47.13
Resíduos	29	50.6	22.4	
+Silt	1	990.32	990.32	40.02
Resíduos	29	717.62	24.75	
+Orgmat	1	168.09	168.09	3.06
Resíduos	27	1480.75	54.84	
+PH	1	99.81	99.81	1.80
Resíduos	29	1608.13	55.45	
+Sand	1	21.65	21.65	0.37
Resíduos	29	1686.29	58.15	

Comparando com uma $F_{0,05}(1,29) \approx 4,18$, temos que os únicos modelos, onde não se rejeita a hipótese de que o coeficiente da variável explicativa seja igual a zero são os do ajuste de Gypsum com Orgmat ou com PH ou com Sand. Sendo que o ajuste feito com as variáveis Electrical condutivity ou Clay ou Silt são significativos, ou seja, explicam uma boa parte da variabilidade dos dados.

No quadro abaixo, analisaremos o "*stepwise*", onde verificaremos qual a combinação de variáveis compõe o melhor ajuste para os dados em estudo.

Stepwise Regression

F-to-Enter: 4.00 F-to-Remove: 4.00

Response is gypsum2 on 6 predictors, with N = 29
N(cases with missing obs.) = 21 N(all cases) = 50

Step	**1(UM)**	**2(DOIS)**
Constant	-3.251	4.623
EC2	7.22	5.30
T-Ratio	9.86	6.56
Clay2		-0.192
T-Ratio		-3.60

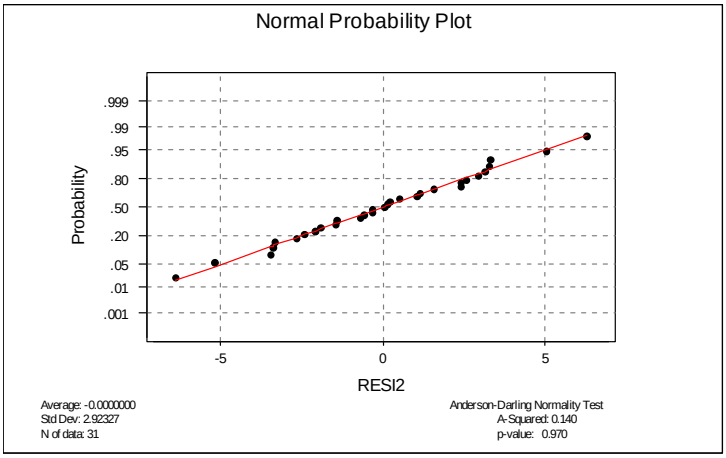
S	3.64	3.03
R-Sq	78.27	85.49

Pelos resultados do *stepwise*, concluímos que o melhor modelo para os dados considerados como normalmente distribuídos é :

Gypsum = 4,623 + 5,30EC - 0,192Clay

Observamos que a variável Electrical condutivity é a que mais influencia na variável resposta Gypsum, o aumento de uma unidade nesta variável acarreta num aumento de 5,30 unidades na variável resposta. A variável Clay se comporta de forma inversa, um aumento de uma unidade nesta variável acarreta numa diminuição de 0,192 unidades na variável resposta. Note que este resultado é análogo ao obtido para o **nível 1**.

Gráfico de Probabilidade Normal para os resíduos do modelo final



Observando o gráfico acima, obtemos a mesma conclusão já referenciada no **nível 1**, a suposição de normalidade do resíduo não foi violada.

Pelo mesmo sentimento expresso com relação ao **nível 1**, vamos buscar ajustar um modelo sem o intercepto.

Stepwise Regression

F-to-Enter: 4.00 F-to-Remove: 4.00

Response is gypsum2 on 6 predictors, with N = 29
N(cases with missing obs.) = 21 N(all cases) = 50

Step No constant	1(UM)	2(DOIS)
EC2 T-Ratio	5.80 16.54	6.65 18.48
clay2 T-Ratio		-0.107 -3.88
S	3.88	3.16

O melhor modelo sem o intercepto é:

Gypsum = 6,65EC - 0,107Clay

Observamos que o modelo é composto pelas mesmas variáveis do modelo com intercepto, porém o coeficiente da variável Electrical condutivity aumentou e o da variável Clay diminuiu.

Na tese da professora Cristina Augustin [*Augustin, 1981*], as suposições básicas da regressão múltipla e a adequacidade do modelo foram avaliadas. São elas:

1- os resíduos são distribuídos aleatoriamente,
2- a distribuição dos resíduos é normal,
3- ausência de correlação significativa entre as variáveis independentes,
4- ajuste satisfatório da regressão.

Que nós aqui avaliaremos com base nos dados em estudo.

A suposição de aleatoriedade foi testada construindo-se gráficos bidimensionais dos resíduos *versus* as variáveis independentes. Nestes gráficos costuma-se detectar tipos mais comuns de modelos inadequados. A segunda suposição foi testada através do gráfico de probabilidade normal dos resíduos onde um teste similar ao de Shapiro-Wilk foi realizado.

O terceiro item foi testado pela estatística de Durbin-Watson e finalmente, o ajuste da regressão pela construção da técnica de ANOVA.

Uma vez que os gráficos de resíduos foram insatisfatórios pois revelavam falta de aleatoriedade, os dados foram separados em dois grupos com base nos altos valores (2mmhos/cm) e baixos (1.5mmhos/cm) da variável Electrical Conductivity. Esta variável foi usada como referência para a separação dos dados porque foi a que mais contribuiu para a variabilidade da regressão. Além disso, tanto os gráficos de dispersão quanto os de resíduos evidenciaram claramente o efeito de agrupamento que também estava presente nos gráficos de resíduos de outras variáveis.

A partir destas mudanças alguns procedimentos usados até então foram repetidos pois nesta etapa do trabalho começamos a priorizar o item 2 acima citado, já que o objetivo

era incorporar outras técnicas à análise estatística feita em [*Augustin, 1981*].

Neste momento, dar-se-á enfoque à questão sobre a suposição de normalidade dos resíduos e até que ponto isto poderia afetar a determinação e o ajuste do modelo de regressão.

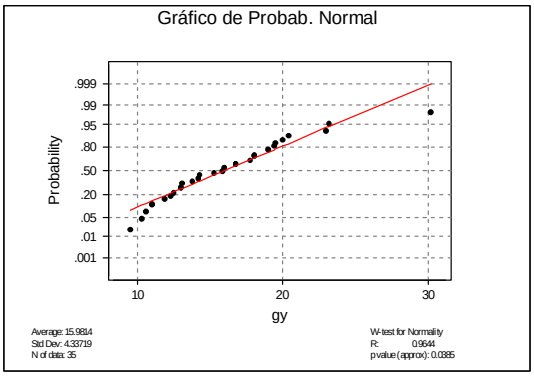
Nível 1 e Categoria 1
(Alta condutividade)

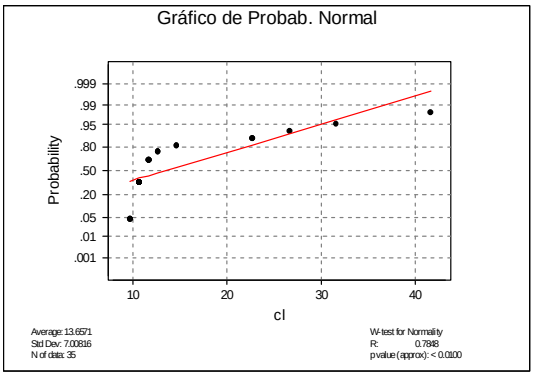
A correlação das variáveis com Gypsum é muito baixa. As maiores correlações são com respeito a Clay (negativamente) e Silt (positivamente), sendo que entre estas duas variáveis ocorre o maior valor, em módulo, da correlação. As outras correlações entre as variáveis são relativamente pequenas.

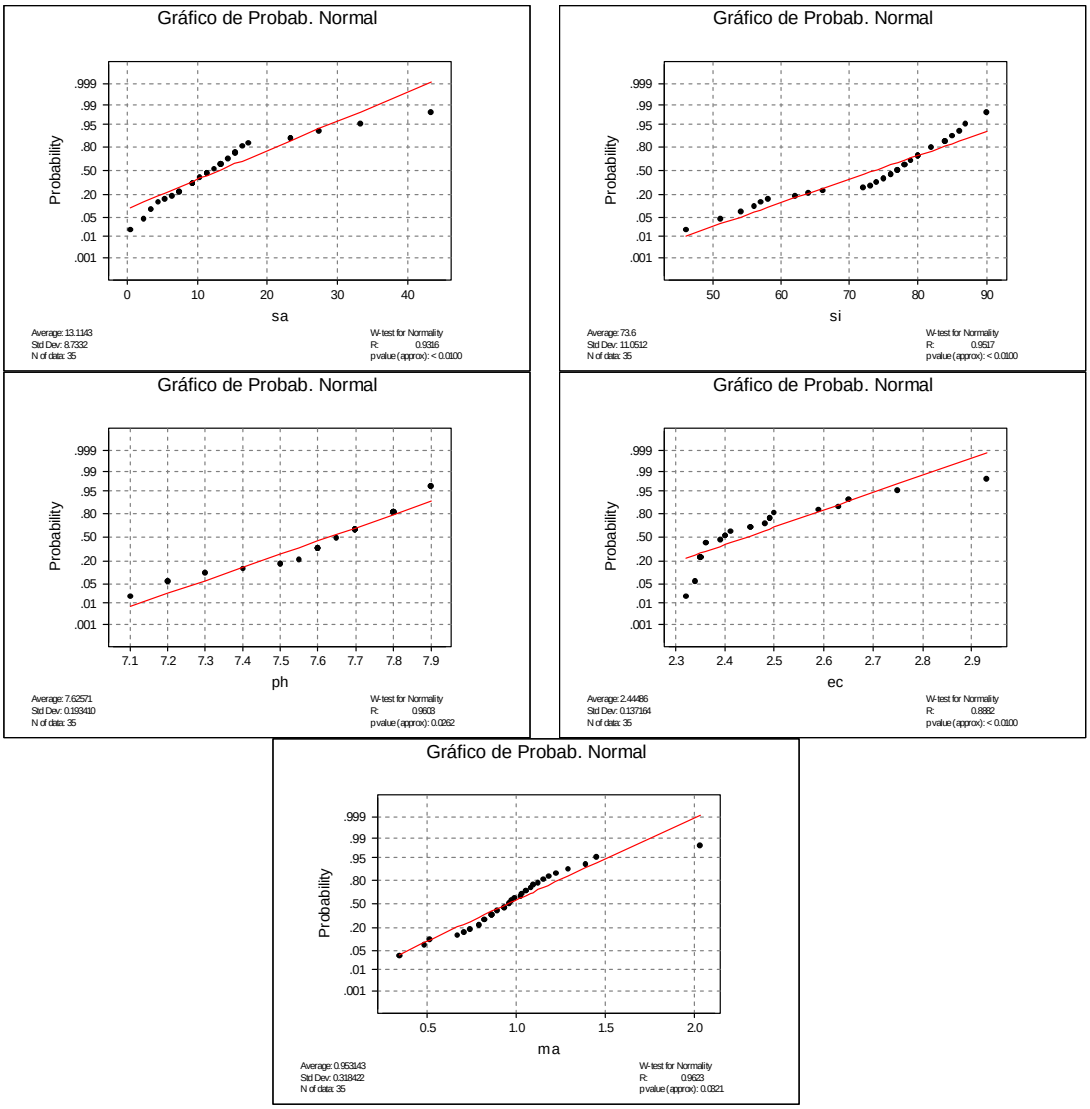
Correlations (Pearson)

	Clay	Sand	Silt	PH	EC	MO
Gypsum	-0.283	-0.191	0.342	0.102	-0.079	0.017
Clay		-0.055	-0.606	0.252	-0.039	0.061
Sand			-0.746	0.101	-0.051	0.049
Silt				-0.237	0.072	0.023
PH					0.008	0.173
EC						0.084

Os gráficos abaixo avaliam a suposição de normalidade dos dados. Visualmente quando os pontos estão localizados, de forma aproximada, ao longo de uma reta isto indica que a amostra é proveniente de uma população normalmente distribuída. Para confirmar esta suposição um teste similar ao de Shapiro-Wilk é feito, onde rejeitamos a hipótese de normalidade. Assim mesmo, optamos por seguir a forma apresentada em [*Augustin, 1981*], para podermos, mais adiante, comparar com a nossa análise.







Gráficos de Probabilidade Normal para as 7 variáveis categorizadas (alta condutividade)

Stepwise Regression

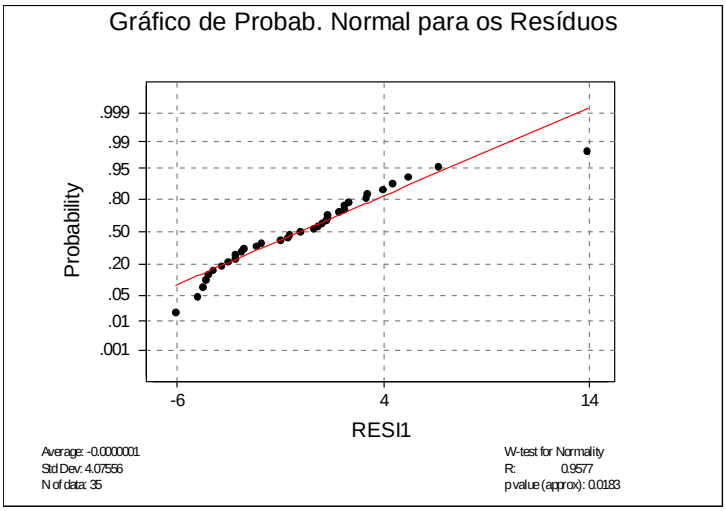
F-to-Enter: 4.00 F-to-Remove: 4.00
Response is gy on 6 predictors, with N = 35

Step	1(UM)
Constant	6.101
Silte	0.134
T-Ratio	2.09
S	4.14
R-Sq	11.70

Pela regressão *stepwise*, concluímos que o melhor modelo para os dados considerando que o componente aleatório seja normal é:

Gypsum = 6.10 + 0.134Silte

Em [*Augustin, 1981*], foram feitos gráficos de resíduos onde novamente os efeitos de agrupamento apareceram. Como nosso objetivo é outro, preferimos os gráficos de probabilidade normal aos gráficos de resíduos. A normalidade assumida pelo componente aleatório não é confirmada por este modelo.

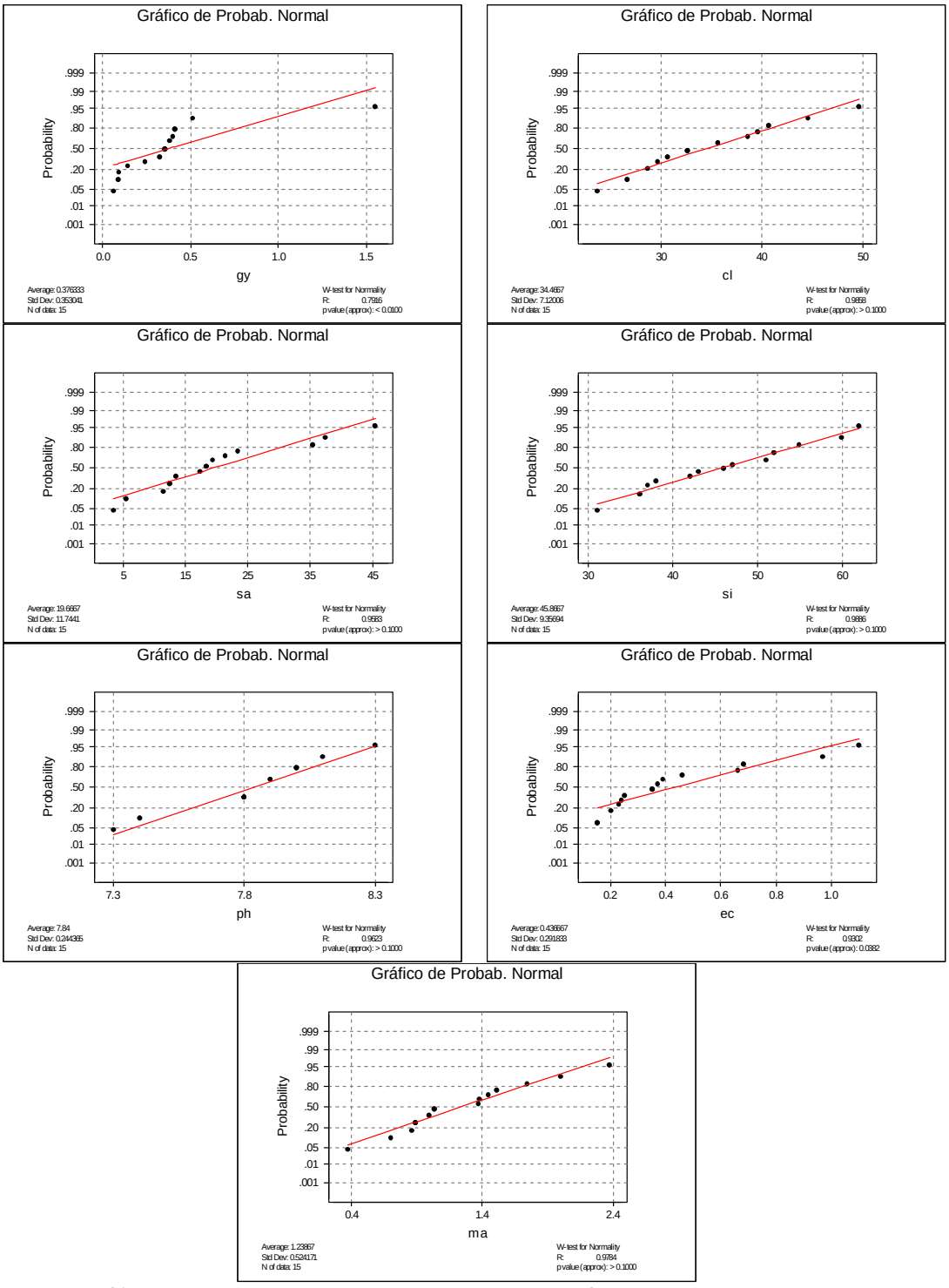


Nível 1 Categoria 2
(Baixa condutividade)

As correlações entre as variáveis explicativas e Gypsum continuam baixas, exceto a variável EC que inclusive possui o maior valor em módulo. Nesta nova situação o que se observa é um aumento nos valores das correlações que antes eram baixas e um decréscimo onde antes eram altas. Isto pode justificar a formação de grupos quando analisamos os dados integralmente.

Correlations (Pearson)

	Clay	Sand	Silt	PH	EC	MO
Gypsum	0.139	-0.011	-0.092	0.161	0.682	-0.212
Clay		-0.604	-0.002	-0.370	0.629	-0.552
Sand			-0.795	0.086	-0.329	0.460
Silt				0.174	-0.065	-0.158
PH					-0.038	0.164
EC						-0.393



Gráficos de Probabilidade Normal para as 7 variáveis categorizadas (baixa condutividade)

Todos os gráficos pelo teste aceitaram a suposição de normalidade das variáveis exceto para as variáveis Gypsum e Electrical Condutivity.

Stepwise Regression

F-to-Enter: 4.00 F-to-Remove: 4.00
Response is gy on 6 predictors, with N = 15

Step	1(UM)	2(DOIS)
Constant	0.01588	0.67766
EC	0.83	1.19
T-Ratio	3.37	4.22
Cl		-0.024
T-Ratio		-2.06
S	0.268	0.240
R-Sq	46.56	60.52

Pela regressão *stepwise*, concluímos que o melhor modelo para os dados considerando o componente aleatório normal é:

Gypsum = 0.678 + 1.19 EC - 0.0238 Clay

Para este modelo o componente aleatório é normalmente distribuído. Nesta situação os dados se ajustaram melhor ao modelo.

Na tese [*Augustin, 1981*], a professora Cristina Augustim chama a atenção para o fato de que possivelmente outra variável que não tenha sido medida também seja a responsável pelo agrupamento dos dados mesmo categorizando-os segundo a variável Electrical Condutivity.

Nível 2 Categoria 1
(Alta condutividade)

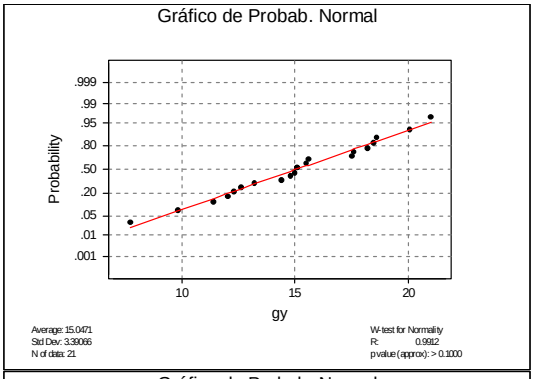
O mesmo procedimento realizado anteriormente para o **nível 1** foi repetido para o **nível 2**.

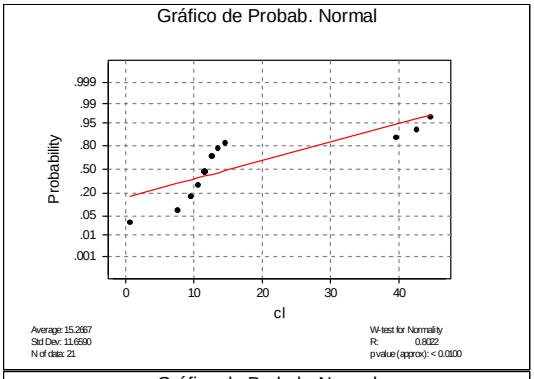
A variável resposta está bem correlacionada com Clay (negativamente) e Silt (positivamente), sendo que a correlação entre as duas variáveis é a maior em módulo. Isto também foi observado para o **nível 1** nestas mesmas condições.

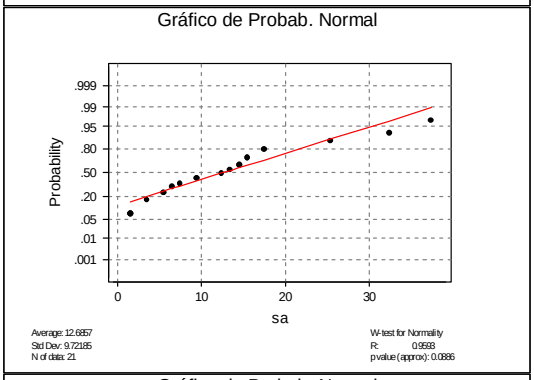
Correlations (Pearson)

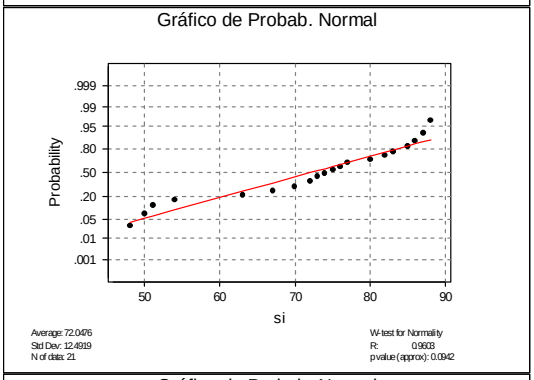
	Clay	Sand	Silt	PH	EC	MO
Gypsum	-0.493	-0.159	0.506	0.204	-0.072	0.022
Clay		-0.191	-0.750	-0.292	-0.290	0.207

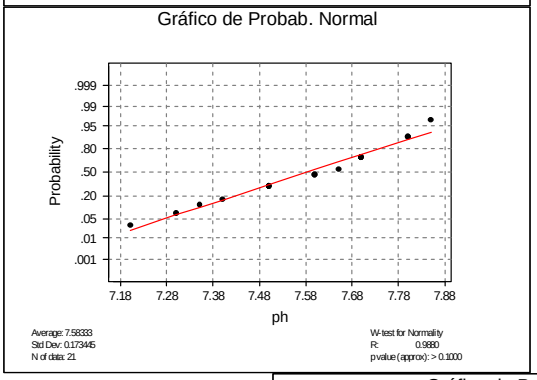
Sand			-0.472	0.025	0.071	-0.356
Silt				0.288	0.130	0.001
PH					0.396	-0.264
EC						-0.204

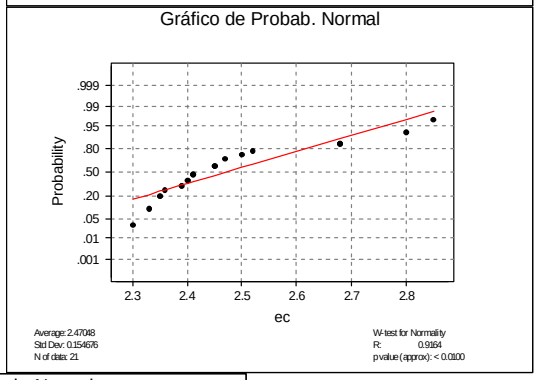


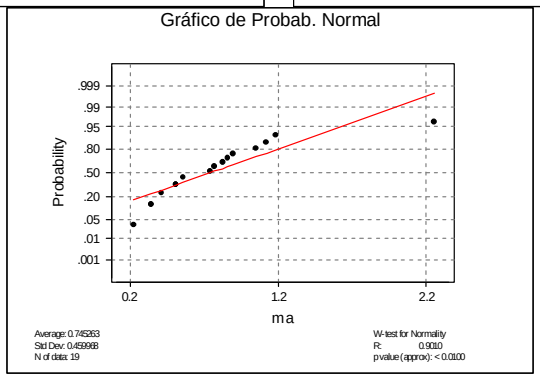












Gráficos de Probabilidade Normal para as 7 variáveis categorizadas (alta condutividade)

Pelo teste as variáveis Gypsum, Sand, PH e Silt podem ser consideradas normalmente distribuídas uma vez que as probabilidades de significância foram altas. Já o teste para as variáveis Clay e EC rejeitou a hipótese de normalidade. Este resultado difere bastante do obtido no **nível 1**.

Stepwise Regression

F-to-Enter: 4.00 F-to-Remove: 4.00

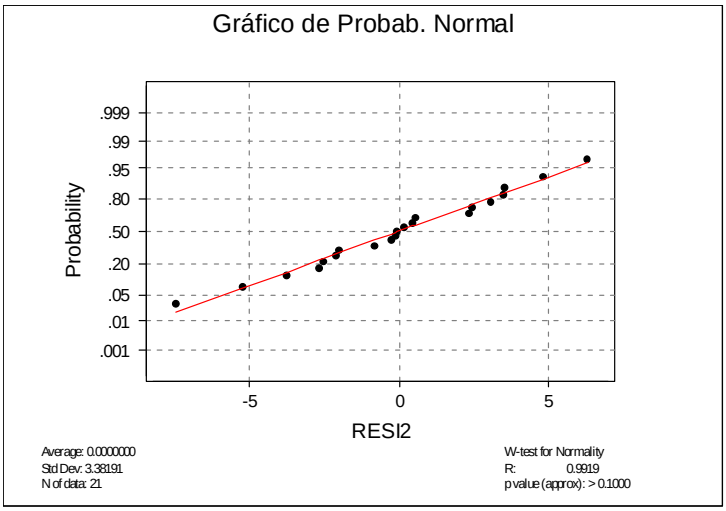
Response is gy on 6 predictors, with N = 19
N(cases with missing obs.) = 21 N(all cases) = 40

Step	1(UM)
Constant	17.34
Cl T-Ratio	-0.146 -2.48
S R-Sq	3.06 26.58

Pela regressão *stepwise*, concluímos que o melhor modelo para os dados considerando o componente aleatório normal é:

Gypsum = 17.34 - 0.146 Clay

O componente aleatório deste modelo pode ser considerado normalmente distribuído:



Nível 2 Categoria 2
(Baixa condutividade)

As análises nestas condições ficaram prejudicadas pelo excesso de valores omissos e

pela inadequacidade do procedimento stepwise. Na construção do modelo nenhuma variável permaneceu no modelo. Este fato estatísticamente não tem fundamento.

VI. Uma análise utilizando o GLIM

Agora iremos tentar atribuir às variáveis uma distribuição que melhor se adapte à sua natureza. Como sabemos algumas das variáveis, envolvidas neste problema, são estocásticas, no sentido de que as quantidades das variáveis sejam proporcionais à porção de solo estudada. Talvez a consideração de uma distribuição estocástica do tipo **Gama** se faça aqui mais adequada.

Para que fique mais claro, podemos usar uma interessante relação entre a distribuição **Gama** e a distribuição de **Poisson** [*Dobson, 1990*]**.** Sabemos que a distribuição **Gama** é descrita por meio de dois parâmetros, r e α, dos quais se exige que $r>0$ e $\alpha>0$. Para evidenciar esta relação, r deve ser um inteiro. Quando tratamos de uma **Poisson**, no nosso caso, estamos essencialmente interessados no número de ocorrências de um evento numa determinada área ou volume. E a distribuição **Gama** surge quando indagamos quanto a distribuição da área necessária para obter um número específico de ocorrências do evento. Desta forma, justificado pela natureza das variáveis em estudo, na tentativa de associar a granulometria a variáveis físico-químicas do solo, analisado através de volumes pré-estabelecidos, indicamos neste trabalho uma análise alternativa baseada na distribuição **Gama** para modelagem.

Para determinarmos um modelo no qual as variáveis são consideradas bem ajustadas, foi utilizada uma técnica chamada "*backward*". Esta técnica consiste de passos que testam a retirada de variáveis do modelo proposto sob H_0. Começamos com o modelo completo, com toda as variáveis do estudo, e vamos retirando variáveis segundo critérios pré-determinados (nível de significância).

Pela justificativa dada acima, consideramos as variáveis como tendo uma distribuição **Gama**, porém o teste que usamos para testar os modelos no "b*ackward*", é o teste *F*, que por sua vez é construído supondo normalidade das variáveis. Para que este teste seja viável vamos usar a seguinte justificativa [*Jϕrgensen*, 1992]:

- Seja uma variável aleatória pertencente à classe das famílias exponenciais com dispersão $X \in ED(\mu, \sigma^2)$, contínua com função de densidade de probabilidade dada por:

$$p(y;\theta,\lambda) = a(\lambda, y)\exp\left[\lambda\{\theta y - \kappa(\theta)\}\right], y \in \Re$$

onde $\mu=\tau(\theta)$ e $\sigma^2= 1/\lambda$. Existe um teorema que afirma que se $\sigma^2 \to 0$ então:

$$p(y;\theta,\lambda) \approx \{2\pi\sigma^2 V(y)\}^{-1/2}\exp\{-D(y,\mu)/(2\sigma^2)\},\ y \in \Omega,$$

onde Ω é o domínio do espaço de variáveis. Esta é chamada de **aproximação de ponto de sela.** (Bent, 1987)

Observe que a **aproximação de ponto de sela** é exata para a distribuição **Normal**. Para verificá-lo, basta que seja usada uma expansão quadrática de $D(y,\mu)$ como uma função de y em torno de μ produzindo, para $y \in \Omega$,

$$D(y,\mu) \approx (y-\mu)^2/V(\mu) \text{ para } \sigma^2 \to 0,$$

que substituindo na equação acima, nos fornece a função densidade de uma distribuição **Normal** - $N(\mu,\sigma^2 V(\mu))$.

Sabemos que a função densidade da **Gama**, Ga (μ,σ^2) é:

$$p(y;\theta,\lambda) = \frac{\lambda^{\lambda}}{\Gamma(\lambda)} y^{\lambda-1} \exp[\lambda\{\theta y + \log(-\theta)\}], y > 0 \quad \textbf{(equação1)}$$

cuja função de variância é dada por $V(\mu) = \mu^2$ e o desvio (deviance)

$$D(y, \mu) = 2\{y/\mu - 1 + \log(\mu/y)\}.$$

Considerando que a variância tende a zero, podemos aplicar o teorema citado acima, onde

$$\begin{aligned} p(y;\theta,\lambda) &\approx (2\pi\sigma\ y^2)^{-1/2} \exp\{-D(y,\mu)/(2\sigma^2)\} \\ &= \lambda^{1/2}(2\pi)^{-1/2} y^{-1} \exp[-\lambda\{-1+\log(1/y)\} + \lambda\{y\theta + \log(-\theta)\}] \\ &= \lambda^{1/2} e^{\lambda} (2\pi)^{-1/2} y^{\lambda-1} \exp[\lambda\{\theta y + \log(-\theta)\}] \textbf{(equação2)} \end{aligned}$$

Este resultado nos garante que, se temos uma amostra proveniente de uma distribuição **Gama**, cuja variância seja pequena (tendendo a zero), existe uma convergência em distribuição para a Normal. Observe que a **aproximação de ponto de sela** é exata após uma renormalização para distribuição **Gama**, isto quer dizer que se dividirmos **equação2** por sua integral com relação a *y*, obtemos a densidade exata relativa à **equação1.**

Pelo resultado acima podemos afirmar:

$$D_1/\sigma^2 \approx \chi^2(n-k_1), \text{ para } \sigma^2 \to 0 \text{ , sob } H_0;$$

$$(D_2 - D_1)/\sigma^2 \approx \chi^2(k_1-k_2), \text{ para } \sigma^2 \to 0, \text{ sob } H_1.$$

Sendo que D_1 e D_2-D_1 são assintoticamente independentes, sobre a hipótese alternativa, para $\sigma^2 \to 0$, temos:

$$\hat{\sigma}^2 \approx D_1/(n-k_1)$$

$$\underline{(D_2 - D_1)}/\underline{(k_1-k_2)} \approx F(k_1-k_2; n-k_1)$$

$D_1/(n-k_1)$

Baseado neste marco teórico, mostraremos, agora, alguns resultados práticos do "**backward**" [*Heavly*, 1990] nos **níveis 1 e 2** nas tabelas abaixo.

Nestas tabelas todas as variáveis foram sendo retiradas até o modelo final, onde permanece apenas o intercepto e a variável **EC** (Electrical Condutivity). Por exemplo, na primeira linha da coluna "modelo" temos o modelo completo, na segunda temos o mesmo modelo sem a variável **Sand**, na terceira o modelo anterior sem a variável **Silt**, e assim por diante.

Nível 1

Modelo	Deviance	G.L.	Dif. Deviance	Dif. G.L.	F
Completo	45,407	43			
-sand	45,407	44	0	1	0,00
-silt	45,411	45	0,004	1	0,0039
-clay	47,953	46	2,542	1	2,52
-ph	48,124	47	0,171	1	0,164
-orgmat	51,359	48	3,235	1	3,16
-ec	112,37	49	61,011	1	57,02

Nível 2

Modelo	Deviance	G.L.	Dif. Deviance	F
Completo	28,440	24		
-orgmat	28,502	25	0,062	0,0523
-Silt	32,412	26	3,91	3,429
-clay	32,638	27	0,226	0,181
-ph	32,777	28	0,139	0,115
-orgmat	34,889	29	2,112	1,804
-ec	60,320	30	25,431	21,138

Percebemos que a única variável que não foi retirada do modelo foi "EC" (*Electrical Condutivity*), nos dois níveis. Não explicitaremos o modelo final deste ajuste, pois este não nos traz nenhuma informação adicional, uma vez que permaneceu apenas uma das variáveis. Talvez possamos melhorar o ajuste se não usarmos o intercepto:

Nível 1

Modelo	Deviance	G.L.	Dif. Deviance	F
Completo (sem intercepto)	45,465	44		
-silt	46,067	45	0,602	0,583
-sand	46,068	46	0,001	0,0009
-clay	48,513	47	2,445	2,441
-orgmat	49,993	48	1,48	1,434

-ph	145,900	49	95,907	92,084
-ec	88,684	49	38,691	37,148

Nível 2

Modelo	Deviance	G.L.	Dif. Deviance	F
Completo (sem intercepto)	28,920	25		
-clay	32,136	26	3,216	2,780
-silt	32,723	27	0,587	0,475
-sand	32,780	28	0,057	0,047
-orgmat	32,789	29	0,009	0,007
-ph	93,028	30	60,239	53,278
-ec	60,035	30	27,246	24,097

Assim, temos os modelos nos dois níveis com as mesmas variáveis, diferenciados apenas pelos valores dos coeficientes:

Nível 1 → Gypsum = 0,06708PH - 0,1747EC

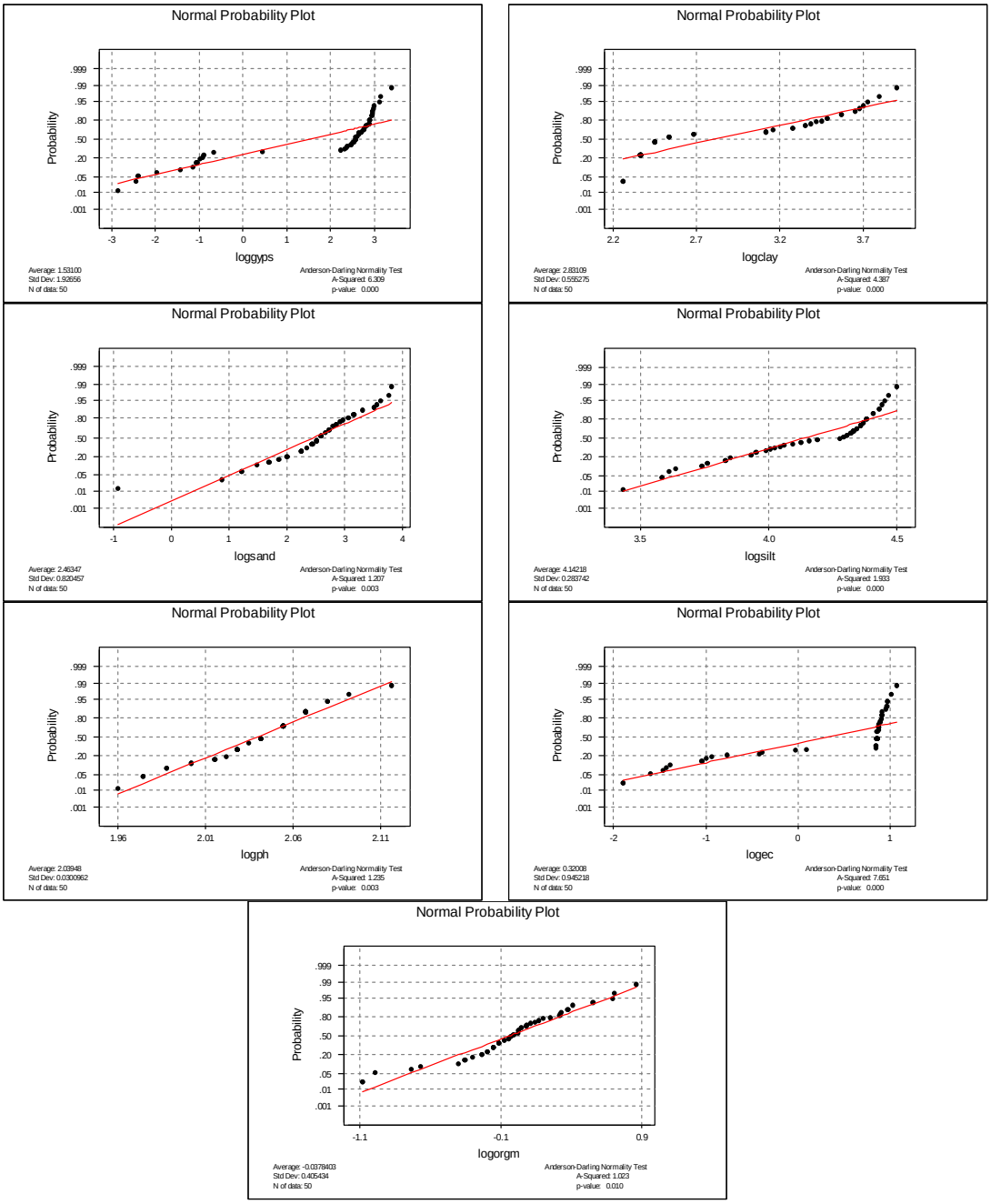
Nível 2 → Gypsum = 0,08963PH - 0,2378EC

Os modelos apresentados acima apresentam maior coerência, dado que usam dois tipos de informações do solo: uma **física** e uma **química**. Note que não usamos nenhuma das variáveis de composição (textura) do solo.

Continuaremos a tentar melhores ajustes para os dados. Como até agora nenhum dos ajustes foi inteiramente satisfatórios, talvez se transformássemos os dados melhoraremos o ajuste. Lembre-se que sempre transformaremos de forma a obter a mesma distribuição em todas as variáveis.

Transformaremos os dados agrupando os níveis, já que não notamos muitas diferenças nas conclusões obtidas até agora para cada nível separadamente.

Primeiramente vamos buscar uma transformação logarítmica onde estamos interessados em homogeneizar as variáveis.



Percebemos que a transformação logarítmica apesar de ter melhorado a forma da distribuição dos dados ao longo da reta central, não foi suficiente para considerarmos os dados transformados como normalmente distribuídos.

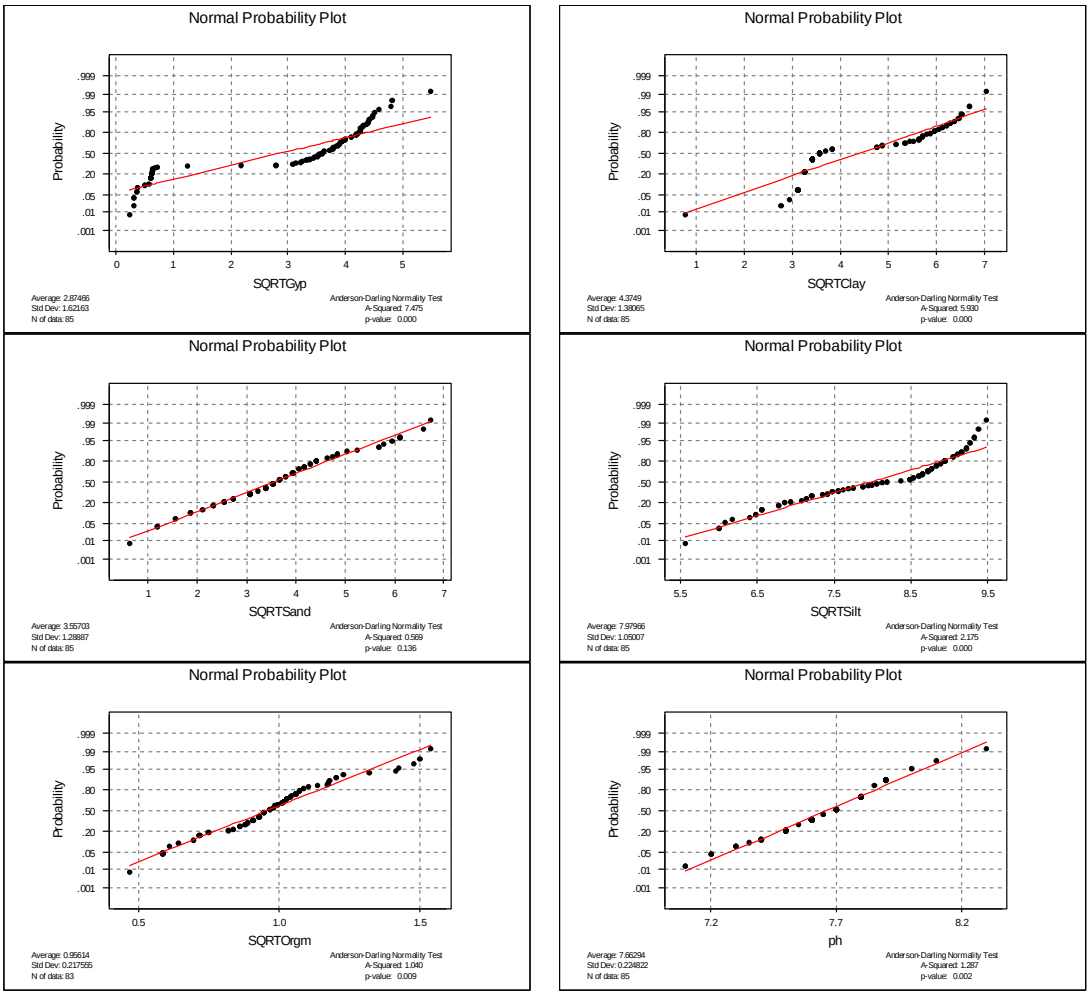
Pensando no caso em que a distribuição **Qui-quadrado** é um caso particular da distribuição Gama, podemos fazer transformações diferenciadas nas variáveis de forma a obter sempre uma distribuição normal.

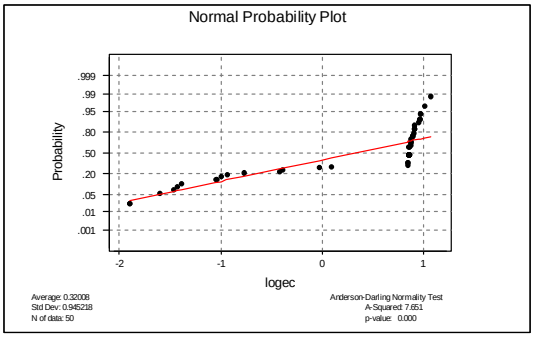
Levando em consideração a natureza dos dados faremos as seguintes transformações:

* Como as variáveis *Gypsum*, *Clay*, *Sand*, *Silt* e *Organic Matter*, são consideradas como cumulativas, a distribuição **Qui-Quadrado** se encaixa melhor, lembrando que esta é uma particularidade da **Gama**.

* Na variável *PH* não será feita nenhuma transformação pois desde o início já mostramos podermos considerá-la como **Norma**l.

* À primeira vista, a variável *Electrical Condutivity*, pela sua essência, pode ser considerada como tendo distribuição **Normal**. Todavia, notemos que em todos os gráficos temos dois grupos bem definidos, o que sugere a presença de duas populações diferenciadas. Apesar de não parecer razoável, primeiramente podemos tentar apenas a homogeneização dos valores desta variável operando-se o logaritmo.





Apesar de haver alguma melhora, notamos que as transformações não foram suficientes para sustentarmos a suposição de normalidade. Mas como já citado anteriormente, a variável *Electrical condutivity* apresenta uma configuração especial, onde observamos dois grupos de dados, no qual um grupo apresenta valores mais altos para esta variável e o outro apresenta valores mais baixos. Percebemos também que na variável resposta *Gypsum* também ocorre um agrupamento semelhante. O que nos leva a pensar que se tratássemos estes dados como duas populações, talvez teríamos um melhor ajuste dos dados.

Para esta análise utilizou-se o programa ***GLIM*** [*Heavly*, 1990)]. Os dados foram estratificados pela variável Electrical Condutivity e a análise prosseguiu pela diferenciação dos níveis.

Nível 1 Categoria 1
Alta condutividade elétrica

Model	Deviance	DF	Dif. Deviance	F
Completo (saturado)	1.8642	28		
-clay	1.8642	29	0.000000406	0.000006
-sand	1.9148	30	0.05059	0.786991
-silt	2.2944	31	0.3796	5.947357
+silt-ph	2.0003	31	0.0855	1.339565
-ec	2.0265	32	0.02624	0.406659
-orgmat	2.0265	33	0.000006183	0.0001

MODELO: **Gypsum = 0.1053 - 0.0005732 Silt**

O modelo proposto por este procedimento definiu um ajuste com as mesmas variáveis apontadas pelo procedimento *stepwise*. No entanto os coeficientes deste modelo têm uma magnitude muito inferior àquela do modelo anterior.

Nível 1 Categoria 2
Baixa condutividade elétrica

Model	Deviance	DF	Dif. Deviance	F

Completo (saturado)	2.8919	9		
-clay	2.8919	9	0	0
-sand	4.0925	10	1.201	3.737668
-silt	4.11	11	0.01750	0.04276
-ph	5.0154	12	0.9054	2.422
-ec	8.6539	13	3.638	8.70439
+ec-orgmat	5.1328	13	0.1174	0.28089

MODELO: **Gypsum = 4.934-3.727 EC**

Neste modelo permaneceu apenas a variável EC, não adicionando nenhuma informação.

Nível 2 Categoria 1
Alta condutividade elétrica

Model	Deviance	DF	Dif. Deviance	F
Completo (saturado)	28.44	24		
-clay	32.109	25	3.67	3.097
-sand	32.636	26	0.53	0.4126
-silt	32.775	27	0.14	0.1115
-ph	34.569	28	1.79	1.4746
-ec	57.996	29	23.43	18.9777
+ec-orgmat	34.889	29	0.32	0.2592

MODELO: **Gypsum = 0.5934 - 0.2023 EC**

Observe que o modelo proposto por este procedimento definiu um ajuste com uma variável diferente daquela apontada pelo procedimento *stepwise*.

Nível 2 Categoria 2
Baixa condutividade elétrica

Model	Deviance	DF	Dif. Deviance	F
Completo (saturado)	1.9481	3		
-clay	1.9611	4	0.013	0.02
-sand	2.2562	5	0.295	0.6017
-silt	2.6126	6	0.3564	0.7898
-ph	2.7064	7	0.0938	0.2154

-ec	2.8472	8	0.1408	0.3642
-orgmat	2.8979	9	0.0508	0.1427

MODELO: Gypsum = 4.934-3.727 EC

Note que no procedimento *stepwise*, não foi possível ajustar um modelo.

Análises dos modelos sem intercepto:

Nível 1 Categoria 1

Model	Deviance	DF	Dif. Deviance	F
Saturado	1.9281	29		
-Clay	2.0830	30	0.1549	2.3298
-Sand	2.0871	31	0.0041	0.0590
-Silt	2.3448	32	0.2576	3.8262
-Ph	2.3555	33	0.0107	0.1460
-Ec	6.5559	34	4.2000	58841
+Ec-Orgamat	2.3557	34	0.0002	0.0028

Modelo: Gypsum = 0.02562 EC

Obtivemos um modelo totalmente diferente do anterior (com intercepto), onde a variável Silt permaneceu no ajuste final.

Nível 1 Categoria 2

Model	Deviance	DF	Dif. Deviance	F
Saturado	2.8919	9		
-Clay	4.8033	10	1.9110	5.9473
+Clay-Sand	3.3779	10	0.4860	1.5125
-Silt	3.3931	11	0.0151	0.0447
-Ph	3.8154	12	0.4223	1.3690
-Ec	8.7730	13	4.957	15.5905
+Ec-Orgamat	3.8400	13	0.0246	0.0774

Modelo: Gypsum = 0.1727 Clay - 5.638 EC

Obtivemos um modelo mais informativos em relação ao ajuste feito com intercepto. Aqui temos uma variável granulométrica e outra física.

Nível 2 Categoria 1

Model	Deviance	DF	Dif. Deviance	F
Saturado	28.920	25		
-Clay	32.136	26	3.22	2.7835
-Sand	32.651	27	0.52	0.4207
-Silt	32.780	28	0.13	0.1075
-Ph	62.868	29	30.09	25.7023
+Ph-Ec	57.811	29	25.031	21.381
+Ec-Orgamat	32.789	29	0.009	00077

Modelo: Gypsum = 0.08963 PH - 0.2378 EC

Este modelo se difere do anterior, com intercepto, pela inclusão da variável Ph. De todos os modelos ajustados até o momento é o mais informativo uma vez que compõe-se de uma variável química e outra física.

Nível 2 Categoria 2

Model	Deviance	DF	Dif. Deviance	F
Saturado	1.9602	4		
-Clay	1.9611	5	0.001	0.002
-Sand	2.2564	6	0.2952	0.7526
-Silt	2.7875	7	0.5311	1.4122
-Ph	4.6398	8	1.852	4.6508
-Ec	4.6994	9	0.06	0.1035
-Orgamat				

No **nível 2** com baixa condutividade foi impossível o ajuste de um modelo, provavelmente, devido ao pequeno número de observações. Pela Análise de *Deviance* nenhuma das seis variáveis eram significativas ao nível de significância de 5%.

VI. Conclusões

Num primeiro momento, considerando as variáveis como normalmente distribuídas, ao

buscarmos um melhor ajuste para os dados, obtivemos modelos diversos formados pelas variáveis EC, Clay e Silt. O que já era esperado, pois são variáveis que possuem as mais altas correlações com a variável resposta Gypsum. Forma que se evidencia na análise de conglomerados, onde formou-se 2 grupos, onde estas variáveis se dissociaram, ficando cada uma num grupo diferente.

Ao dividir os dados em dois grupos de EC (alta e baixa condutividade), percebemos de forma clara o agrupamento existente com relação à granulometria. Nestes modelos a variável EC não domina tanto no ajuste quanto anteriormente, sendo que outras variáveis passam a contribuir mais na explicação da variável resposta, Gypsum.
Agora, considerando as variáveis como tendo distribuição Gama, e sem intercepto, os modelos ajustados se mostraram mais coerentes com a natureza dos dados. Pois dentre os modelos ajustados em cada situação, o que melhor se enquadra às nossas expectativas é onde a variável resposta Gypsum é explicada por uma variável associada à uma propriedade física do solo - EC , e outra química - PH.

VII. Referências Bibliográficas

Jϕrgensen, Bent (1992). **The Theory of Exponential Dispersion Models and Analysis of Deviance**. *Monografias de Matemática* **51**. Rio de Janeiro: Instituto de Matemática pura e aplicada.

Heavly, M.J.R. (1990). **Glim: An Introduction**. Oxford: Clarendon Press.

Dobson, Annette, J. (1990). **An Introduction to Generalized Linear Models**. Chapman and Hall.

Augustin, Cristina H.H. Rocha (1981) **An Investigation of the Relationship between Plant Species Occurence and Soil Gypsum Content Near Alcantarilla, Southeast Spain**. Tese de Mestrado, Universidade de Sheffield.

Guillaume, A. (1977). **Introduction à la Géologie Quantitative**. Masson.

APÊNDICE A

Análise de sistemas fechados

Análise Fatorial das variáveis Gypsum, Clay, Sand e Silt

Primeiramente fizemos a análise fatorial para o sistema granulométrico fechado das variáveis de solo analisadas em laboratório, levando-se em consideração os três níveis sem discriminá-los.

Factor Analysis

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

85 cases used 65 cases contain missing values

Variable	Factor1	Factor2	Factor3	Factor4	Communality
gypsum	0.895	0.209	0.395	-0.000	1.000
clay	-0.871	-0.436	0.223	0.049	1.000
sand	-0.482	0.875	-0.012	0.035	1.000
silt	0.972	-0.148	-0.170	0.061	1.000
Variance	2.7367	1.0217	0.2343	0.0073	4.0000
% Var	0.684	0.255	0.059	0.002	1.000

Podemos observar que com apenas dois fatores explicamos, aproximadamente, 93% da proporção de variabilidade total do processo. Note que, referente ao fator 1 os coeficientes de correlação entre as variáveis Gypsum, Clay e Silt e este fator são altos. Numa primeira análise poderíamos tomar apenas o fator 1 como uma "covariável" que resumiria estas quatro variáveis em uma só, posto que esta estaria explicando aproximadamente 68% da proporção de variabilidade total da granulometria.

Nesta análise nós pudemos evidenciar as nossas suspeitas iniciais de que existia a formação de dois grupos granulométricos neste sistema fechado. Através dos coeficientes associados ao fator 1, construímos a primeira combinação possível, dada por: **Grupo 1** – Gypsum e Silt, oposto ao **Grupo 2** – Clay e Sand. Desta forma, o fator 1 pode ser interpretado como uma comparação entre estes dois grupos.

Nível 1

Factor Analysis

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Gypsum1	0.894	0.234	0.382	-0.000	1.000
Clay1	-0.849	-0.477	0.224	0.033	1.000
Sand1	-0.580	0.814	-0.007	0.027	1.000
Silt1	0.975	-0.146	-0.159	0.045	1.000
Variance	2.8088	0.9659	0.2215	0.0039	4.0000
% Var	0.702	0.241	0.055	0.001	1.000

Como na análise feita anteriormente, observamos que com apenas dois fatores explicamos, aproximadamente, 94% da proporção de variabilidade total do processo. Ocorrendo também, no fator 1, valores altos dos coeficientes de correlação entre as variáveis Gypsum, Clay e Silt e este fator. Numa primeira análise poderíamos tomar apenas o fator 1 como uma "covariável" que resumiria estas quatro variáveis em uma só, posto que esta estaria explicando aproximadamente 70% da proporção de variabilidade total da granulometria.

Esta análise sustenta as nossas suspeitas iniciais de que existia a formação de dois grupos granulométricos neste sistema fechado. Através dos coeficientes associados ao fator 1, construímos a primeira combinação possível, dada por: **Grupo 1** – Gypsum e Silt, oposto ao **Grupo 2** – Clay e Sand. Desta forma, o fator 1 pode ser interpretado como uma comparação entre estes dois grupos.

Nível 2

Factor Analysis

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

31 cases used 19 cases contain missing values

Variable	Factor1	Factor2	Factor3	Factor4	Communality
gypsum2	0.901	-0.156	0.405	-0.000	1.000
clay2	-0.910	0.347	0.216	-0.067	1.000
sand2	-0.280	-0.959	-0.025	-0.040	1.000
silt2	0.961	0.196	-0.182	-0.075	1.000
Variance	2.6413	1.1028	0.2441	0.0118	4.0000
% Var	0.660	0.276	0.061	0.003	1.000

Para o nível 2 temos um resultado análogo ao do nível 1, sendo que com apenas dois fatores explicamos, aproximadamente, 93% da proporção de variabilidade total do processo. No fator 1, os valores dos coeficientes de correlação entre as variáveis Gypsum, Clay e Silt e este fator são mais altos. Numa primeira análise poderíamos tomar apenas o fator 1 como uma "covariável" que resumiria estas quatro variáveis em uma só, posto que esta estaria explicando 66% da proporção de variabilidade total da granulometria.

E mais uma vez observamos a formação de dois grupos granulométricos neste sistema fechado. Através dos coeficientes associados ao fator 1, construímos a primeira combinação possível, dada por: **Grupo 1** – Gypsum e Silt, oposto ao **Grupo 2** – Clay e Sand. Desta forma, o fator 1 pode ser interpretado como uma comparação entre estes dois grupos.

Concluímos, assim, que no que concerne a análise fatorial das variáveis granulométricas deste sistema fechado, a categorização em níveis de profundidade se mostrou não sensível com relação à análise global.

Nível 1

Factor Analysis

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Commnlty
clay1	0.787	0.616	0.999
sand1	0.694	-0.719	0.999
silt1	-0.999	-0.014	0.998
Variance	2.0990	0.8971	2.9961
% Var	0.700	0.299	0.999

Apenas com dois fatores conseguimos explicar praticamente toda a variabilidade do processo no primeiro nível. Associado ao primeiro fator, temos valores bem elevados para as três variáveis, principalmente para Silt, o que pode ser interpretado como uma medida global da composição do solo. Este fator explica 70% da variabilidade dos dados. No segundo fator, dominam as variáveis Clay e Sand, que pode ser interpretado basicamente, como uma comparação entre as quantidades de destas variáveis em oposição. Este segundo fator explica apenas 29,9% da variabilidade dos dados, o que nos leva a crer que talvez apenas o primeiro fator seja suficiente numa análise estatística multivariada. Mas esta decisão tem que ser tomada junto ao pesquisador, pois pode ser que para ele o segundo fator seja interessante.

Nível 2

Factor Analysis

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

31 cases used 19 cases contain missing values

Variable	Factor1	Factor2	Factor3	Commnlty
clay2	-0.873	0.483	0.067	1.000
sand2	-0.409	-0.912	0.040	1.000
silt2	0.996	0.049	0.075	1.000
Variance	1.9210	1.0672	0.0118	3.0000
% Var	0.640	0.356	0.004	1.000

Para o segundo nível temos uma situação semelhante a anterior, ressaltando que o primeiro fator explica uma menor variabilidade do processo baseado nestes dados (64%), e o segundo passa a explicar um pouco mais (35,6%).

Afim de nos munir de uma justificativa mais simples, porém tão concisa quanto à anterior, poderíamos lançar mão de uma técnica eficaz para a determinação de agrupamentos que se segue. Esta análise evidencia a presença das duas formações: Silte e Gypsum, Clay e Sand.

Análise de conglomerados

Correlation Coefficient Distance, Centroid Linkage

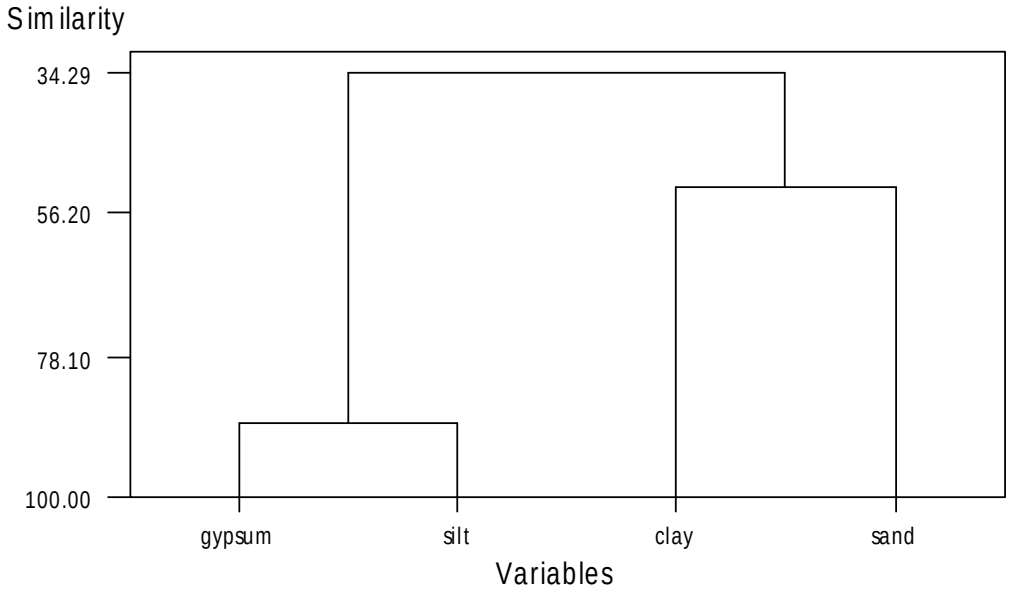
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of Obs in new cluster
1	3	88.61	0.228	1	4	1	2
2	2	51.87	0.963	2	3	2	2
3	1	34.29	1.314	1	2	1	4

Final Partition

Cluster 1
gypsum silt

Cluster 2
clay sand



Correlation Coefficient Distance, Centroid Linkage

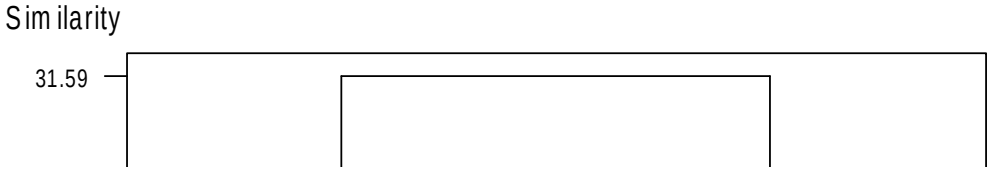
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of Obs in new cluster
1	3	88.85	0.223	1	4	1	2
2	2	55.20	0.896	2	3	2	2
3	1	31.59	1.368	1	2	1	4

Final Partition

Cluster 1
Gypsum1 Silt1

Cluster 2
Clay1 Sand1



Nível 2

Correlation Coefficient Distance, Centroid Linkage

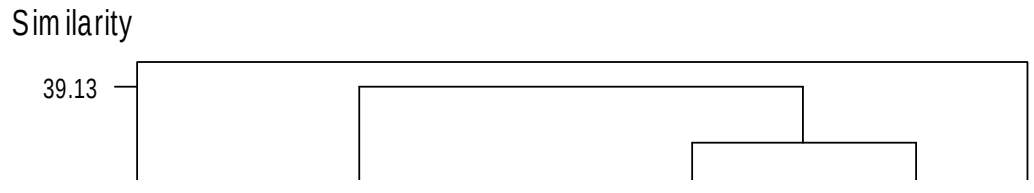
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of Obs in new cluster
1	3	88.07	0.239	1	4	1	2
2	2	45.95	1.081	2	3	2	2
3	1	39.13	1.217	1	2	1	4

Final Partition

Cluster 1
gypsum2 silt2

Cluster 2
clay2 sand2



Apêndice B

Entropia associada às variáveis Gypsum, Clay, Sand e Silt

O conceito de entropia em probabilidade, tem suas raízes na termodinâmica e na teoria das comunicações em Ciências da Terra.

Seja um conjunto sedimentar granulométrico compreendendo Clay (c%), Silt (si%), Sand (sa%) e Gypsum (g%), onde *c%+si%+as%+g%=100%* - um sistema fechado em cada observação. Por definição, a informação total fornecida por este conjunto é

$$I_t = -[c\% \ln c\% + si\% \ln si\% + sa\% \ln sa\% + g\% \ln g\%] .$$

De forma geral e introduzindo nesta expressão uma constante arbitrária *K*, a entropia *H* deste conjunto é dada por

$$H = -K.I_t ,$$

comumente esta constante é considerada igual a 100.

Note que a entropia é máxima se os quatro componentes, "independentes", têm a mesma probabilidade de ocorrência no conjunto: $H_M = \ln 4$; como o logaritmo é tomado à base *e*, a unidade de entropia é o "*nit*".

Observe que, se X_1, X_2, X_3, X_4 são quatro eventos independentes, a entropia do conjunto é

$$H_c = H_{X_1} + H_{X_2} + H_{X_3} + H_{X_4},$$

se estes eventos não forem independentes:

$$H_c < H_{X_1} + H_{X_2} + H_{X_3} + H_{X_4}.$$

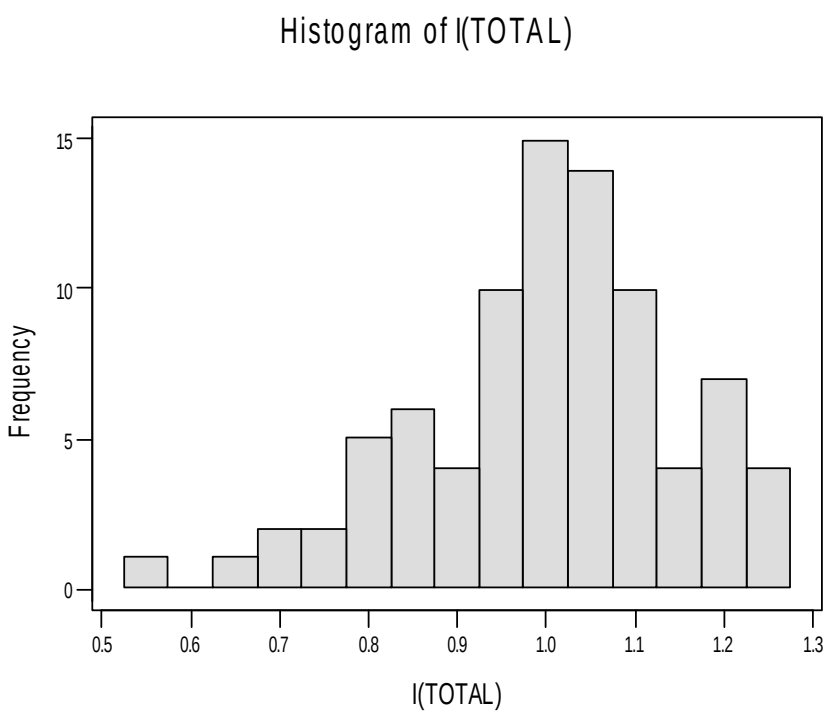
A entropia relativa é definida como $H_r = H / H_M$, onde H_M é a máxima entropia. A entropia relativa é, geralmente, expressa em percentagem. Por interesses inerentes aos processos geológicos, a entropia relativa tem um papel importante em estudo de sistemas granulométricos fechados.

Uma entropia elevada, denota um alto grau de mistura dos componentes (proporções vizinhas). Uma baixa entropia, denota a predominância de um termo - ou uma combinação destes – sobre os outros. Não obstante, a entropia não nos permite distinguir o termo predominante. Note que, historicamente, a entropia, aparece como sinônimo de probabilidade freqüentista.

Descrição da informação total do sistema fechado em estudo:

Variable	N	N*	Mean	Median	Tr Mean	StDev	SE Mean
I(TOTAL)	85	65	0.9998	1.0117	1.0049	0.1466	0.0159

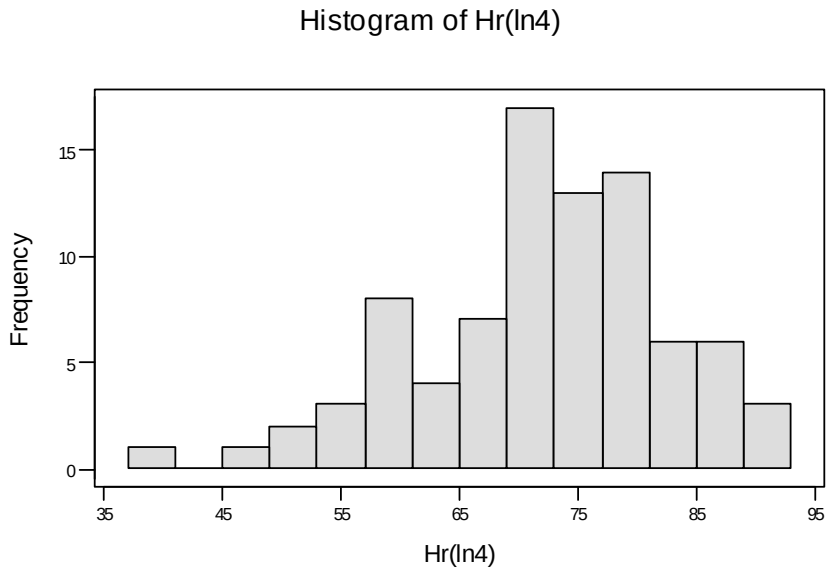
Variable	Min	Max	Q1	Q3
I(TOTAL)	0.5307	1.2670	0.9260	1.1011



Descrição da entropia relativa associada ao sistema fechado em estudo:

Variable	N	N*	Mean	Median	Tr Mean	StDev	SE Mean
Hr(ln4)	85	65	72.12	72.98	72.49	10.58	1.15

Variable	Min	Max	Q1	Q3
Hr(ln4)	38.28	91.39	66.80	79.43



Onde nós concluímos que a entropia relativa média é de 72,12%.

Assim, a equivalência entre entropia e probabilidade pode ser estabelecida de forma mais ampla. Sejam quatro fontes de informação não independentes, X_1, X_2, X_3, X_4 - referentes a Clay (c%), Silt (si%), Sand (sa%) e Gypsum (g%) - constituindo o conjunto granulométrico (G). Sejam n_1, n_2, n_3, n_4, o número de "mensagens" que podem ser "emitidas" por X_1, X_2, X_3, X_4, respectivamente. Se as probabilidades de aparição de n_1, n_2, n_3, n_4 são as mesmas, o número total de "mensagens" recebidas pelo conjunto granulométrico (G), n_G é inferior ao produto do número de mensagens individuais

$$n_G < \prod_{i=1}^{4} n_i \ .$$

A quantidade total de informação proveniente de (G) é uma função $f(n_G)$; ela é inferior à soma das informações elementares "emitidas" por cada componente:

$$f(n_G) < \sum_{i=1}^{4} f(n_i) \ .$$

Uma solução possível deste sistema é:

$$f(n_G) < K.\ln n_G \ ;$$

ou, sob uma outra forma: $f(n_G) < f(\prod_{i=1}^{4} n_i) = K.\ln(\prod_{i=1}^{4} n_i) = \sum_{i=1}^{4} (K.\ln n_i) = \sum_{i=1}^{4} f(n_i)$.

Se a fonte de informações (G) comporta n_G mensagens possíveis e equiprováveis, a informação associada a uma mensagem é $f(n_G) \le K.\ln n_G$.

Agora, nos colocamos no núcleo de nosso problema. Suponha que uma fonte de informação compreendendo quatro mensagens elementares, de igual probabilidade individual de aparição, porém reagrupados em dois subconjuntos disjuntos $G = G_1 \bigcup G_2$, de probabilidades de aparição diferentes p_{G_1} e p_{G_2} - no nosso estudo, relativo aos resultados anteriores, podemos supor $G_1 = \{Gypsum, Silte\}$ e $G_2 = \{Clay, Sand\}$. A informação máxima é $\ln 4$, com $K = 1$.

Para uma mensagem do grupo G_1, a informação é $\ln n_{G_1}$, e para uma mensagem do grupo G_2, a informação é $\ln n_{G_2}$. Estas informações, apenas são emitidas proporcionalmente às relações $n_{G_1}/4$ e $n_{G_2}/4$.

A diferença *H* entre a informação máxima e a informação disponível é

$$H = \ln 4 - [\frac{n_{G_1}}{4} \ln n_{G_1} + \frac{n_{G_2}}{4} \ln n_{G_2}] .$$

Desta forma, se nós estudarmos a proporção elementar, teremos

$$H = -[\, p_{G_1} \ln p_{G_1} + p_{G_2} \ln p_{G_2} \,],$$

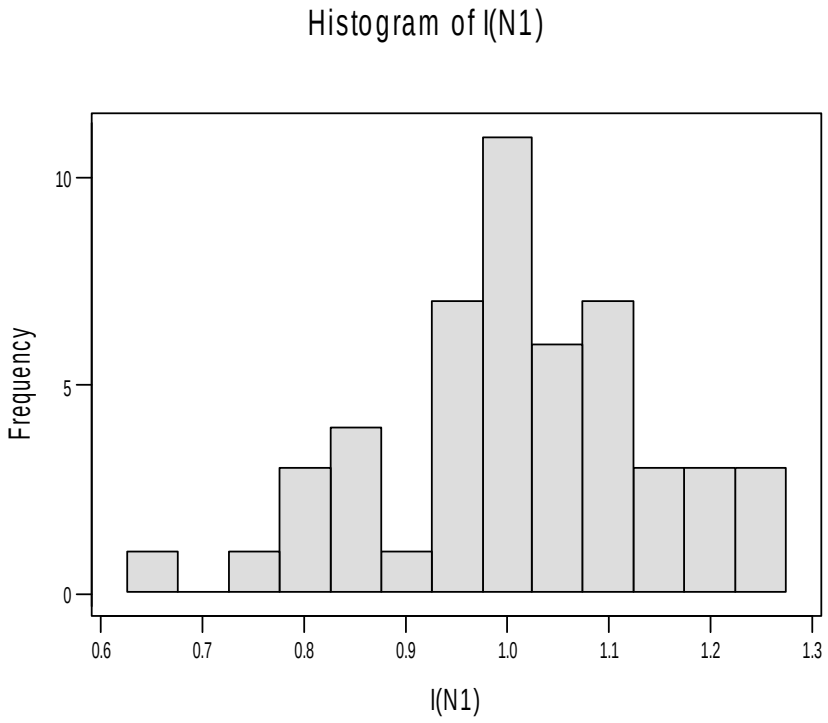
ou, $H = -\sum_{i=1}^{4} p_i \ln p_i$.

Assim analisamos:

Nível 1

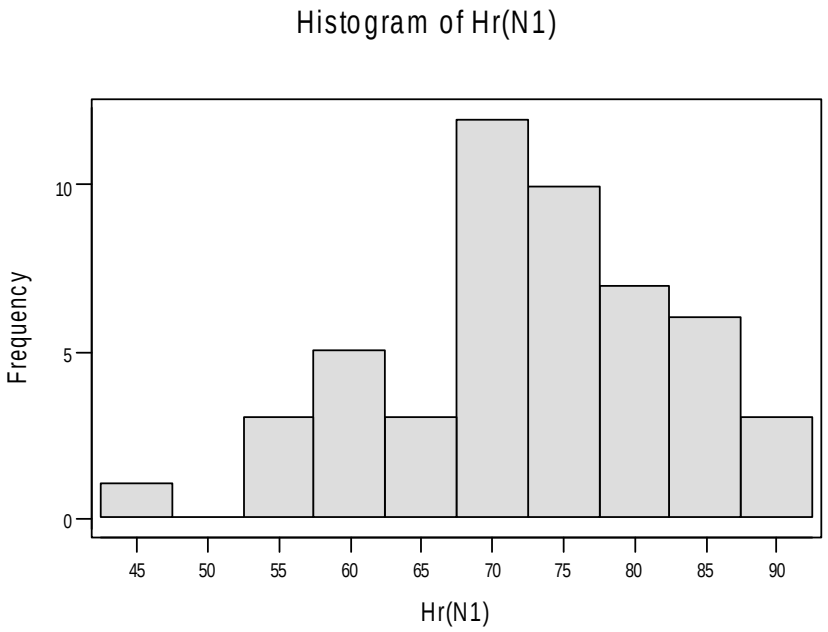
Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
I(N1)	50	1.0077	1.0130	1.0108	0.1333	0.0188

Variable	Min	Max	Q1	Q3
I(N1)	0.6574	1.2495	0.9443	1.0981



Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
Hr(N1)	50	72.69	73.07	72.91	9.61	1.36

Variable	Min	Max	Q1	Q3
Hr(N1)	47.42	90.13	68.12	79.21

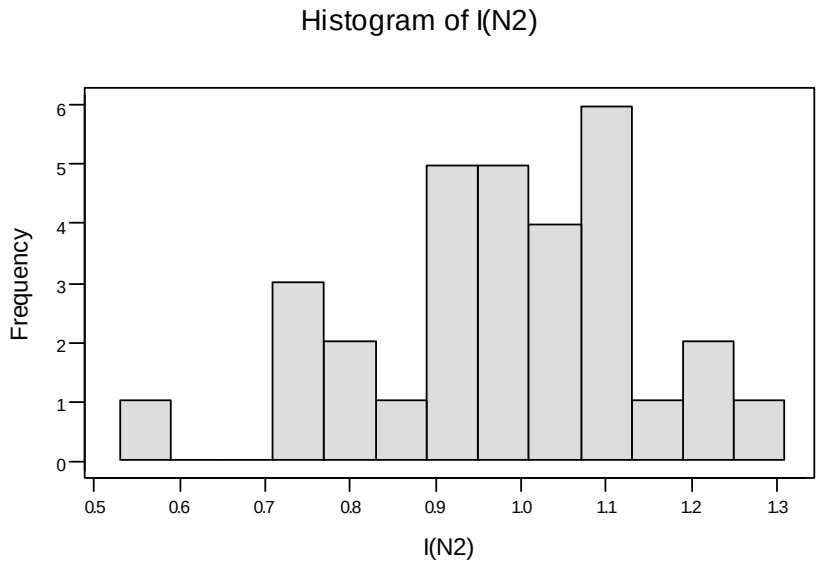


Onde nós concluímos que a entropia relativa média é de 72,69%.

Nível 2

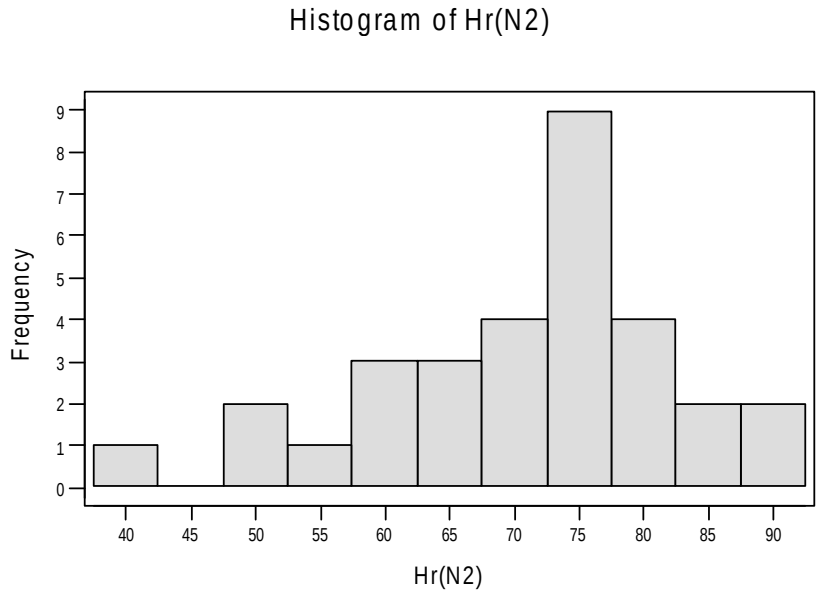
Variable	N	N*	Mean	Median	Tr Mean	StDev	SE Mean
I(N2)	31	19	0.9797	1.0070	0.9866	0.1644	0.0295

Variable	Min	Max	Q1	Q3
I(N2)	0.5307	1.2670	0.8947	1.0749



Variable	N	N*	Mean	Median	Tr Mean	StDev	SE Mean
Hr(N2)	31	19	70.67	72.64	71.17	11.86	2.13

Variable	Min	Max	Q1	Q3
Hr(N2)	38.28	91.39	64.54	77.54



Onde nós concluímos que a entropia relativa média é de 70,67%.