1 MODELAGENS UNIVARIADAS

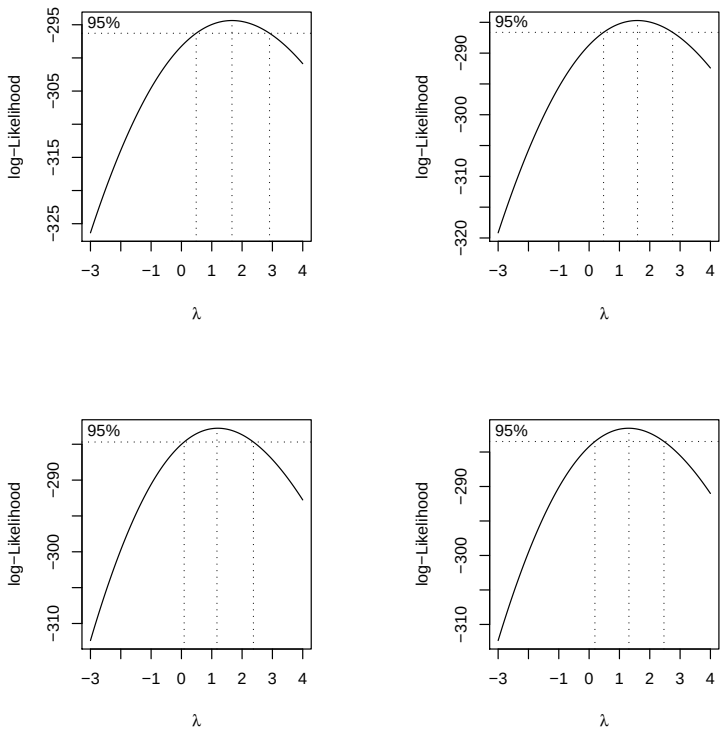
Nesta seção serão apresentados os resultados da modelagem geoestatística proposta para cada uma das variáveis resposta.

1.1 Saturação por bases

Com relação a esta variável, da análise exploratória inicial, suspeita-se que existe um padrão espacial nos dados, além disso, suspeita-se que a média do processo, aparentemente, é influenciada ou pela coordenada x ou pela área de manejo. Sendo assim, serão propostos modelos que consideram estacionariedade da função de correlação, mas com diferentes tendências para as médias, sendo assim, o modelo pode ter mais ou menos parâmetros relativos a média.

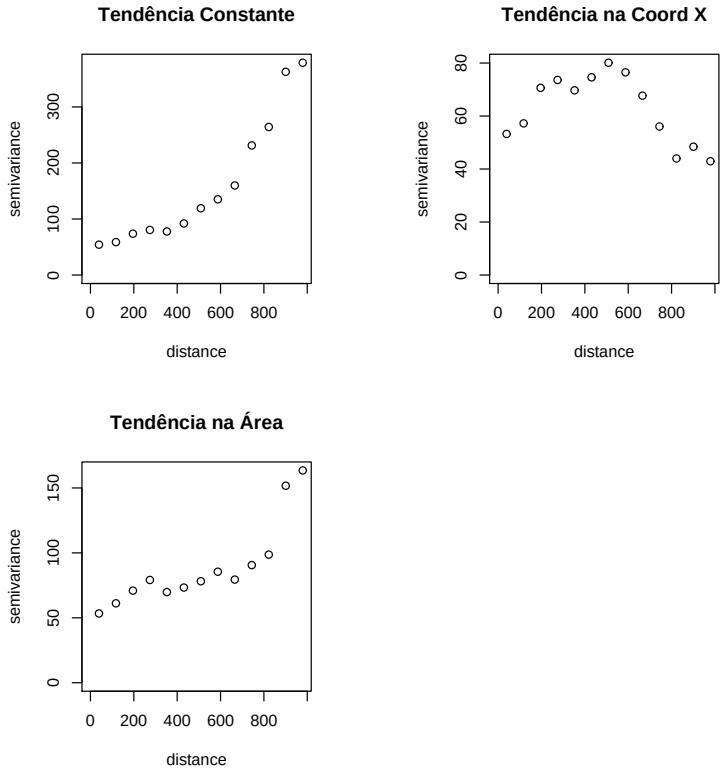
Para toda a modelagem foi utilizada a família Matérn de funções de correlações válidas, essa escolha foi feita por conta dessa família possuir funções deriváveis e não deriváveis em todo o domínio, ou seja, essa família engloba funções suaves e não suaves para as correlações, e essa suavidade do processo é determinada através do parâmetro κ da função, kappas maiores que 1.5 são as funções deriváveis.

No entanto, antes de propor alguma modelagem, é atribuído ao campo aleatório e ao ruído branco distribuições gaussianas, além disso é suposto estacionariedade das variâncias e covariâncias, conforme visto na revisão bibliográfica, sendo assim, esses pressupostos devem ser testados, seguem os gráficos dos λ's estimados para a transformação da família de Box-Cox para cada tendência estudada:



Os gráficos acima representam os intervalos de confiança para os lambdas estimados para a transformação de Box-Cox, sendo que o primeiro não considera tendência alguma, o segundo considera tendência na área de manejo, o terceiro considera tendência na coordenada X e o último considera tendência na área e na coordenada X. Como todos os intervalos de confiança contem λ=1, não será conduzida nenhuma transformação nos dados.

Agora o próximo passo é fazer a estimação dos parâmetros para alguns modelos, para tal será utilizado o método da máxima verossimilhança, no entanto, devido a complexidade do sistema de derivadas que deve ser resolvido, esse método utiliza métodos numéricos para calculas as estimativas, e como todos métodos numéricos precisam de um valor inicial para começar as iterações serão apresentados gráficos de semivariogramas empíricos, os quais serão utilizados para dar chutes iniciais aos parâmetros, cabe ressaltar que foram fixados alguns κ's distintos de forma que as funções de correlações englobadas na modelagem sejam mais ou menos suvaes:



Os gráficos acima mostram que, para cada tendência considerada os valores de semivariograma empírica são bem distintos, além disso, tem-se que para distância grande entre as localizações os valores de semivariograma empírico se comportam de forma estranha, nos casos de tendência na área de manejo e tendência constante, o semivariograma não pára de crescer, ou seja, se fosse utilizado esse método para estimação dos parâmetros, deveria ser considerando um alcance prático de forma que a partis de uma certa distância seria considerado que as localizações não possuem mais correlação. No caso da tendência em X, o comportamento é ainda mais estranho, pois o ruído branco é maior que o sinal e tem-se ainda que, o semivariograma vai aumentando conforme a distância aumenta e em um certo

ponto a estatística cai novamente, ou seja, para valores mais distantes a correlação entre as observações volta a crescer, essa característica destoa totalmente dos pressupostos da função de covariância, que quanto maior as distâncias menor a correlação entre os valores do campo aleatório. No entanto, o semivariograma empírico não é uma boa medida para estabelecer os parâmetros estimados, ou seja, não é muito adequado tentar ajustar um modelo aos valores do semivariograma empírico e considerar que esse ajuste são as estimativas para os parâmetros envolvidos nos modelos, esse método não deve ser utilizado por conta do acaso amostral ou pelo tamanho da amostra, pois se existem poucas observações, alguns semivariogramas serão calculados com poucas observações que estarão dentro da distância considerada. Sendo assim, os gráficos acima têm caráter exploratório e serão utilizados para dar os valores iniciais para os estimadores de máxima verossimilhança.

A tabela abaixo refere-se aos parâmetros estimados dos modelos com todas tendências levadas em consideração, com todos os κ's utilizados e os valores maximizados dos logaritmos das funções de verossimilhança:

β	τ^2	σ^2	ϕ	κ	log-verossim.
48.53	59.57	120.62	625.58	1	-239.8
47.98	62.40	124.31	516.78	1.5	-239.8
47.27	63.29	149.69	489.58	2	-239.7
47.32	63.61	122.37	326.39	3	-239.7

Tabela 1: Estimativas de Máxima Verossimilhança - Tendência constante

Com os resultados acima, tem-se que independente dos valores fixados para κ as estimativas de verossimilhança se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 2 e 3 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
47.92	8.83	38.78	39.28	59.18	1	-238.00
47.92	8.86	43.35	34.69	50.33	1.5	-237.98
47.92	8.88	45.53	32.47	44.34	2	-237.96
47.44	4.85	64.96	34.57	190.39	3	-239.27

Tabela 2: Estimativas de Máxima Verossimilhança - Tendência na área de manejo

Agora com tendência na área de manejo, tem-se que os resultados acima, independente dos valores fixados para κ, se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 1.5 e 2 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

Com tendência na coordenada X, tem-se que os resultados acima, independente dos valores fixados para κ, se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 2 e 3 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
-14300.66	0.025	27.56	41.46	45.42	1	-234.72
-14300.91	0.025	33.53	35.53	39.51	1.5	-234.70
-14300.17	0.025	36.40	32.68	35.33	2	-234.68
-14297.97	0.025	39.14	29.97	29.74	3	-234.66

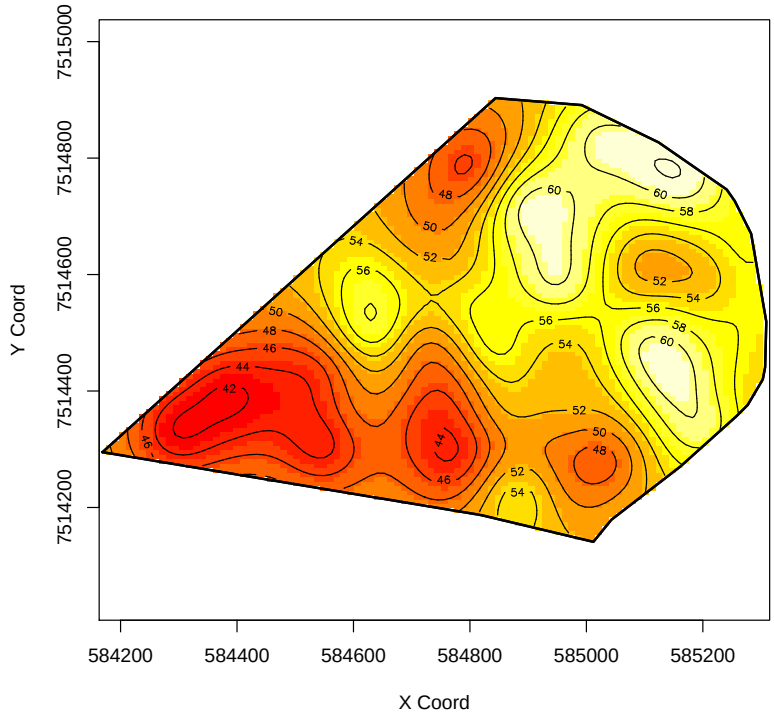
Tabela 3: Estimativas de Máxima Verossimilhança - Tendência na coordenada X

O próximo passo é escolher entre os modelos com mais ou menos parâmetros na média, ou seja, devemos fazer a seleção de covariáveis importantes ao modelo. Para tal, não se pode comparar os máximos das funções de verossimilhança, uma vez que, os valores das mesmas são alterados conforme o número de parâmetros no modelo, sendo assim, como os modelos possuem números de parâmetros de média distintos, se deve utilizar outro critério para seleção, sendo assim, será utilizado o critério da informação de Akaike, esse critério faz uma ponderação entre a explicação do modelo e o número de parâmetros usados, ou seja, esse critério é uma espécie de punição ao modelo pelo número de parâmetros utilizados para explicar uma determinada variabilidade, logo, quanto menor o valor da estatística melhor o modelo:

Tendência	κ	AIC
Constante	2	487.499
Constante	3	487.328
Area	1.5	485.9588
Area	2	485.922
Coord.X	2	479.361
Coord.X	3	479.321

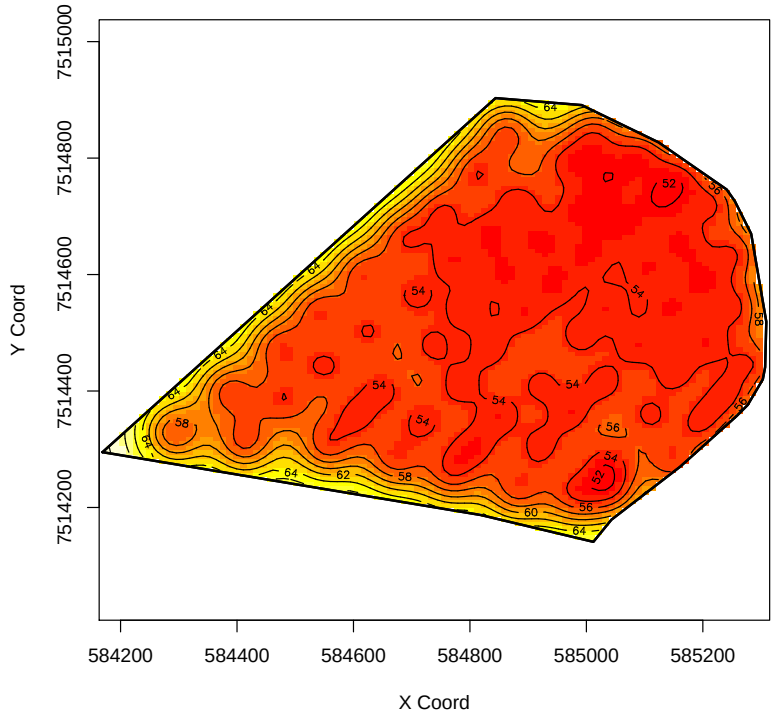
Tabela 4: Critério de informação de Akaike

Com os resultados acima, tem-se que o modelo com tendência na coordenada X e κ igual a 3 é o que melhor se ajustou aos dados, sendo assim, o próximo passo é fazer a predição ou krigagem para todo o espaço da fazenda, sendo assim, com a estimação dos parâmetros feita, é possível prever o valor do campo aleatório para localizações não amostradas, essa predição é feita através da média estimada para a localização ponderada pelos valores estimados para a variância e covariância do campo aleatório. Segue o gráfico da krigagem:



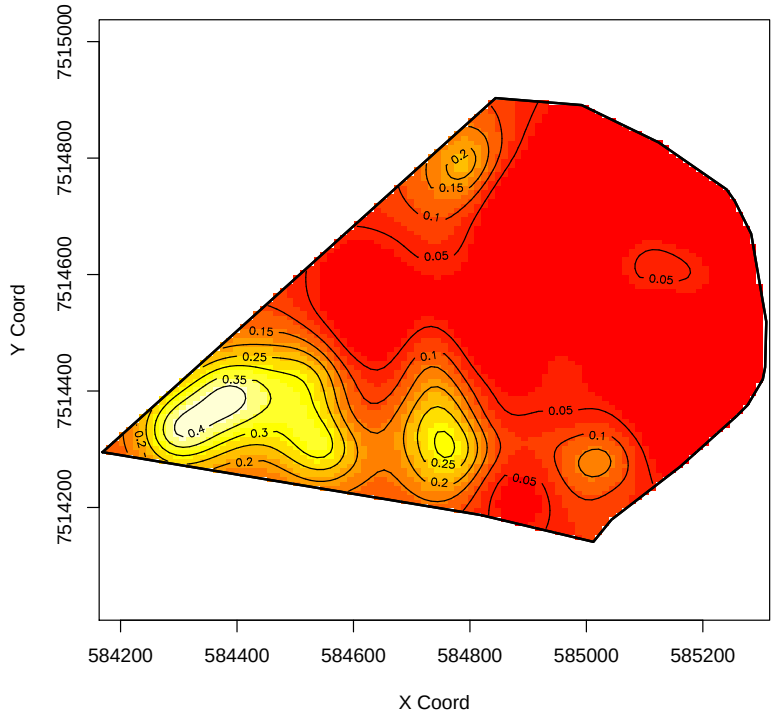
No gráfico acima as cores mais próximas do branco indicam valores mais elevados para a saturação por bases e conseqüentemente cores próximas do vermelho indicam valores menores para a saturação. Analisando os valores observados nas localizações amostradas, tem-se que as predições se aproximaram refletiram os valores reais e suavizou para o restante do espaço o campo aleatório.

O próximo passo é analisar o gráfico das variâncias das estimativas do campo aleatório, segue o gráfico:



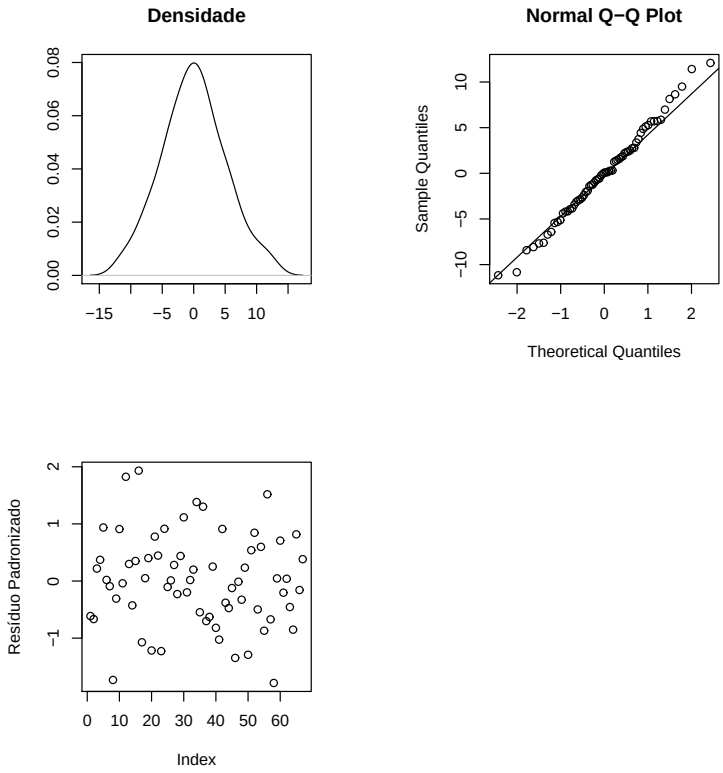
Com o gráfico acima tem-se que existe o padrão espacial esperado para as variâncias, ou seja, as variâncias mudam no espaço todo, mas essa mudança é suave, o que de certa forma corrobora a idéia de estacionariedade da função de covariância.

Com a definição dos parâmetros estimados, e com os valores preditos para cada posição do grid definido, é possível fazer análises a partir da distribuição normal de cada posição onde foi feita predição, assim é possível calcular a probabilidade prevista de que a saturação por bases seja menor que 40 em cada posição predita, ou seja, se pode fazer um gráfico das probabilidades de cada valor predito ser inferior a 40, que é um valor abaixo do esperado para uma boa qualidade do solo da fazenda sob estudo, abaixo segue o gráfico:



Com o gráfico acima tem-se que valores preditos em regiões com menor valor para a coordenada X possuem maior probabilidade da saturação ser inferior a 40, informação que corrobora a análise descritiva onde se tem a clara impressão de que a média da saturação é inferior em regiões com menores coordenadas X.

Uma outra análise que pode ser conduzida é a validação do pressuposto de normalidade univariada do ruído branco, abaixo seguem análises gráficas do tipo:



O primeiro gráfico acima mostram que a densidade dos resíduos brancos estimados se aproximam de uma distribuição normal, além disso o qqplot mostra que existe uma pequena fuga da normalidade na cauda superior dos resíduos, mas nada que afete o pressuposto e por último não existe valores para os resíduos padronizados fora do intervalo de [-3,3], o que indica não existir valores discrepantes dos erros brancos. Após a análise gráfica o teste de normalidade de Shapiro-Wilk foi conduzido, o qual gerou um p-valor de 0.9674, logo a normalidade não é rejeitada e esse pressuposto está atendido.

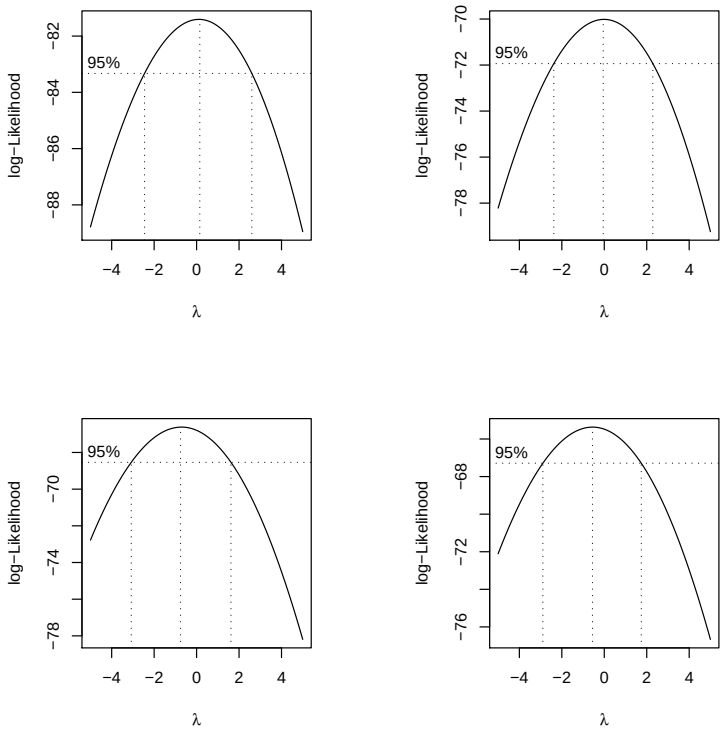
1.2 Ph

Com relação a esta variável, da análise exploratória inicial, suspeita-se que existe um padrão espacial nos dados, além disso, suspeita-se que a média do processo, aparentemente, é influenciada ou pela coordenada x ou pela área de manejo. Sendo assim, serão propostos modelos que consideram estacionariedade da função de correlação, mas com diferentes tendências para as médias, sendo assim, o modelo pode ter mais ou menos parâmetros relativos a média.

Para toda a modelagem foi utilizada a família Matérn de funções de correlações válidas, essa escolha foi feita por conta dessa família possuir funções deriváveis e não deriváveis em todo o domínio, ou seja, essa família engloba funções suaves e não suaves para as correlações, e essa suavidade do processo é determinada através do parâmetro κ da função, kappas maiores que 1.5 são as funções deriváveis.

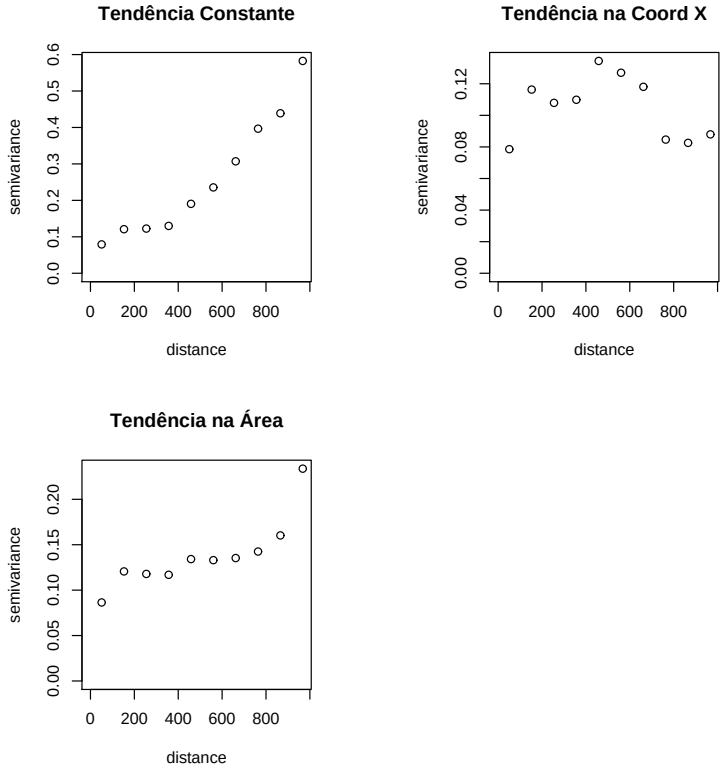
No entanto, antes de propor alguma modelagem, é atribuído ao campo aleatório e ao ruído branco distribuições gaussianas, além disso é suposto estacionariedade das variân-

cias e covariâncias, conforme visto na revisão bibliográfica, sendo assim, esses pressupostos devem ser testados, seguem os gráficos dos λ's estimados para a transformação da família de Box-Cox para cada tendência estudada:



Os gráficos acima representam os intervalos de confiança para os lambdas estimados para a transformação de Box-Cox, sendo que o primeiro não considera tendência alguma, o segundo considera tendência na área de manejo, o terceiro considera tendência na coordenada X e o último considera tendência na área e na coordenada X. Como todos os intervalos de confiança contem λ=1, não será conduzida nenhuma transformação nos dados.

Agora o próximo passo é fazer a estimação dos parâmetros para alguns modelos, para tal será utilizado o método da máxima verossimilhança, no entanto, devido a complexidade do sistema de derivadas que deve ser resolvido, esse método utiliza métodos numéricos para calculas as estimativas, e como todos métodos numéricos precisam de um valor inicial para começar as iterações serão apresentados gráficos de semivariogramas empíricos, os quais serão utilizados para dar chutes iniciais aos parâmetros, cabe ressaltar que foram fixados alguns κ's distintos de forma que as funções de correlações englobadas na modelagem sejam mais ou menos suaves:



Os gráficos acima mostram que, para cada tendência considerada os valores de semivariograma empírica são bem distintos, além disso, tem-se que para distância grande entre as localizações os valores de semivariograma empírico se comportam de forma estranha, nos casos de tendência na área de manejo e tendência constante, o semivariograma não pára de crescer, ou seja, se fosse utilizado esse método para estimação dos parâmetros, deveria ser considerando um alcance prático de forma que a partis de uma certa distância seria considerado que as localizações não possuem mais correlação. No caso da tendência em X, o comportamento é ainda mais estranho, pois o ruído branco é maior que o sinal e tem-se ainda que, o semivariograma vai aumentando conforme a distância aumenta e em um certo ponto a estatística cai novamente, ou seja, para valores mais distantes a correlação entre as observações volta a crescer, essa característica destoa totalmente dos pressupostos da função de covariância, que quanto maior as distâncias menor a correlação entre os valores do campo aleatório. No entanto, o semivariograma empírico não é uma boa medida para estabelecer os parâmetros estimados, ou seja, não é muito adequado tentar ajustar um modelo aos valores do semivariograma empírico e considerar que esse ajuste são as estimativas para os parâmetros envolvidos nos modelos, esse método não deve ser utilizado por conta do acaso amostral ou pelo tamanho da amostra, pois se existem poucas observações, alguns semivariogramas serão calculados com poucas observações que estarão dentro da distância considerada. Sendo assim, os gráficos acima têm caráter exploratório e serão utilizados para dar os valores iniciais para os estimadores de máxima verossimilhança.

A tabela abaixo refere-se aos parâmetros estimados dos modelos com todas tendências levadas em consideração, com todos os κ's utilizados e os valores maximizados dos logaritmos das funções de verossimilhança:

β	τ^2	σ^2	ϕ	κ	log-verossim.
4.905	0.1006	0.1493	510.58	1	-25.69
4.904	0.1055	0.2513	650.50	1.5	-25.57
4.903	0.1061	0.2880	569.74	2	-25.45
4.902	0.1063	0.2961	430.36	3	-25.31

Tabela 5: Estimativas de Máxima Verossimilhança - Tendência constante

Com os resultados acima, tem-se que independente dos valores fixados para κ as estimativas de verossimilhança se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 2 e 3 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores, no entanto, quanto maior o κ mais o máximo da verossimilhança está aumentando, o que pode indicar que existe alguma covariável não considerada.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
4.7094	0.4421	0.1220	0.0000	0.0000	1	-24.60
4.7094	0.4421	0.1220	0.0000	0.0000	1.5	-24.60
4.7794	0.2272	0.1100	0.1100	349.35	2	-24.86
4.7880	0.1948	0.1092	0.1092	327.03	3	-24.79

Tabela 6: Estimativas de Máxima Verossimilhança - Tendência na área de manejo

Agora com tendência na área de manejo, tem-se que os resultados acima mostram não existir padrão espacial for considerada essa tendência na média, pois os melhores modelos foram os com variabilidade e parâmetro de correlação do campo aleatório igual a zero. Sendo assim esse tipo de tendência será desconsiderado do estudo.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
-584.69	0.001	0.1124	0.0000	0.0000	1	-21.86
-586.88	0.001	0.1050	0.0104	171.76	1.5	-21.71
-608.32	0.001	0.0081	0.1040	24.508	2	-19.36
-584.69	0.001	0.1124	0.0001	822.95	3	-21.89

Tabela 7: Estimativas de Máxima Verossimilhança - Tendência na coordenada X

Com tendência na coordenada X, tem-se que os resultados acima, independente dos valores fixados para κ, se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 1.5 e 2 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

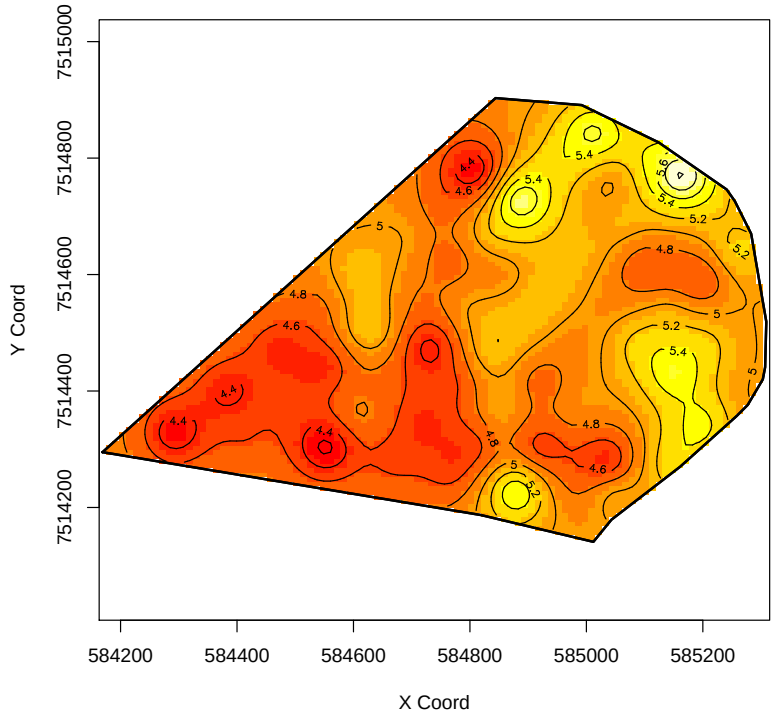
O próximo passo é escolher entre os modelos com mais ou menos parâmetros na média, ou seja, devemos fazer a seleção de covariáveis importantes ao modelo. Para tal, não se pode comparar os máximos das funções de verossimilhança, uma vez que, os valores

das mesmas são alterados conforme o número de parâmetros no modelo, sendo assim, como os modelos possuem números de parâmetros de média distintos, se deve utilizar outro critério para seleção, sendo assim, será utilizado o critério da informação de Akaike, esse critério faz uma ponderação entre a explicação do modelo e o número de parâmetros usados, ou seja, esse critério é uma espécie de punição ao modelo pelo número de parâmetros utilizados para explicar uma determinada variabilidade, logo, quanto menor o valor da estatística melhor o modelo:

Tendência	κ	AIC
$Constante$	1.5	59.147
$Constante$	2	58.902
$Constante$	3	58.612
$Coord.X$	1	53.722
$Coord.X$	1.5	53.429
$Coord.X$	2	48.727

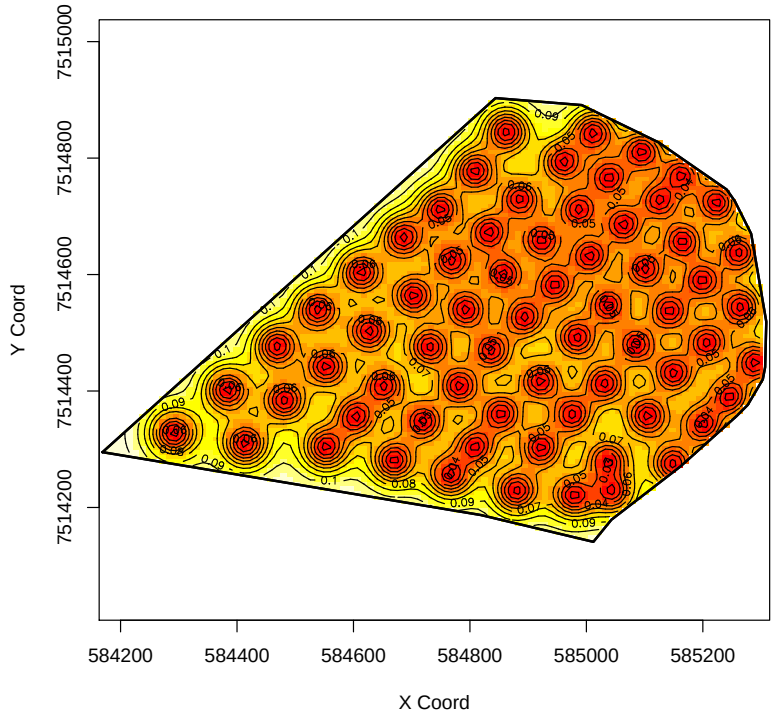
Tabela 8: Critério de informação de Akaike

Com os resultados acima, tem-se que todos os modelos com tendência a coordenada X foram melhores, pois obtiveram menor AIC, do que os modelos sem tendência na média, logo, o modelo com tendência na coordenada X e κ igual a 2 é o que melhor se ajustou aos dados, sendo assim, o próximo passo é fazer a predição ou krigagem para todo o espaço da fazenda, sendo assim, com a estimação dos parâmetros feita, é possível prever o valor do campo aleatório para localizações não amostradas, essa predição é feita através da média estimada para a localização ponderada pelos valores estimados para a variância e covariância do campo aleatório. Segue o gráfico da krigagem:



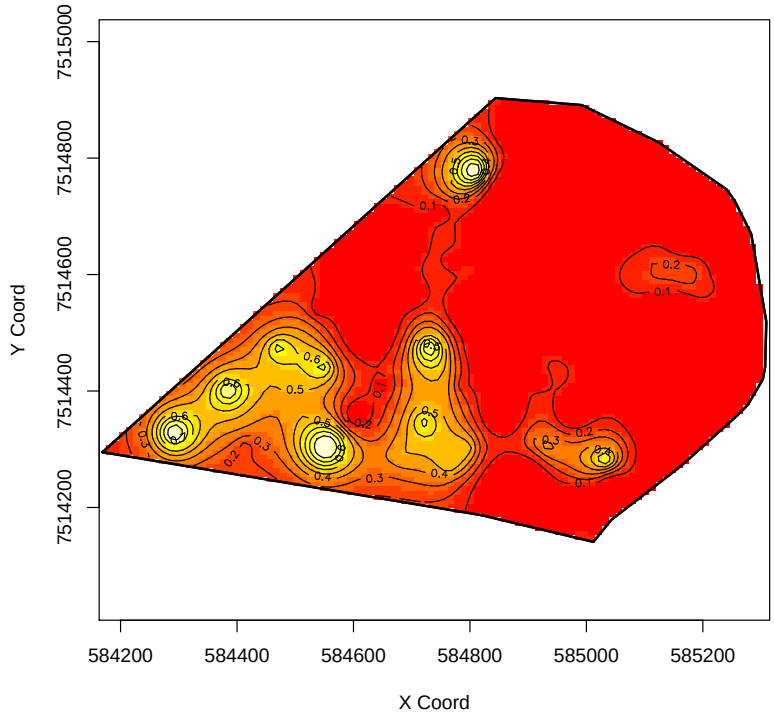
No gráfico acima as cores mais próximas do branco indicam valores mais elevados para a saturação por bases e conseqüentemente cores próximas do vermelho indicam valores menores para a saturação. Analisando os valores observados nas localizações amostradas, tem-se que as predições se aproximaram refletiram os valores reais e suavizou para o restante do espaço o campo aleatório.

O próximo passo é analisar o gráfico das variâncias das estimativas do campo aleatório, segue o gráfico:



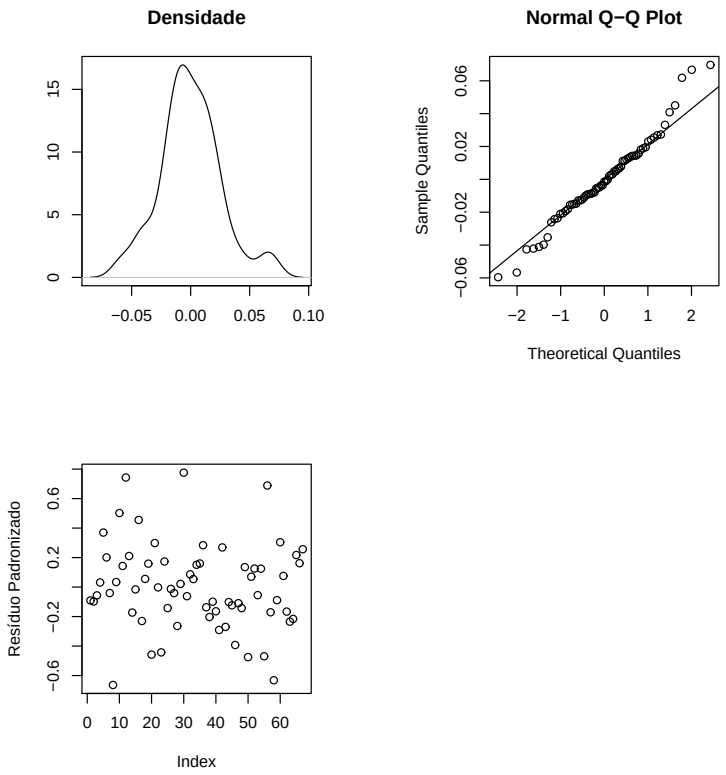
Com o gráfico acima tem-se que existe o padrão espacial esperado para as variâncias, ou seja, as variâncias mudam no espaço todo, mas essa mudança é suave, o que de certa forma corrobora a idéia de estacionariedade da função de covariância, além disso fica bem claro que valores de variância onde foi feita amostra são menores do que localizações não amostradas.

Com a definição dos parâmetros estimados, e com os valores preditos para cada posição do grid definido, é possível fazer análises a partir da distribuição normal de cada posição onde foi feita predição, assim é possível calcular a probabilidade prevista de que a saturação por bases seja menor que 4.5 em cada posição predita, ou seja, se pode fazer um gráfico das probabilidades de cada valor predito ser inferior a 4.5, que é um valor abaixo do esperado para uma boa qualidade do solo da fazenda sob estudo, abaixo segue o gráfico:



Com o gráfico acima tem-se que valores preditos em regiões com menor valor para a coordenada X possuem maior probabilidade do ph ser inferior a 4.5, informação que corrobora a análise descritiva onde se tem a clara impressão de que a média da saturação é inferior em regiões com menores coordenadas X.

Uma outra análise que pode ser conduzida é a validação do pressuposto de normalidade univariada do ruído branco, abaixo seguem análises gráficas do tipo:



O primeiro gráfico acima mostram que a densidade dos resíduos brancos estimados se aproximam de uma distribuição normal, além disso o qqplot mostra que existe uma pequena fuga da normalidade na cauda superior dos resíduos, mas nada que afete o pressuposto e por último não existe valores para os resíduos padronizados fora do intervalo de [-1,1], o que indica não existir valores discrepantes dos erros brancos. Após a análise gráfica o teste de normalidade de Shapiro-Wilk foi conduzido, o qual gerou um p-valor de 0.2185, logo a normalidade não é rejeitada e esse pressuposto está atendido.