Revista Brasileira de Cartografia (2016), Nº 68/4, Edição Especial Geoinformação e Análise Espacial: 745-758
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto
ISSN: 1808-0936

INFLUÊNCIA DO DELINEAMENTO AMOSTRAL NA INFERÊNCIA ESPACIAL POR GEOESTATÍSTICA APLICADA A DADOS DE CLOROFILA-*A* ADQUIRIDOS EM TRANSECTOS

Influence of Sample Design for Spatial Inference by Geostatistics Applied to Chlorophyll-A Data Acquired in Transects

Gabrielle Gomes dos Santos Ribeiro[1], Vilma Mayumi Tachibana[1] & Maria de Lourdes Bueno Trindade Galo[1]

[1]Universidade Estadual Paulista – UNESP
Programa de Pós-graduação em Ciências Cartográficas
R. Roberto Simonsen, 305, 19060-900 - Presidente Prudente, SP - Brasil
gabyy_14pp@hotmail.com, vilma@fct.unesp.br, mlourdes@fct.unesp.br

Recebido em 5 de Agosto, 2015/ Aceito em 11 de Janeiro, 2016
Received on August 5, 2015/ Accepted on January 11, 2016

RESUMO

A disposição dos elementos amostrais na área de estudo e sua influência nos resultados de análises espaciais é algo que vem sendo discutido frequentemente por pesquisadores da área de geociências, já que a qualidade de uma inferência espacial vai depender do tamanho da amostra e da distribuição espacial dos pontos amostrais. Nesse sentido, este trabalho tem o objetivo de analisar o impacto que diferentes delineamentos amostrais podem causar nos resultados da inferência espacial por Krigagem Ordinária. Para isso, primeiramente utilizou-se um conjunto de dados coletado em forma de transectos em uma parte do Reservatório de Nova Avanhandava, composto por 978 observações. Esse conjunto foi submetido a reduções sistemáticas no número de elementos amostrais, com o intuito de analisar o efeito do tamanho da amostra no resultado da inferência espacial. Em seguida, para analisar os resultados por diferentes tipos de amostras, simulou-se um grande conjunto de dados à partir da Krigagem Ordinária, utilizando o conjunto de dados em transecto. Nesse conjunto, aplicaram-se as técnicas de Amostragem Simples, Amostragem Sistemática e Amostragem Estratificada para obter variados tipos e tamanhos de amostras. A análise espacial foi realizada através do processo de Krigagem Ordinária, possibilitando obter diversos mapeamentos da variável clorofila-*a* na região de interesse. A validação da inferência foi realizada pela análise comparativa do Erro Quadrático Médio e do Índice Kappa. Os resultados mostraram que uma redução gradativa no tamanho da amostra original dos dados em transectos não causou alterações no resultado da inferência. Na análise dos diferentes métodos de amostragem, verificou-se que as amostras obtidas pela técnica de Amostragem Sistemática foram as mais eficazes ao mapear a variável clorofila-*a* por Krigagem Ordinária.

Palavras-chaves: Delineamento Amostral, Dados Espaciais, Transectos, Krigagem.

ABSTRACT

The arrangement of sampling units in the study area and its influence on the results of spatial analysis have been frequently discussed by researchers of the geoscience area, since the quality of a spatial inference will depend on sample size and spatial distribution of sample points. In this sense, this work aims to analyze the impact that different sampling designs may cause in the results of spatial inference by ordinary kriging. For this, first we used a data set collected in the form of transects in a section of Nova Avanhandava Reservoir, consisting of 978 observations. This set was

subjected to systematics reductions in the number of sampling units, with the aim to analyze the effect of sample size in the results of spatial inferences. Then, it was simulated up a dense amount of data by ordinary kriging, using the original data set obtained in transects. In this set, were applied Simple Sampling, Systematic Sampling and Stratified Sampling techniques and were obtained various types and sizes of samples. The spatial analysis was performed through the process of Ordinary Kriging, allowing obtaining mappings of the variable chlorophyll-a in the region of interest. The validation of inference was performed by comparative analysis of the Mean Error Squared and the Kappa Index. The results showed that the gradual reduction in the size of the sample of data on transects did not cause changes in the results of inferences. In addition, in the analysis of different sampling methods, it was found that the samples obtained by systematic sampling technique were the most effective in mapping the variable chlorophyll-a by ordinary kriging.

Keywords: Sampling Designs, Spatial Data, Transects, Kriging.

1. INTRODUÇÃO

Em muitas abordagens de análise espacial é necessário coletar uma quantidade considerável de dados georreferenciados para produzir um mapeamento da região de estudo e, dependendo do tamanho e da localização da região em questão, a aquisição das informações pode demandar tempo e recursos financeiros consideráveis. Além disso, situações que requerem predição ou estimativa da quantidade total ou concentração de uma ou mais variáveis de interesse em uma região de estudo, demandam um planejamento cuidadoso na definição de um delineamento amostral que garanta a qualidade da inferência espacial. Segundo Yamamoto e Landim (2013), essa qualidade vai depender do tamanho da amostra e da distribuição espacial dos elementos amostrais na área de interesse. Assim, o plano de amostragem adotado, além de gerar uma amostra representativa que forneça boas estimativas, deve ser operacionalmente viável.

A definição de um delineamento amostral que seja adequado para a coleta de dados espaciais é algo que tem sido estudado em vários trabalhos. Para propósitos gerais Wang *et al.* (2012) comentam que os tipos de amostragem frequentemente utilizados em aplicações práticas são: amostragem por grade, por transecto, sequencial e aninhada. Já Englund (1988) e Yamamoto e Landim (2013) mostraram que na presença de autocorrelação espacial, a amostragem sistemática produz resultados mais eficientes. Para autores como Haining (1990), independentemente do método de amostragem escolhido, é necessário certo cuidado ao trabalhar com esse tipo de dados já que são correlacionados espacialmente e não são provenientes de populações independentes e identicamente distribuídas (i.i.d.), ou seja, violam a suposição de independência.

Dentre os trabalhos que estudam delineamentos amostrais para coleta *in situ* de parâmetros de qualidade da água está o trabalho de Samizava *et al.* (2008), que propôs duas metodologias diferentes para delineamento amostral em planícies de inundação (lagoas e rios): a primeira baseada em análise de agrupamentos em dados limnológicos para definir as lagoas com características limnológicas semelhantes e a segunda baseada em imagens multiespectrais do sensor TM/Landsat, que possibilitou gerar regiões espectralmente homogêneas (estratos) para a aplicação de uma amostragem aleatória estratificada. Ambas as abordagens utilizadas mostraram-se adequadas no reconhecimento das observações que apresentam padrões de homogeneidade no espaço.

Outro fator importante de ser estudado é o tamanho da amostra, sabe-se que amostras maiores geram estimativas mais precisas, porém, erros menores têm seu custo: maiores complexidades, mais equipes, mais equipamentos, mais tempo de trabalho em campo, etc. Isso implica custos financeiros mais elevados. Por outro lado, deve-se levar em conta que os resultados extraídos de amostras menores estão sujeitos a grandes variabilidades, transmitem pouca confiança e, portanto, não são consideradas adequadas para tomadas de decisões estratégicas. Portanto, o ideal seria encontrar um ponto de equilíbrio entre o erro permitido pelo pesquisador e a precisão requerida nos resultados.

O plano de amostragem adotado, além de gerar uma amostra representativa que forneça boas estimativas, deve ser operacionalmente viável, ou seja, sua aplicação em campo deve ser efetiva. Para isso, o delineamento escolhido deve fornecer uma disposição dos elementos amostrais na área de estudo que respeite alguns

fatores importantes no momento da coleta, como: o difícil acesso a algumas áreas da região, acidentes geográficos, o tamanho da amostra relacionado ao tempo disponível para a coleta, a distância entre os pontos, etc. Mas, mesmo tomando todos os cuidados possíveis, sabe-se que o pesquisador das Ciências Naturais sempre encontrará algumas adversidades em campo, e por isso adaptações no esquema amostral quase sempre são necessárias.

Wang *et al.* (2012) comentam que os métodos de maior custo-benefício são: amostragem estratificada, amostragem por conglomerado, amostragem em dois estágios e amostragem sequencial.

Devido à complexidade da definição do esquema de amostragem, da coleta de dados, neste trabalho considerou-se uma amostra densa de dados, com mais de 900 observações adquiridas em transecto e coletadas por Cicerelli (2013), e a variável a ser analisada foi clorofila – *a*. Algumas vezes, observações tomadas a uma distância muito pequena podem dificultar a análise, principalmente na análise espacial em que o efeito "pepita" pode indicar variação nos dados, que na realidade são aleatórias não controláveis. Por isso, resolveu-se realizar reduções no tamanho desse conjunto de dados, obtendo várias amostras através da amostragem sistemática, a fim de responder as seguintes perguntas: É possível utilizar apenas parte dessa amostra densa sem prejudicar a análise? E quanto é possível reduzir?

Para fazer afirmações sobre um universo, a partir de uma amostra, a inferência espacial foi aplicada através da técnica de Krigagem Ordinária**,** com a criação de uma função matemática que forneça informações sobre pontos não coletados referentes à concentração de clorofila-*a* em uma pequena área do Reservatório de Nova Avanhandava, possibilitando gerar uma representação da distribuição espacial dessa variável.

Com o objetivo de estudar diferentes delineamentos amostrais propostos neste estudo com diferentes tamanhos de amostra e analisar qual a melhor disposição espacial das amostras para realizar a inferência por krigagem, a área original de estudo de Cicerelli (2013) foi reduzida e simulou-se através da Krigagem ordinária a distribuição da concentração da clorofila-*a* nesta área, formada por uma malha com 7384 células de 30 x 30 metros.

Burrough e McDonnel (1998) identificam algumas questões que devem ser consideradas quando se pretende obter mapas interpolados: o número de pontos da amostra, a distribuição dos mesmos, o tamanho da área de estudo, topologia da vizinhança das amostras e o método matemático a ser usado. Por isso, este trabalho pretende contribuir na discussão de: quanto é possível diminuir uma amostra em transecto sem afetar a qualidade dos resultados, qual o melhor delineamento amostral para aplicação da krigagem ordinária e como o tamanho da amostra pode influenciar nos resultados da inferência.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma revisão dos principais temas abordados nessa pesquisa.

2.1 Geoestatística

Os métodos geoestatísticos, ou simplesmente geoestatística, tiveram origem a partir de estudos desenvolvidos na França no final da década de 50 e início da década de 60, baseados em dados referentes às atividades mineradoras na África do Sul (ANDRIOTTI, 2003).

A geoestatística tem por objetivo a caracterização espacial de uma variável de interesse por meio do estudo da sua variabilidade espacial, permitindo o mapeamento, a quantificação e a modelagem dessa variável, através da interpolação dos pontos amostrados no espaço.

Segundo Webster e Oliver (2007), a geoestatística permite a estimativa de valores em locais não amostrados, de modo que não haja tendência e com um erro mínimo. Assim, pode-se lidar com propriedades que variam de modo não sistemático e em diferentes escalas.

Landim (1998) define a geoestatística como um ramo da estatística que trata de problemas referentes às variáveis regionalizadas (variáveis distribuídas no espaço ou tempo) que possuem características tanto de variáveis verdadeiramente casuais quanto totalmente determinísticas. A variável tende a apresentar valores muito similares em dois pontos vizinhos e, à medida que os pontos começam a se distanciar, os valores estimados se tornam mais distintos. Nesse sentido, as principais características de

uma variável regionalizada são: localização, anisotropia (apresentam variações graduais em uma determinada direção e irregulares em outras direções) e continuidade (CAMARGO, 1997).

Segundo Burrough (1986), a variação espacial de uma variável regionalizada pode ser expressa pela soma de três componentes: a) uma componente estrutural, associada a um valor médio constante ou a uma tendência constante; b) uma componente aleatória, espacialmente correlacionada; e c) um ruído aleatório ou erro residual.

O semivariograma experimental é uma medida do grau de dependência espacial entre os elementos da amostra de uma variável, que pode ser representado por um gráfico da covariância ($\gamma(h)$) versus a distância (h), que pode ser visto na Figura 1. O seu padrão representa o que se espera que aconteça quando se utiliza dados de campo, isto é, que as diferenças dos valores obtidos pela variável Z nos pontos x_i e x_{i+h} ($Z(x_i) - Z(x_{i+h})$ decresçam à medida que a distância h entre eles diminui. Os parâmetros do semivariograma são: alcance (a - definido pelo intervalo no qual as amostras apresentam correlação espacial), patamar (C - valor do semivariograma correspondente ao seu alcance, a partir do qual a função se estabiliza e não existe mais dependência espacial entre as amostras), efeito pepita (C_0 - refere-se à descontinuidade na origem do semivariograma) e contribuição (C_1 - diferença entre o patamar (C) e o efeito pepita (C_0)).

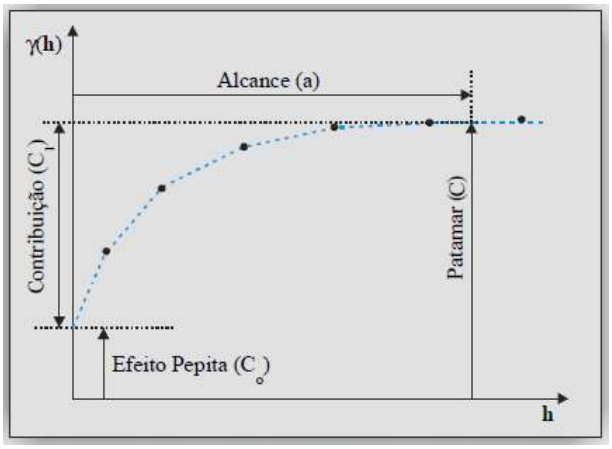

Fig. 1 - Semivariograma Experimental. Fonte: Camargo *et al.*, (2000, p.3.)

Embora existam diversos modelos de variogramas teóricos com patamar, alguns são mais comuns e podem explicar a maioria dos fenômenos espaciais. São eles: o exponencial, o esférico e o gaussiano. Após a construção dos semivariogramas experimentais em diferentes direções e do mapa variográfico, é possível verificar se o fenômeno espacial é isotrópico ou não. Quando os semivariogramas tiverem o mesmo comportamento, diz-se haver isotropia da variável, caso contrário tem-se anisotropia. Se constatada a anisotropia, devem-se selecionar quais são as direções de maior e menor alcance para que sejam modelados os semivariogramas relativos a essas duas direções, por meio de ajuste de um modelo teórico.

As vantagens da geoestatística sobre outras técnicas convencionais de predição estão relacionadas com a possibilidade de: estudar a variabilidade espacial, a suavização, o desagrupamento de observações, a determinação da anisotropia, a precisão e a estimativa do erro (ANDRIOTTI, 2003).

2.2 Inferência por Krigagem

As propriedades naturais da superfície terrestre são espacialmente contínuas, por isso encontram-se algumas restrições ao tentar descrevê-las por meio de simples funções matemáticas, sendo necessárias funções numéricas ordinárias que assumem um valor definido em cada ponto no espaço. A Krigagem é um dos modelos inferenciais propostos com esse objetivo.

Segundo Thompson (2002), a Krigagem tem capacidade de produzir melhores estimativas em termos de interpolação, porque está embasada em duas premissas: não-tendenciosidade do estimador e variância mínima das estimativas.

Os interpoladores são funções matemáticas usadas na construção de superfícies contínuas a partir de um conjunto de pontos coletados (BURROUGH & MACDONNEL, 1998).

Na Krigagem, o procedimento é semelhante ao de interpolação por média móvel ponderada, exceto que, nesse caso, recorre-se ao semivariograma para encontrar pesos ótimos a serem associados às amostras, para a estimativa de um valor para um determinado ponto (CÂMARA *et al.*, 2004).

Existem diversos tipos de Krigagem, cada uma com suas especificidades, entre elas a simples, a universal, a ordinária e a Co-

Krigagem.

A Krigagem Ordinária é um método local de estimativa, cujo valor estimado em um ponto não amostrado resulta da combinação linear dos valores encontrados na vizinhança próxima (YAMAMOTO & LANDIM, 2013). O estimador da Krigagem Ordinária é:

$$Z(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \qquad (1)$$

em que os λ_i são os pesos e $Z(x_i)$ são os valores de Z para i = 1, 2, 3, ..., n.

Os pesos ótimos são calculados de modo que o estimador não seja tendencioso e a variância de estimativa seja mínima.

O sistema de Krigagem Ordinária também pode ser escrito em forma de notação matricial:

$$\mathrm{K}.\lambda = \mathrm{k} \rightarrow \lambda = \mathrm{K}^{-1}.\mathrm{k} \qquad (2)$$

em que K e k são as matrizes de covariâncias e λ o vetor dos pesos.

2.3 Amostragem

As vantagens dos métodos por amostragem em relação ao de contagem integral (censos) são, segundo Cochran (1977), menor custo na aquisição dos dados, maior rapidez e maior amplitude.

Os métodos de amostragem mais utilizados em análises estatísticas, quando se trata de dados espaciais ou não, são: Amostragem Aleatória Simples (AAS), Amostragem Aleatória Estratificada (AE) e Amostragem Sistemática (AS).

A Amostragem Aleatória Simples (AAS) é o método mais simples e estatisticamente consistente para a definição da posição dos elementos amostrais. Na AAS, uma amostra é escolhida de modo que cada elemento amostral tenha a mesma probabilidade de ser incluído na amostra.

A AAE usa informação existente sobre a população para dividi-la em grupos bem definidos, chamados estratos. Em cada um desses estratos é selecionado um elemento amostral, mediante um processo aleatório simples. Esse método de amostragem tem a vantagem de fornecer resultados com menor probabilidade de erro associada.

A aplicação da Amostragem Estratificada no espaço consiste em dividir a área de estudo em subáreas, de modo que essas subáreas tenham a maior homogeneidade possível no seu interior, mas que sejam heterogêneas entre si. Diferente de um esquema aleatório simples em que podem ocorrer algumas regiões não amostradas, esse esquema de amostragem assegura que todas as subáreas que compõe o local em estudo sejam amostradas. Para que seja feita a subdivisão na área é importante levar em consideração informações prévias da região.

As duas formas de amostragem aleatória, simples ou estratificada, não levam em consideração a continuidade e correlação espacial entre os pontos. Por isso, quando se realiza uma distribuição aleatória dos pontos de amostragem, dois pontos podem ser localizados muito próximos gerando informação redundante e desperdiçando os recursos empregados.

Finalmente, a Amostragem Sistemática constitui uma variação da amostragem aleatória, porém sua aplicação requer que a população seja ordenada de modo tal que cada um de seus elementos possa ser identificado pela sua posição.

Ela é recomendada para contornar os problemas de uma amostragem aleatória usando um esquema com distribuição sistemática dos pontos de amostragem (malha de amostragem) que, além de evitar a coleta de amostras em pontos muito próximos, apresenta as mesmas vantagens da subdivisão da área. Assim, a amostra sistemática oferece um melhor resultado, porém nem sempre é possível a sua obtenção, que depende de uma série de fatores, tais como: acesso, acidentes geográficos (rios, lagos, topografia), vegetação, etc. (YAMAMOTO & LANDIM, 2013).

O objetivo primário ao estabelecer um plano de amostragem é promover um levantamento de dados o mais representativo da área avaliada, considerando-se um custo de investigação já fixado, ou se possível que seja minimizado. O segundo objetivo é a adoção de um esquema de amostragem simples e eficiente, que facilite a análise dos dados e a sua implantação em campo. Um terceiro ponto se refere ao tamanho da amostra.

Não há dúvida de que uma amostra não representa perfeitamente uma população, por

isso, a definição do seu tamanho implica na aceitação de uma margem de erro (erro amostral) especificado pela diferença entre um resultado amostral e o verdadeiro resultado populacional (NETO, 2004).

Além disso, a determinação do tamanho da amostra deve considerar que amostras grandes acarretam desperdício de tempo e recursos e amostras muito pequenas podem levar a resultados não confiáveis. Por isso, a necessidade de se decidir qual o erro aceitável e o nível de confiança apropriado para cada aplicação.

2.4 Dados adquiridos em transectos

O método de coleta em transecto consiste em amostrar dados em fluxo contínuo. Os transectos utilizados devem ser distribuídos de forma que toda a área ou grande parte dela seja amostrada.

Transecto tanto pode estar associado a uma aquisição contínua de dados, com todo o comprimento da linha sendo amostrado, ou os elementos amostrais podem ser tomados em determinados pontos ao longo da linha. Para ambas as abordagens, o intervalo no qual os elementos amostrais são tomados depende do habitat em particular, bem como do esforço e do tempo que pode ser atribuído ao levantamento.

Esse método possui baixo custo operacional e permite a obtenção de um grande número de medidas. Porém, quando medidas pontuais são tomadas a intervalos muito grandes, informações relevantes podem não ser detectadas, obscurecendo padrões de zoneamento por falta de observações. Por outro lado, intervalos pequenos demais, podem ocasionar superabundância de dados (ou mesmo sobreposição de medidas), prejudicando a inferência. É importante certificar-se de que o intervalo escolhido não coincida com algum padrão da variável em estudo, para que a amostra não fique viciada, o que pode resultar em conclusões equivocadas sobre a região.

2.5 Validação da Inferência Espacial

Na inferência espacial de qualquer variável, é fundamental obter informações acerca da medida em que a previsão pode se desviar do valor real dessa variável, de modo a fornecer uma melhor percepção sobre qualidade da previsão. Indicações dessa qualidade, muito usadas para validar os resultados de uma inferência são o Erro Quadrático Médio (EQM) e o Índice de Concordância Kappa.

O erro quadrático médio (EQM) é um dos indicadores mais comuns de erro de previsão e que consiste na diferença entre o valor real e a previsão do valor:

$$e_i = Z_i - \widehat{Z}_i \tag{3}$$

sendo:
e_i = Erro da i-ésima observação.
Z_i = Valor real da i-ésima observação.
$\widehat{Z}_i$ = Previsão para a i-ésima observação.

O EQM é determinado somando os erros de previsão ao quadrado e dividindo pelo número de erros usados no cálculo. O erro quadrático médio pode ser expresso pela seguinte equação:

$$EQM = \frac{\sum_{i=1}^{n} {e_i}^2}{n} \tag{4}$$

Cohen (1960) propôs o Coeficiente Kappa que pode ser definido como uma medida de associação usada para descrever e testar o grau de concordância (confiabilidade e precisão) entre dois ou mais métodos de classificação. Ele é baseado no número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre os métodos. Para isso, é necessário que as respostas sejam classificadas por grupos específicos, caracterizados pela natureza da aplicação e do conjunto de dados disponível. Trata-se de uma importante medida para determinar o quão bem funciona uma aplicação de alguma medição ou de algum método de predição.

Para variáveis categóricas ordinais que consideram o fato da discordância estarmais ou menos próxima da diagonal principalna tabela de contingência, Cohen (1968) recomenda usar Kappa ponderado. Nesse caso, por exemplo, para um objeto da classe I, o Kappa ponderado dá mais peso a discordância em que o modelo classifica esse objeto como classe IV do que à discordância em que o modelo o classifica como classe II. Duas medidas Kappa ponderadas independentes podem ser testadas para ver se elas são significativamente diferentes. Os detalhes

dos testes podem ser encontrados em Fleiss *et al.* (1969) e Congalton e Green (2009).

Landis e Koch (1977) sugeriram os intervalos de valores expressos na Tabela 1 para a interpretação dos valores do Índice de Concordância Kappa.

Tabela 1: Interpretação dos valores do Índice de Concordância Kappa.

Valores de Kappa	Interpretação
<0	Sem concordância
0 – 0,19	Concordância pobre
0,20 – 0,39	Concordância regular
0,40 – 0,59	Concordância Moderada
0,60 – 0,79	Concordância Satisfatória
0,80 - 1	Concordância Excelente

Fonte: Landis e Koch (1977).

3. MATERIAIS E MÉTODO

Este capítulo contém uma descrição completa dos materiais utilizados e da metodologia adotada neste trabalho.

3.1 Dados da área de estudo

O conjunto de dados utilizados nesta investigação refere-se a medidas da concentração de clorofila-*a* obtidas no Reservatório de Nova Avanhandava, localizado no rio Tietê, Estado de São Paulo, cuja localização e forma geral são mostradas na Figura 2.

Os dados de clorofila-*a* foram adquiridos em um levantamento fluorimétrico de campo realizado no dia 19 de Setembro de 2012, conforme descrito por Cicerelli (2013). Paraa coleta dos dados foi utilizado um fluorômetro de campo 10AU™ (Turner Designs Inc. – Sunnyvale, CA, EUA) operando de forma contínua, em transectos transversais (margem a margem), de modo a garantir uma ampla representação da área. No total, foram medidas as concentrações de clorofila-*a* em 978 pontos amostrais georreferenciados (CICERELLI, 2013).

No processamento e análise dos dados foram utilizados os aplicativos: o *software* estatístico *R* (R Core-development Team), *SAS* (SAS Institute Inc.) e o aplicativo *VARIOWIN* (PANNATIER), usado nas análises geoestatísticas.

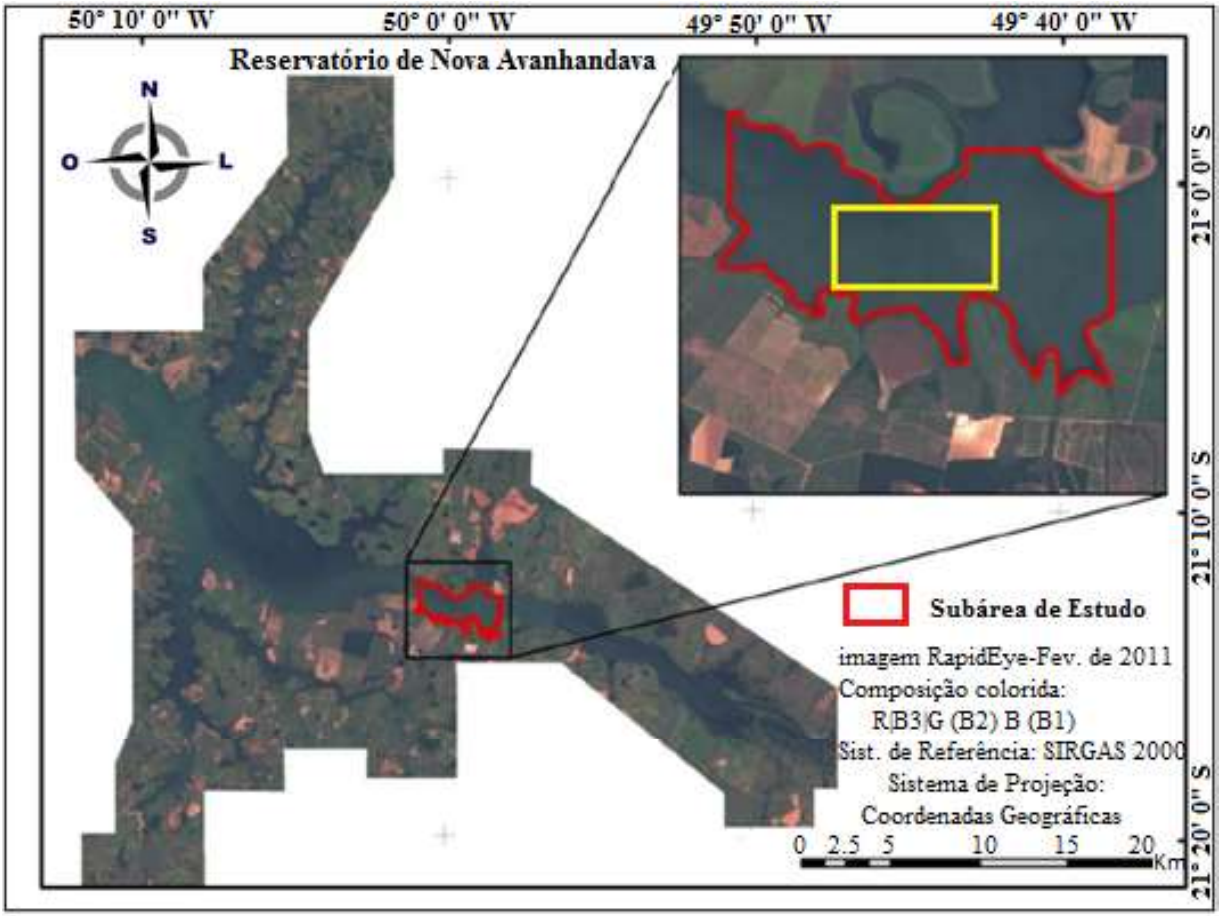

Fig. 2 – Localização e formato da área de estudo. Fonte: Cicerelli (2013).

3.2 Desenvolvimento metodológico

Considerando o propósito de avaliar a influência do delineamento amostral na inferência espacial por Krigagem foram desenvolvidas duas abordagens para tratamento e análise da amostra original de 978 medidas da concentração de clorofila-*a*, obtidas em transectos realizados em uma parte do reservatório de Nova Avanhandava. Na primeira, a amostra original foi dividida em dois conjuntos (um para inferência e outro para validação) e submetida a reduções sucessivas no número de elementos amostrais, realização das inferências *e análise* do erro com base na amostra de validação. Uma segunda abordagem foi aplicada, considerando dados simulados gerados a partir da amostra inicial, utilizando diferentes métodos para a aquisi*ção* de amostras desses dados.

3.2.1 Reduções sucessivas do tamanho da amostra obtida em transectos

A distribuição espacial dos 978 elementos amostrais obtidos na forma de transectos é mostrada na Figura 3. Essa configuração indica a estratégia adotada de realizar transectos transversalmente ao corpo d'água e a dificuldade em manter um deslocamento uniforme do barco, sujeito à ação do vento e do próprio meio aquático, para assegurar uma dispersão mais adequada dos elementos amostrais na área de estudo.

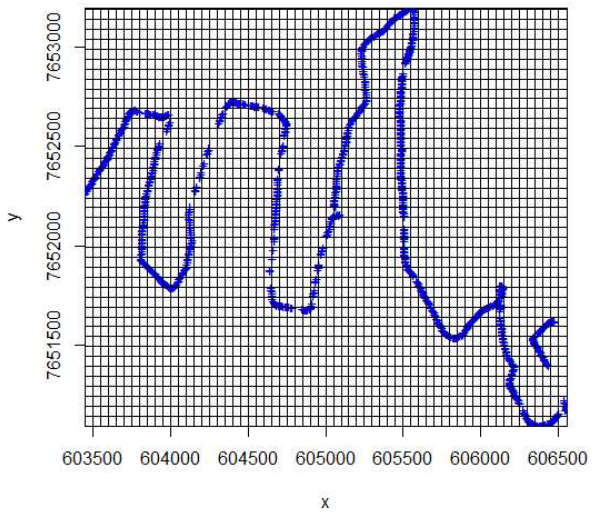

Fig. 3 - Configuração espacial do conjunto de dados.

Previamente à redução da amostra para analisar sua influência no processo de inferência espacial da concentração de clorofila-*a*, foi separada uma amostra aleatória de 200 elementos, equivalente a aproximadamente 20% do conjunto total, para validação das inferências. Com o conjunto remanescente de 778 observações, iniciou-se o processo de retirada sistemática de pontos, com o intuito de reduzir ao máximo esse conjunto de dados, sem alteração significativa no formato característico dos transectos iniciais e sem que a qualidade da inferência fosse afetada.

Foram realizadas quatro reduções sucessivas no número de elementos amostrais dos transectos, definindo tamanhos de amostra de 400, 300, 200 e 100 observações. Foi analisada a distribuição espacial de cada amostra e, para cada uma delas realizou-se o processo de inferência por Krigagem Ordinária. Com isso, foi possível o mapeamento da clorofila-*a* com as quatro amostras de tamanho diferentes.

A validação das inferências, realizada com base na amostra definida previamente para esse fim (200 elementos amostrais),foi feita comparando pontualmente os valores interpolados com as concentrações reais de clorofila-*a*, nos mesmos pontos amostrais georreferenciados, por meio do cálculo do Erro Quadrático Médio (EQM) e do Índice Kappa.

3.2.2 Experimento com dados simulados

A fim de avaliar a influência da distribuição espacial dos elementos amostrais na inferência, foi gerada uma superfície de referência para a concentração de clorofila-*a* e sobre ela foram aplicados os diferentes delineamentos amostrais e adquiridos os valores da variável em cada ponto especificado. A superfície de referência foi gerada sobre grade regular, com resolução de 30 metros, definindo um retângulo de 71 linhas por 104 colunas. Aplicou-se a Krigagem Ordinária, utilizando todos os 978 pontos amostrais, para estimar as 7384 predições de concentração de clorofila-*a* (para cada uma das células da grade regular). Uma parte dos dados foi separada para realizar a validação do processo de inferência espacial, por meio da seleção aleatória de 1.500 predições, equivalente à aproximadamente 20% dos dados. Ao conjunto remanescente de predições (5884) foram aplicados os três diferentes métodos de amostragem: AAS, na qual o os elementos amostrais foram selecionados por sorteio; AAE e AS. Para cada método foram consideradas amostras com 1000, 400, 300, 200 e 100 observações.

No caso da Amostragem Estratificada, os estratos foram definidos sobre o mapa resultante da Krigagem Ordinária (Figura 4) nos três intervalos de valores demarcados no mapa: valores baixos de concentração de clorofila-*a*, entre 5 a 8 µg L^{-1} (indicados na cor verde), valores intermediários, entre 8 a 10 µg L^{-1} (cor amarela) e valores altos, de 10 a 14 µg L^{-1} (cor vermelha). Com isso, utilizando o *software* R e com o auxílio da imagem, foi possível delimitar a área de cada estrato e assim definir em quais estratos os pontos cairiam de acordo com suas localizações geográficas.

Posteriormente, a amostragem foi realizada proporcionalmente à quantidade de observações dentro de cada estrato.

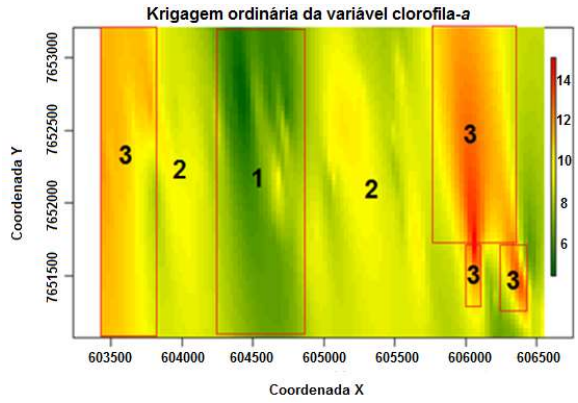

Fig. 4 – Estratos considerados na AAE.

Após a seleção dos elementos da amostra, realizou-se o processo de inferência espacial da clorofila-*a* usando Krigagem Ordinária. Os resultados foram validados e comparados por meio do EQM e do Índice Kappa, a fim de avaliar o efeito causado nos resultados da inferência ao adotar amostras de tamanhos diferentes.

4. RESULTADOS E DISCUSSÃO

O presente capítulo trata da apresentação e discussão dos resultados.

4.1 Dados em transecto

A fim de comparar as estimativas de média e variância de cada amostra de tamanho reduzido em relação ao conjunto total de pontos georreferenciados que compõem os transectos, foi elaborada a Tabela 2 que indica, para cada amostra de tamanho reduzido (n), a concentração média de clorofila-*a* (µg L^{-1}), o coeficiente de redução no tamanho da amostra (k), dado pela razão entre o número de elementos amostrais original (778) e o número de elementos da amostra reduzida, além da variância da amostra (µg L^{-1}).

Tabela 2: Indicadores estatísticos da concentração de clorofila-*a* obtidos para as amostras de tamanho reduzido

Número de elementos	Média (µg L^{-1})	k (coeficiente de redução da amostra)	Variância (µg L^{-1})
400	8,895	1,945	4,509
300	8,919	2,593	4,418
200	8,870	3,890	4,025
100	8,846	7,789	4,296

Considerando que a média e variância da concentração de clorofila-*a*, obtidos para o conjunto original de 778 medidas (excluindo os dados de validação), eram 8,891µg L^{-1} e 4,608µg L^{-1}, respectivamente, verifica-se que a redução do conjunto original de dados em amostras menores não afetou as propriedades estatísticas básicas dos dados em qualquer um dos casos.

Na Figura 5 apresenta-se o mapa resultante da Krigagem Ordinária para as diferentes amostras de tamanho reduzido, indicados pelas letras (a), (b), (c) e (d) para 400, 300, 200 e 100 elementos amostrais, respectivamente. A escala de cores à direita dos mapas refere-se às concentrações de clorofila-*a* que assumem valores menores que 6µg L^{-1} (tons esverdeados) até níveis superiores a 12µg L^{-1} (tons avermelhados).

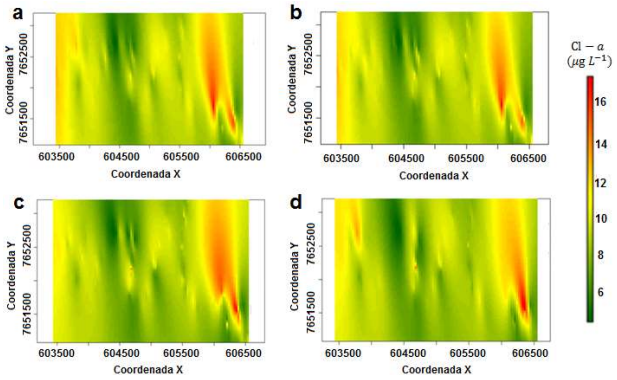

Fig. 5 – Krigagem Ordinária da clorofila-*a* (µg L^{-1}) para as amostras de tamanhos 400 (a), 300 (b), 200 (c) e 100 (d).

Ao comparar visualmente os quatro mapas da Figura 5, verifica-se que a redução no número de elementos amostrais dos dados originais obtidos por transecto parece não afetar os resultados da Krigagem Ordinária, uma vez que a distribuição espacial da concentração de

clorofila-*a* se mantém quase similar nas quatro representações.

De modo geral, observam-se poucas áreas em tons mais escuros de vermelho, que são as regiões em que ocorrem concentrações de clorofila-*a* próximas a 14µg L^{-1}. Os níveis de concentração predominantes estão entre 5 e 10µg L^{-1}, representados no mapa pelas regiões com coloração entre verde escuro e amarelo.

Uma segunda visualização da configuração da dispersão espacial dos níveis de concentração de clorofila-*a* na área de estudo foi gerada pelo fatiamento dos valores preditos da variável, em intervalos de 2µg L^{-1}, produzindo cinco classes de saída.

Na Figura 6 apresentam-se os resultados do fatiamento aplicado às estimativas obtidas pela Krigagem Ordinária, para as amostras de tamanhos gradativamente reduzidos. Novamente, as letras (a), (b), (c) e (d) são usadas para indicar 400, 300, 200 e 100 elementos amostrais, respectivamente, porém a escala de cores refere-se agora aos intervalos de valores fatiados de clorofila-*a*. Nessa visualização percebe-se melhor que, à medida que o tamanho da amostra é reduzido, as variações na concentração tendem a ficar suavizadas.

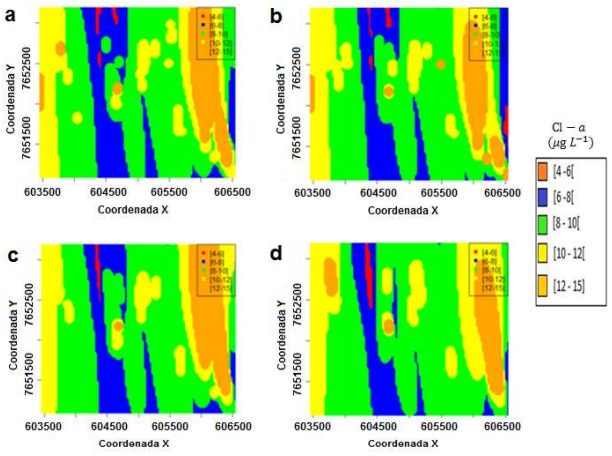

Fig. 6 – Mapa de classes das estimativas de concentração de clorofila-*a* (µg L^{-1}) para as amostras de tamanho 400 (a), 300 (b), 200 (c) e 100 (d).

As estimativas do Erro Quadrático Médio (EQM) e Índice Kappa, obtidas a partir da comparação entre dados de validação e valores inferidos pela Krigagem Ordinária nas mesmas posições, para cada tamanho de amostra, são mostrados na Tabela 3.

Tabela 3: Erro Quadrático Médio e Índice Kappa estimados para os mapas resultantes das amostras reduzidas

	400	**300**	**200**	**100**
EQM	0,92	1,06	1,30	2,55
Kappa	0,68	0,66	0,58	0,55

Conforme esperado, a redução gradativa no tamanho da amostra implica em aumento também gradativo no valor do EQM e diminuição do Índice Kappa. Porém, a amostra de tamanho 400 obteve uma leve vantagem em relação às outras amostras.

A partir da redução do conjunto, percebeu-se também que talvez uma amostra menor seria suficiente para representar o conjunto de dados e alcançar os objetivos desejados, uma vez que ao comparar as médias e as variâncias de todas as amostras com as da população, seus valores eram muito semelhantes.

De forma geral, os valores de Índice Kappa resultaram em um grau de concordância apenas moderado para as amostras de tamanho 100 e 200 e satisfatória com tamanho 300 e 400. Esses resultados são decorrentes do fato de que na coleta de dados em transecto há uma grande densidade de medidas coletadas ao longo dos transectos e nenhuma medida entre eles. Com isso, mesmo que haja um cuidado no planejamento da coleta de dados, ocorrem muitos espaços sem observações, fazendo com que a variância do erro de estimação seja alta nessas regiões.

4.2 Dados simulados

Conforme especificado na seção 3.2, foram obtidas cinco amostras de tamanhos diferentes para cada tipo de amostragem (aleatória, estratificada e sistemática). Nas figuras a seguir são mostrados os mapas resultantes da Krigagem Ordinária aplicada aos conjuntos de dados resultantes da amostragem aleatória (Figura 7), estratificada (Figura 8) e sistemática (Figura 9) para as diferentes amostras de tamanho reduzido, indicados pelas letras (a), (b), (c), (d) e (e) para 1000, 400, 300, 200 e 100 elementos amostrais, respectivamente. Também nessas representações, a escala de cores à direita dos mapas refere-se às concentrações de clorofila-*a* que assumem valores de 6µg L^{-1} (tons esverdeados) até 14µg L^{-1} (tons avermelhados).

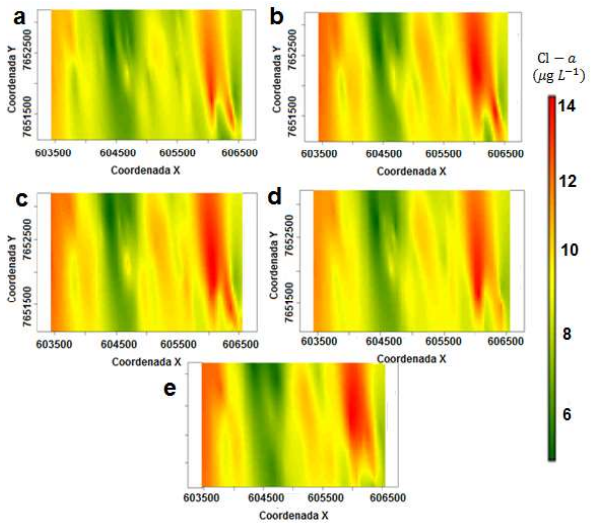

Fig. 7 – Krigagem Ordinária da clorofila-*a* (μg L^{-1}) aplicada aos dados da AAS contendo 1000 (a), 400 (b), 300 (c), 200 (d) e 100 (e) elementos amostrais.

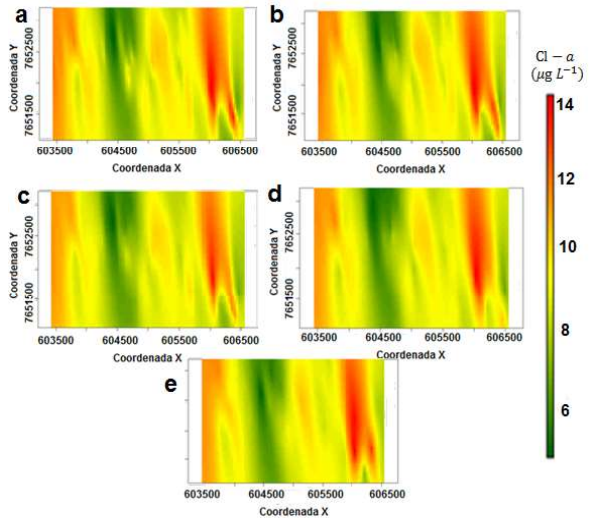

Fig. 8 – Krigagem Ordinária da clorofila-*a* (μg L^{-1}) aplicada aos dados da AAE contendo 1000 (a), 400 (b), 300 (c), 200 (d) e 100 (e) elementos amostrais.

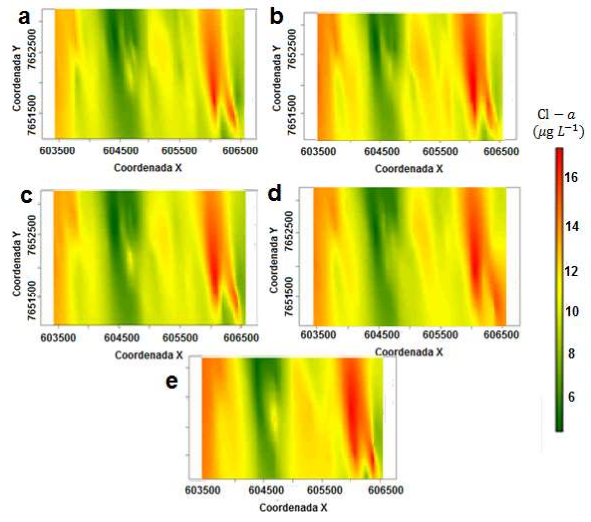

Fig. 9 – Krigagem Ordinária da clorofila-*a* (μg L^{-1}) aplicada aos dados da AS contendo 1000 (a), 400 (b), 300 (c), 200 (d) e 100 (e) elementos amostrais.

Visualmente, o resultado da inferência espacial para os três tipos de amostragem adotados apresentam similaridade em termos de distribuição espacial das variações na concentração de clorofila-*a*. As estimativas do Erro Quadrático Médio (EQM) e Índice Kappa, obtidas a partir da comparação, ponto a ponto, entre dados de validação ("assumidos como verdade terrestre") e valores inferidos nas mesmas posições para cada tamanho de amostra, são mostrados na Tabela 4.

De acordo com a interpretação usual dos valores de Índice Kappa (Tabela 1), os resultados foram considerados excelentes para todos os tamanhos de amostras, e os três métodos de amostragem estudados.

Tabela 4: Estimativas do Erro Quadrático Médio (EQM) e Índice Kappa para os diferentes tipos de amostragem e tamanho de amostra

Tipo de amostragem	Tamanho da amostra	EQM	Índice Kappa
Aleatória (AA)	1000	0,008	0,96
	400	0,059	0,93
	300	0,062	0,92
	200	0,084	0,90
	100	0,239	0,80
Estratificada (AE)	1000	0,011	0,97
	400	0,052	0,95
	300	0,061	0,94
	200	0,127	0,87
	100	0,185	0,86
Sistemática (AS)	1000	0,010	0,98
	400	0,038	0,94
	300	0,042	0,93
	200	0,076	0,89
	100	0,170	0,84

Pelas estimativas do EQM e Índice Kappa, todos os tipos de amostragemcom tamanho 1000 apresentaram resultados melhores em relação às amostras menores. O erro padrão assintótico (epa) do Kappa para amostras de tamanho 1000, obtido utilizando o *software SAS*, foi aproximadamente 0,0020 sendo melhor do que as demais amostras, no nível de significância de 95%. A medida de Kappa para amostras de tamanhos 400 e 300 são estatisticamente iguais e melhores que de amostras de tamanho 200, que tem epa igual a 0,036. Essa última amostra

apresenta epa maior que 0,0042, com a menor excelência entre os cinco diferentes tamanhos de amostra.

De modo geral, a Amostragem Sistemática forneceu os melhores resultados na inferência espacial ao definir o menor EQM e o maior Índice Kappa. Vários autores, entre eles Olea (1984) apud Englund (1988) e Yamamoto e Landim (2013), já confirmaram que na presença de autocorrelação espacial, a amostragem sistemática sobre uma malha produz resultados mais eficientes.

Como apresentado anteriormente, a redução no tamanho da amostranão foi considerada significativa, uma vez que para todos os tamanhos de amostras as inferências foram consideradas excelentes de acordo com o Índice Kappa.

Ao comparar os valores de EQM e Índice Kappa obtidos com os dados em transecto e com os dados simulados, nota-se que mesmo o pior resultado obtido com os dados simulados (o da Amostra Simples de tamanho 100, resultando em EQM=0,239 e Índice Kappa=0,80) foi superior ao resultado dos dados em transectos reduzidos a 400 elementos amostrais (EQM=0,92 e Índice Kappa=0,68.

5. CONCLUSÕES

Para os dados em transecto, a redução no tamanho da amostra evitou a utilização de amostras espacialmente sobrepostas, diminuindo a autocorrelação no conjunto de dados de entrada. Os valores médios da concentração de clorofila-*a*, assim como a variância, forneceram estimativas próximas àquelas obtidas a partir do conjunto original das medidas tomadas em transecto. Além disso, a distribuição espacial dos elementos amostrais manteve o formato do deslocamento do barco na aquisição das medidas em transectos, mesmo com a redução para 100 elementos amostrais.

A utilização de dados da predição da concentração de clorofila-*a* para obter valores em locais não amostrados e, a esses dados simulados, aplicar três diferentes métodos de amostragem, definiu melhores estimativas de EQM e Índice Kappa com a utilização da Amostragem Sistemática.

Em ambas as análises, tanto para dados em transecto como paraos dados simulados, observou-se que o tamanho da amostra é muito importante quando se deseja manter baixos os erros cometidos nas inferências, visto que, em termos de resultados do EQM e do Índice Kappa, as amostras maiores forneceram melhores resultados nas inferências espaciais.

Mesmo que a análise visual dos mapas resultantes da inferência por Krigagem apresentem uma aparência similar em todos os casos analisados, as estimativas do EQM e Índice Kappa mostram que a aquisição de dados pontuais, segundo um delineamento amostral prévio, e não em transectos, forneceram menores estimativos de erro e maiores índices Kappa.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRIOTTI, J. L. S. **Fundamentos de estatística e geoestatística**. São Leopoldo: Editora Unisinos, 2003, 165p.

BURROUGH, P. A. **Principles of geographical information systems for land resources assessment**. Oxford: Claredon Press, 1986, 194p.

BURROUGH, P. A.; McDONNELL, R. A. **Principles of geographical information systems: Spatial Information Systems and Geostatistics**. 2. ed. Oxford: Oxford University Press, 1998. 333p.

CÂMARA, G. FUCKS, S. D.; CÂMARA, G.; CAMARGO, E. C. G. **Capítulo 3: Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004 (ISBN: 85-7383-260-6). Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap3-superficies.pdf>. Acesso em: 12 jan. 2013.

CAMARGO, E. C. G. **Desenvolvimento, implementação e teste de procedimentos geoestatísticos (Krigeagem) no Sistema de Processamento de informações Georreferenciadas (SPRING)**. Dissertação de mestrado (Mestrado em Sensoriamento Remoto). Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 1997.124 p.

CICERELLI, R. E. **Estudo da ocorrência de cianobactérias em ambiente aquático continental por meio da inferência espacial do pigmento ficocianina.** 2013. 165 f. Tese

(Doutorado em Ciências Cartográficas) Faculdade de Ciência e Tecnologia - UNESP, Presidente Prudente.

COCHRAN, W. G. **Sampling Techiniques.** 3. ed. New York: John Wiley & Sons, 1977.428p.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement 20** (1): 1960. 37–46pp.

COHEN, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. **Psychological Bulletin**, 1969, 72,213-220pp.

CONGALTON, R. G., GREEN, K. **Assessing the Accuracy or Remotely Sensed Data: Principles and Practices**, 2. ed., Boca Raton, Fl: CRC/ Taylor & Francis. 2009. 183p.

ENGLUND, E. J. **Spatial Autocorrelation: Implications for sampling and estimation.** U.S. Environmental ProtectionAgency, Las Vegas, 1988. Disponível em: <www.epa.gov/esd/cmb/research/papers/ee107.pdf>. Acesso em: 06 jan. 2015.

FLEISS, J. L., COHEN, J., EVERITT, B.S. Large sample standard errors of kappa and weighted kappa. **Psychological Bulletin** 72, 1969, 323-327pp.

HAINING, R. **Spatial data analysis in the social and environmental sciences.** Cambridge: University Press, 1990, p. 432.

KOTZ, S.; JOHNSON, N. L.; READ, C. B. **Encyclopedia of statistical sciences**. New York: John Wiley & Sons, 1983, v. 4, p. 352.

LANDIM, P. M. B. **Análise estatística de dados geológicos**. São Paulo: Editora UNESP, 1998.226p.

LANDIS, J. R., KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, p.159-75, 1977.

NETO, J. C. P. C. **Determinação do tamanho da amostra.** 2004. Disponível em: <http://fesppr.br/~centropesq/Calculo_do_tamanho_da_amostra/Tamanho%20da%20Amostra%20-%201.pdf>. Acesso em: 15 dez. 2014.

OLEA, E. A. Sampling Design Optimization for Spatial Functions. **Mathematical Geology**, v.16, n.4, p. 369-389, 1984.

PANNATIER, Y. ***VARIOWIN*: *Software* for Spatial Data Analysis in 2D**. Springer-Verlag.1996.

PEREIRA, A. C. F.; GALO, M. L. B. T., VELINI, E. D.; NOVO, E. M. L. M. Amostragem em corpos d'água: definição de elementos amostrais, posicionamento e coleta de dados "in situ". In: **II Simpósio Brasileiro de Geomática**, 2007, Presidente Prudente. 866-874pp.

ROSSONI, D. F. **Análise de variância para experimentos com dependência espacial**. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, MG, Brasil, 2011. 108p.

SAMIZAVA, T. M.; IMAI, N. N.; ROTTA, L. H. S.; FERREIRA, M. S.; GALO, M. L. B. T.; ROCHA, R. R. A.; ENNES, R.. Proposta de delineamento amostral para levantamento de medidas de variáveis limnológicas e de dados espectrorradiométricos em planície de inundação. In: **II Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação**, 2008, Recife. **Anais.**

SANTOS, P. S.; EPIPHANIO, J. C. N. Aprimoramento do método de amostragem simples utilizado pelo Projeto Geosafras para estimativa municipal de área plantada com soja. In: Anais **XIII Simpósio Brasileiro de Sensoriamento Remoto**, 2007, Florianópolis.371-378pp.

SAS Institute Inc. 2012.Base SAS® 9.3 Procedures Guide: Statistical Procedures, Second Edition. Cary, NC: SAS Institute Inc.

SILVA, S. U. **AnáliseEstatística Espacial dos Óbitos por Câncer de Traquéia, Brônquios, Pulmão e Estômago Registrados para o Estado de São Paulo de 1995 a 2010**. 2011. Trabalho de Conclusão de Curso (Graduação em Estatística) – Faculdade de Ciência e Tecnologia, Universidade Estadual Paulista Júlio de Mesquita Filho, Presidente Prudente. 82p.

THOMPSON, S. K. **Sampling**. 2. ed. New York: Wiley, 2002.343p.

UEBERSAX, J. **Calculating Kappa with SAS®**. Disponível em <http://www.john-uebersax.com/

stat/saskappa.htm>. Acesso em: 20 nov. 2015

WANG, J. F., STEIN, A., GAO, B. B., & GE, Y. (2012). A review of spatial sampling. **Spatial Statistics**, *2*(1),1–14pp. http://doi.org/10.1016/j.spasta.2012.08.001

WEBSTER, R.; OLIVER, M. A. **Geostatistics for environmental scientists**. England: John Wiley & Sons, 2007.330p.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações.** 1. ed. São Paulo: Oficina de Textos, 2013.215 p.