Anais 7º Simpósio de Geotecnologias no Pantanal, Jardim, MS, 20 a 24 de outubro 2018
Embrapa Informática Agropecuária/INPE, p. 714-723

Uso de programas geoestatísticos no tratamento de grande volume de dados

Sérgio Aparecido Braga da Cruz [1]
Eduardo Antonio Speranza [1]
Inácio Henrique Yano [1]

[1] Embrapa Informática Agropecuária - CNPTIA
Av. André Toselo, 209 - Caixa Postal 6041
13083-886 - Campinas - SP, Brasil
{sergio.cruz,eduardo.speranza,inacio.yano}@cnptia.embrapa.br

Resumo. Métodos geoestatísticos para interpolação de dados demandam uma grande capacidade de processamento computacional. Este trabalho propõe uma estratégia para melhorar o desempenho deste método possibilitando a redução do tempo de execução ou aumento do volume de dados processados. A estratégia pode ser usada em programas geoestatísticos legados com esforço mínimo para adaptações e/ou modificações de seus formatos originais. Os resultados obtidos propiciam uma maior eficiência no uso e ajustes de métodos de interpolação de dados na região do Pantanal, onde a cobertura insuficiente de sensores e a dificuldade de acesso dificultam a coleta de dados.

Palavras-chave: big data, programas legados, geoestatística, krigagem

Abstract. Geostatistical methods for data interpolation require a large computational processing capacity. This work proposes a strategy to improve the performance of this method by reducing the execution time or increasing the volume of processed data. The strategy can be used in legacy geostatistical programs with minimal effort to adapt and / or modify their original formats. The results obtained allow a greater efficiency in the use and adjustments of data interpolation methods in the Pantanal region, where insufficient coverage of sensors and difficulty in access make data collection difficult.

Keywords: geostatistics, kriging, big data, legacy programs

1. Introdução

Métodos de interpolação geoestatística de dados são utilizados normalmente na inferência de dados faltantes necessários para execução de procedimentos de análise geoespacial e geoprocessamento. A inferência é necessária quando, por exemplo, existem poucos dispositivos sensores disponíveis, ou a resolução dos mesmos não é adequada para uma determinada análise. Dificuldades no acesso aos pontos de amostragem de dados podem inviabilizar a sua coleta e consequentemente também diminuir a quantidade de dados disponíveis para análise. As restrições de acessibilidade no Pantanal e a pouca disponibilidade de sensores tem justificado a aplicação de interpoladores para complementação de dados de uma série de variáveis sobre a região (Marcuzzo et al, 2011;Tieppo et al, 2010, Muñoz et al, 2013).

Por outro lado, a necessidade de análises geoespaciais de grandes áreas e a análise de pequenas áreas com nível de detalhamento maior como aqueles necessários na agricultura de precisão têm demandado a adoção de novas estratégias de processamento de dados para produção resultados de qualidade em tempo hábil. Além de novas estratégias de processamento, novos desafios têm sido encontrados na realização das atividades de coleta, organização, armazenamento de dados a partir do surgimento do *Big Data* (HEY, 2009). *Big Data*, pode ser visto como um conjunto de dados caracterizado pelo seu grande volume, sua alta velocidade de atualização e pela sua abrangência ou variedade de temas (Goodchild, 2013). No *Big Data* a grandeza destas características dificulta o seu tratamento por meio de técnicas e ferramentas tradicionais.

Desta forma, interpoladores geostatísticos podem ser utilizados sempre que a quantidade disponível de dados é insuficiente para suporte à análises geoespaciais e geoprocessamento nas escalas e nível de detalhamento desejados.

Porém, em muitos casos, já existe existe um grande investimento em recursos humanos e computacionais no desenvolvimento de soluções específicas tratando escalas menores de volumes de dados. A necessidade de aproveitar este código legado, com redução nos impactos na adoção de novas plataformas para suporte a alta capacidade de processamento é um requisito importante na adoção de soluções no *Big Data*.

Neste artigo descrevemos uma técnica baseada na virtualização de processos que permite a melhoria do desempenho na execução de programas geoestatísticos, com a redução do seu tempo de processamento. Programas previamente implementados podem ser adaptados para utilização neste novo contexto. Para ilustrar a viabilidades desta solução um estudo de caso foi realizado no processamento de interpolação usando o método de krigagem de dados geoespaciais do sensor *Sentinel* da região do Pantanal. Neste estudo de caso um *script* R inicialmente desenvolvido para uso em um computador *desktop* foi adaptado para execução em um ambiente de *data center*, que desta forma pode ser reutilizado para tratamento de um volume maior de dados.

2. Objetivo

Neste artigo é apresentada uma técnica de processamento de dados que melhora o desempenho da execução de programas geoestatísticos legados, possibilitando redução de tempo de execução e propiciando o seu uso no tratamento de grande volume de dados.

3. Material e Método

3.1 Virtualização de processos

O desenvolvimento de um programa de computador é realizado tendo como requisito uma configuração do ambiente computacional no qual o mesmo deverá ser executado. Esta configuração de ambiente define os pré-requisitos para instalação e execução do programa. Normalmente estes pré-requisitos abrangem o tipo e versão do sistema operacional no qual o programa poderá ser executado e indicam a necessidade de pré-instalação de outros programas do qual o programa principal depende. Estes pré-requisitos restringem o uso de um programa em ambientes diferentes para o qual sua execução não tenha sido planejada, uma vez que estes pré-requisitos estão ligados internamente a estrutura do programa implementado.

Para que um programa possa ser executado em ambientes para os quais ele não tenha sido planejado seria necessário que todas as suas dependências internas permanecessem satisfeitas no novo ambiente de execução. O uso de máquinas virtuais (Laureano, 2006) é uma abordagem utilizada para permitir a execução de programas em diferentes ambientes computacionais com garantia da manutenção destas dependências. Elas consistem em programas emuladores que simulam todas as funcionalidades de um computador em outro, possibilitando a substituição completa de uma máquina real. Esta simulação tão abrangente implica em uma maior exigência de capacidade de memória e de processamento para sua execução. Mais recentemente, com o advento dos ambientes de computação em nuvem um novo tipo de virtualização tem sido adotada, denominada virtualização de processos (Fink, 2018). Na virtualização de processos um conjunto mínimo suficiente de dependências é mantido na ferramenta de virtualização para possibilitar a execução de um programa em diferentes ambientes computacionais, com menores requisitos de memória e processamento. O processo virtualizado é executado, de forma semelhante aos programas executados em máquinas virtuais, ou seja, como se estivesse em computador independente, com sistema operacional, sistema de arquivos e rede próprios. Esta característica possibilita que por exemplo, várias instâncias de um mesmo programa possam ser executados em um mesmo computador, de forma virtualizada, sem que haja interferência entre eles.

Está flexibilidade na execução de programas em diferentes ambientais computacionais com baixa exigências de recursos e com isolamento impõe, porém restrições em quais tipos de programas podem ser virtualizados.

A ferramenta *Docker* (Docker, 2018), permite a construção de processos virtualizados e foi utilizada no estudo de caso implementado neste artigo. Através dela, um programa implementado por exemplo, para executar no ambiente DOS , poderá ser executado em um ambiente Linux, sem que haja necessidade de sua modificação. Apesar desta flexibilidade existem restrições impostas para uso da virtualização de processos com o *Docker*:

- Aplicações que apresentam interface gráfica para interação com o usuário não são adequadas para este tipo de virtualização. Aplicações do tipo *batch* que são executadas somente com invocação via linha de comando são mais apropriadas para este uso;
- Mecanismos para exportação de dados internos produzidos pelo programa devem ser previamente configurados, uma vez que todos estes dados serão perdidos ao final da

execução do processo virtualizado. A ferramenta *Docker* possui funcionalidades que permitem esta configuração;

– Programas com funcionalidades mais especificas, autônomos e pouco complexos são mais adequados a este tipo de virtualização uma vez que requerem menos ou nenhuma comunicação com outros programas externos e geram menos subprocessos durante sua execução.

Um processo virtualizado é denominado *docker container* no ambiente *Docker* e corresponde a execução de um programa base armazenado em uma imagem *docker*. Uma imagem *docker* é uma estrutura de dados que armazena todas as informações necessárias para executar o programa na forma de um processo virtualizado, desde arquivos do sistema operacional até bibliotecas e outros programas auxiliares, abrangendo todas as suas dependências para execução. A montagem de uma imagem *docker* é realizada no processo de construção ou *build.* Neste processo um arquivo denominado *Dockerfile*, apresenta um conjunto de instruções de como e quais arquivos devem ser incorporados a uma nova imagem.

As características de isolamento de execução propiciadas pela execução de programas utilizando o recurso de virtualização de processo juntamente com a baixa exigência de recursos computacionais e independência de configuração computacional desta abordagem serve de base para construção de soluções para processamento de grande volume de dados.

A ideia é paralelizar o processamento dos dados, através da criação de múltiplas réplicas de um mesmo programa em sua forma virtualizada. As réplicas podem ser criadas até o limite da capacidade do ambiente computacional disponível. A existência de diferentes configurações computacionais para execução é superada pelas próprias características da virtualização de processos. A ferramenta *Docker Swarm*, fornecida em conjunto com o *Docker* permite a gestão da execução destas réplicas.

3.2 Interpolação por Krigagem

Neste estudo de caso foi utilizado o método de interpolação por krigagem para avaliar o aumento da capacidade de processamento proposta no trabalho. A interpolação por krigagem (Krige, 1951) é um método recorrente no tratamento de dados geoespaciais e exige muito recursos de memória e processamento para a sua execução.

Esse interpolador é baseado em regressão espacial, com o objetivo de estimar valores de um atributo em uma localização não amostrada a partir de valores associados às amostras de sua vizinhança espacial. Essas estimativas são ponderadas de acordo com valores de covariância espacial, obtidos a partir da construção de uma função conhecida como semivariograma (Bohling, 2005). Como é praticamente impossível a determinação exata de que tipo de equação matemática descreve a variabilidade espacial dos dados, o semivariograma se torna uma solução interessante por permitir a interpretação física do fenômeno em questão (Vieira , 2000). Após a sua construção, é necessário que o mesmo seja ajustado a um modelo teórico aproximado (e.g., linear,esférico, exponencial, gaussiano), para que possa ser utilizado como fator de ponderação pelo algoritmo interpolador da krigagem.

Neste estudo de caso o algoritmo de interpolação espacial utilizado foi a krigagem, com ajuste automático de semivariograma, executado pela função autoKrige, disponível no pacote *Automap* para R (versão 1.0.14)[1] (Hiemstra et al, 2008). O *script* R tem como parâmetros o dado original na forma de um *grid* de dados amostrados, a indicação do atributo a ser utilizado na interpolação e, um *grid* de pontos de saída para os quais se deseja calcular o valor interpolado.

1 https://cran.r-project.org/web/packages/automap/

3.3 Dados

Dados de imagem *Sentinel*-2/MSI do dia 09/05/2018, obtidas por meio do provedor *Copernicus Open Acess Hub*[2] com correção nível 1-C foram utilizados no interpolador por krigagem. A correção atmosférica foi realizada utilizando-se o aplicativo oficial *Sen2Cor* (Louis et al, 2016). A imagem *raster* resultante, com resolução de 10 metros, foi convertida para vetorial utilizando-se algoritmos da biblioteca *GDAL* disponíveis no SIG *QGIS* (QGIS, 2018). O dado vetorial formando um *grid* de pontos foi exportado para um servidor de dados geoespaciais *GeoServer* (OSGF, 2018), que segue as especificações da OGC (*Open Geospacial Consortium*). O acesso aos dados via servidor *GeoServer* facilita a paralelização do processamento dos dados uma vez que este servidor possibilita a partição dos dados de diferentes formas com base nos seus filtros de consulta.

A imagem processada corresponde a área do talhão experimental em produção de grãos da Fazenda Farroupilha no município de Pedra Preta, localizado na porção leste do planalto da Bacia do Alto Paraguai, no estado de Mato Grosso. A área possui aproximadamente 50 ha, com coordenadas centrais aproximadas -54,071 (longitude) e -16,878 (latitude), medidas em graus decimais considerando o datum WGS84.

3.4 Infraestrutura de processamento

O *script* R que realiza a krigagem foi utilizado em sua forma original sem alteração. Para acessá-lo como um serviço no *Docker Swarm* foi implementado um conversor que permite a execução do *script* via *Web API*, ou seja, os parâmetros e dados necessários para sua execução são enviados através de um endereço na Web. Ao se criar várias instância destes serviços de krigagem, cada uma se comportará como um serviço na Web instalado em diferentes endereços de internet, correspondendo a diferentes ambientes de execução de um processo virtualizado. No nosso estudo de caso os serviços foram criados tendo como base uma imagem *docker* contendo todos os arquivos necessários para execução do *script* R de krigagem. A montagem desta imagem docker é ilustrada na **Figura 1**.

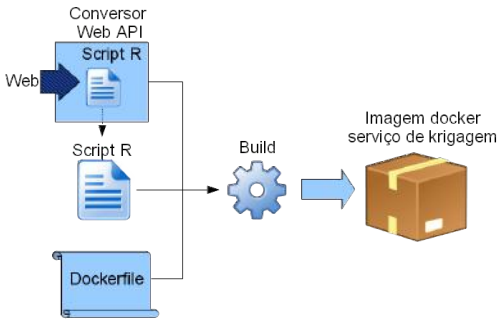


Figura 1 – Construção de imagem docker do serviço de krigagem

A imagem do serviço de krigagem foi então utilizada como base para criação de várias instâncias do s*cript* R, que permitem a paralelização do processamento de interpolação da imagem *Sentinel*. Estas instâncias foram executadas em um infraestrutura virtual criada com o apoio da ferramenta *OpenStack*. A gestão da distribuição e acesso aos diferentes serviços de krigagem foi realizado usando a ferramenta *Docker Swarm*.

O *OpenStack* é uma plataforma, de software livre e de código aberto, para prover ambientes de computação em nuvem, mais especificamente Infraestrutura como Serviço (IaaS), ou seja, para prover serviços de processamento. A partir do *OpenStack* podem ser criadas máquinas

[2] *Copernicus Open Access Hub* (European Spatial Agency): http://scihub.copernicus.eu

virtuais customizadas para os mais diversos fins, de forma amigável e transparente em relação a infraestrutura física do *data center* (**Figura 2**).

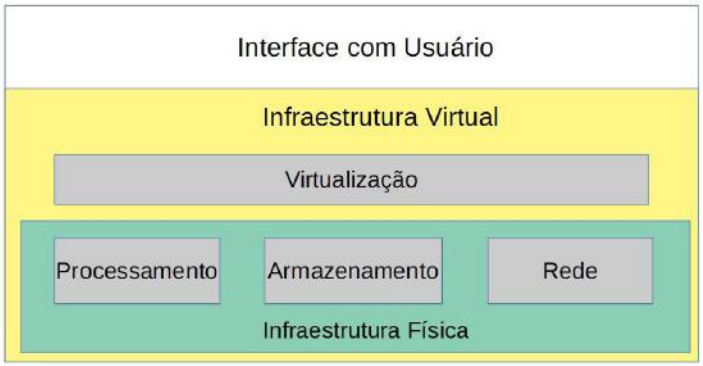


Figura 2 - Representação Gráfica da Plataforma *OpenStack*

Neste trabalho, a partir do *OpenStack*, foram criadas dez máquinas virtuais (**Figura 3**) para criação de uma infraestrutura de processamento distribuído utilizando *Docker Swarm*. A máquina ao centro é a máquina *Manager*, que gerencia a grade de processamento, cuja função é distribuir as tarefas de processamento para si, assim como, para os nove *Workers*, ou seja, acumula as tarefas de gerenciamento e processamento. Todas as máquinas possuíam a mesma configuração, sendo 4 vCPUs de 2,6 GHz, com 4 MB de cache, 16 GB de memória RAM e 50 GB de disco.

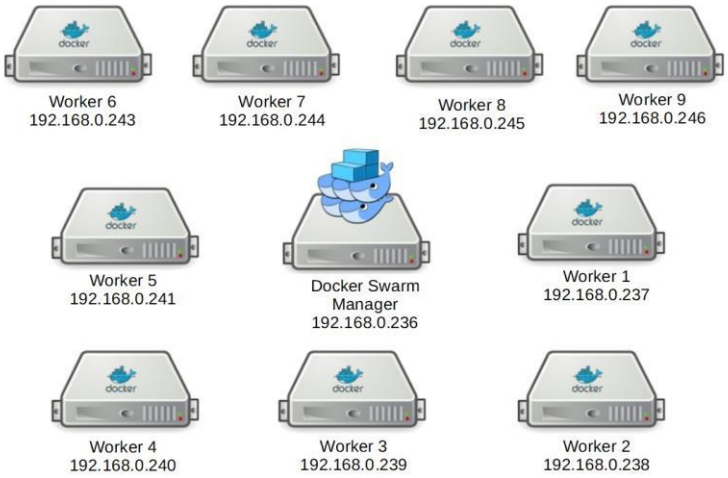


Figura 3 - Infraestrutura Docker Swarm

3.5 Método

Por meio da ferramenta de gestão *Docker Swarm* foram distribuídas 40 instâncias de serviços de krigagem nas diferentes máquinas da infraestrutura computacional sendo que cada máquina acomodou 4 serviços. Para verificar o efeito da paralelização, cada instância de processamento recebeu como parâmetros toda a imagem *Sentinel* original e um *grid* de saída com resolução de 1m particionado em diferentes configurações. Desta forma todos os dados da imagem *Sentinel* estavam disponíveis na fase de modelagem da krigagem e a paralelização efetiva ocorria na fase do cálculo dos valores interpolados para o *grid* parcial de saída. Após o processamento dos dados, os *grids* de saída com valores interpolados eram unidos novamente para formar uma única camada de dados interpolado em um *grid* de pontos de 1x1m. Para cada configuração de

paralelização foi registrado o tempo de download das partições de dados e o de processamento. Foram também calculadas estatísticas de erro para cada configuração de paralelização.

4. Resultados

O método de interpolação por krigagem é um método estatístico que apresenta variações em seus resultados, inerentes a sua concepção como um método de estimativa. Desta forma os resultados interpolados para os pontos de saída apresentam diferenças que são apresentadas na **Tabela 1** e que foram calculadas com base na amostragem de 1% dos resultados obtidos para cada configuração de partição dos dados. Nesta tabela o valor interpolado sem considerar a partição do *grid* (partição 1 x 1) de saída foi considerado como referência.

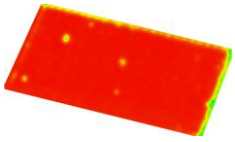
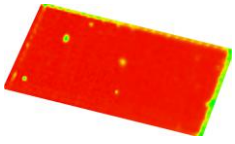
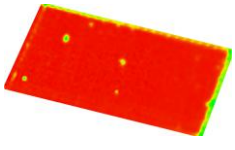
Tabela 1 - Erro

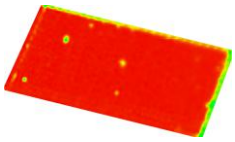
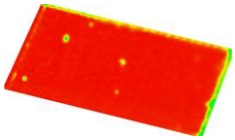
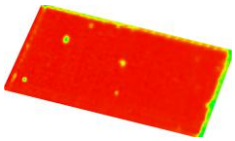
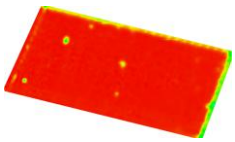
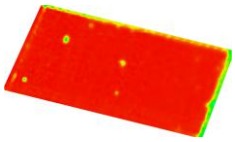
Partição	Média do erro	Erro mínimo	Erro máximo	Variância do erro	Correlação
2x2	0.002128992	0.000000044496	0.001275426	0.000030048	0.98626
3x3	0.002063035	0.000000068814	0.091652997	0.000023787	0.99183
4x4	0.002495103	0.000000234123	0.114136178	0.000042226	0.98549
5x5	0.002151904	0.000000034865	0.108552504	0.000037929	0.98678
6x6	0.002625005	0.000000309868	0.111219722	0.000051142	0.98947
5x8	0.002095078	0.000000034965	0.110569205	0.000029836	0.99013
8x5	0.002220856	0.000000033287	0.011927845	0.000056720	0.98718

O variância e o grau de correlação entre o valores de referência e os valores interpolados indicam um impacto mínimo do processo de paralelização do processamento dos dados no resultado obtido.

A **Tabela 2** abaixo apresenta os dados coletados correspondentes aos tempos de *download* e processamento dos dados para cada configuração das partições. A coluna partições indica em quantas colunas e linhas (X e Y) o *grid* de saída foi particionado. Cada partição do *grid* de saída era obtida do servidor *GeoServer* e enviada para cada instância de processamento juntamente com a imagem *Sentinel*.

Tabela 2 - Paralelização do processo de krigagem

Grid interpolada	Partições (X x Y)	Número Processos	Tempo download (segundos)	Tempo total processamento (minutos)	Tempo max. Krigagem (minutos)
	1x1	1	12	910	889
	2x2	4	13	56	54
	3x3	9	15	23	22

	4x4	16	16	6	5
	5x5	25	19	6	5
	6x6	36	22	5	3
	5x8	40	25	5	3
	8x5	40	24	5	3

5. Conclusão

Pode ser verificada diminuição importante do tempo da geração do dado interpolado final quando comparado com o processamento sem paralelização, indicando a validade da abordagem adotada neste caso de uso. O tempo gasto no particionamento dos dados e envio para o serviço não foi significante, porém neste caso foi utilizado um filtragem simples baseada na seleção por *bounding boxes* da região cujos dados seriam interpolados. Partições mais complexas dos dados a serem tratados podem implicar em um maior tempo de consulta.

O efeito da paralelização diminui quanto mais instâncias paralelas são utilizadas, talvez porque o tempo gasto na fase da geração de modelo da krigagem, que é o mesmo para todas as instâncias, se torne mais importante na geração da interpolação. A grande redução inicial no tempo de processamento, principalmente da partição 1x1 para 2x2 pode indicar que na configuração inicial ocorreu uma sobrecarga dos recursos computacionais para processamento e que neste caso a configuração de 1 CPU a 2.6 GHz e 4 GB de RAM não é adequada para este processamento.

O *script* R foi utilizado em sua forma original sem modificações. Uma possibilidade para melhoria do desempenho do procedimento pode ser a análise do algoritmo implementado e a sua divisão em diferentes partes de forma que o tempo gasto na fase de modelagem da krigagem não precise ser realizado de forma redundante por todas as instâncias de processamento. Outras configurações da infraestrutura computacional podem ser testadas, com a criação de mais máquinas com menos memória e CPU. Na configuração atual do *OpenStack* seria possível criar 40 máquinas com 1VCPU e 4G de memória RAM, onde cada instância poderia ser executada em uma máquina independente com redução de interferência entre os processos virtualizados no nível de infraestrutura computacional.

Foi verificado que em alguns resultados ocorreu um aumento do número de pontos a serem interpolados. Uma hipótese para este aumento esta na precisão na realização do filtro por *bounding box* do *GeoServer*, que em alguns momentos deve coincidir com a linha de pontos do *grid* e desta forma esta linha acaba sendo incluída em duas partições vizinhanças.

Com relação a qualidade dos resultados devem ser realizados mais testes com a utilização de dados com maior variabilidade espacial para verificar a influência da redução de área dos dados interpolados no algoritmo de krigagem paralelizado.

A diminuição do tempo de processamento com a adoção da estratégia proposta no trabalho possibilita a sua utilização no tratamento de dados obtidos de forma sistemática e em grande quantidade, tais como aqueles resultantes do uso de sensores de alta resolução espacial e/ou abrangendo grandes áreas.

6. Agradecimentos

Agradecemos à equipe do Instituto Mato-grossense do Algodão (IMAmt), responsável pela gestão da área da Fazenda Farroupilha onde são realizados experimentos de AP, pelas informações fornecidas a respeito da localização espacial e histórico de produção.

Agradecemos também à Pesquisadora Dr. Célia Regina Greco pelas orientações a respeito dos procedimentos geoestatísticos.

7. Referências

Bohling, G. **Introduction to Geostatistics and Variogram Analysis.** Kansas Geological Survey, Lawrence, 2005.

Docker Inc. **Get Started, Part 1: Orientation and setup** Disponivel em: <https://docs.docker.com/get-started/>. Acesso em: 10 de junho 2018.

Fink, J. **Docker: a Software as a Service, Operating System-Level Virtualization Framework** Disponivel em: <http://journal.code4lib.org/articles/9669?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+c4lj+>. Acesso em: 5 de julho de 2018.

Open Source Geospatial Foundation, **GeoServer** Disponivel em:<http://geoserver.org/> . Acesso em: 28 junho de 2018.

Goodchild, Michael F. **The quality of big (geo) data** Dialogues in Human Geography 3.3 (2013): 280-284.

Hey, T.; Tansley, S.; Tolle, K. (Ed.). **The fourth paradigm: data-intensive scientific discovery**. Redmond: Microsoft Research, 2009. 252 p.

Hiemstra, P.H.; Pebesma, E.J.; Twenhofel, C.J.W.; G.B.M. Heuvelink. **Real-time automatic interpolation of ambient gamma dose rates.** Dutch Radioactivity Monitoring Network. Computers & Geosciences, accepted for publication, 2008

Krige, D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand 1951 **Journal of the Southern African Institute of Mining and Metallurgy**, Volume 52, Issue 6, Dec 1951, p. 119 - 139

Laureano, M. A. P.. **Máquinas Virtuais e Emuladores – Conceitos, Técnicas e Aplicações.** São Paulo: Novatec, 2006. 184 p. Disponível em: <http://www.mlaureano.org/aulas_material/so/livro_vm_laureano.pdf>. Acesso em: 4 de julho de 2018.

Louis, J.; Debaecker, V.; Pflug, B.; Main-Knorn, M.; Bieniarz, J.; Müller-Wilm, U.; Cadau, E.; Gascon, F. . **SENTINEL-2 SEN2COR: L2A Processor for Users**, 2016

Marcuzzo, F.F.N; Andrade, L.R.; Melo, D.C.R. Métodos de interpolação matemática no mapeamento de chuvas no estado do Mato Grosso. **Revista Brasileira de Geografia Física**, v.4, p.793-804, 2011.

Muñoz, V. ; Valeriano, M. D. M.; Bispo, P. Surveying the topographic height from SRTM DATA for canopy mapping in the brazilian Pantanal, Geografia, Vol. 38, Nº Extra 1, 2013, 139-156

QGIS Development Team . **QGIS Geographic Information System**. Open Source Geospatial Foundation Project. http://**qgis**.osgeo.org, 2018

Tieppo, R. C.; Nunes, C. C. P.; Dallacort, R.; Fietz, C. R.; Comunello, E.; Cremon, C. Análise de interpoladores na geração de mapas de precipitação para o Estado de Mato Grosso. In: **SIMPÓSIO DE GEOTECNOLOGIAS DO PANTANAL**, 3., 2010, Cáceres. Anais... Campinas: Embrapa Informática Agropecuária: São José dos Campos INPE, 2010. 3. Geopantanal.

Vieira, S.R. Geoestatística em estudos de variabilidade espacial do solo. In: Novais, R.F.; Alvarez V., V.H. & Schaefer, G.R., eds. **Tópicos em ciência do solo**. Viçosa, Sociedade Brasileira de Ciência do Solo, 2000. v.1, p.1-54.