UNICAMP

NÙMERO: 430/2011

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GEOCIÊNCIAS
ÁREA DE GEOLOGIA E RECURSOS NATURAIS

FÁBIO IWASHITA

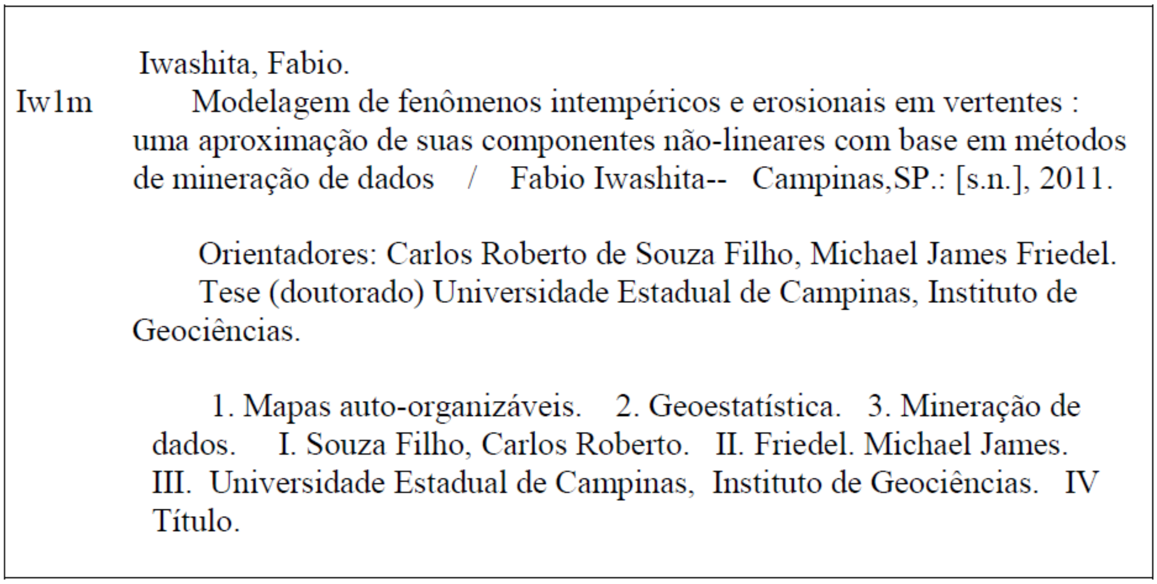
Modelagem de fenômenos intempéricos e erosionais em vertentes: uma aproximação de suas componentes não-lineares com base em métodos de mineração de dados

Tese apresentada ao Instituto de Geociências como parte dos requisitos para obtenção do título de Doutor em Geociências.

Orientador: Prof. Dr. Carlos Roberto de Souza Filho
Co-orientador: Dr. Michael James Friedel

CAMPINAS - SÃO PAULO
Maio – 2011

Catalogação na Publicação elaborada pela Biblioteca
do Instituto de Geociências/UNICAMP



Iwashita, Fabio.
Iw1m Modelagem de fenômenos intempéricos e erosionais em vertentes : uma aproximação de suas componentes não-lineares com base em métodos de mineração de dados / Fabio Iwashita-- Campinas,SP.: [s.n.], 2011.

Orientadores: Carlos Roberto de Souza Filho, Michael James Friedel.
Tese (doutorado) Universidade Estadual de Campinas, Instituto de Geociências.

1. Mapas auto-organizáveis. 2. Geoestatística. 3. Mineração de dados. I. Souza Filho, Carlos Roberto. II. Friedel. Michael James. III. Universidade Estadual de Campinas, Instituto de Geociências. IV Título.

Título em ingles: Modeling of soil weathering on hillslopes : coping with nonlinearity and coupled processes using a data-driven approach.
Keywords: - Self-Organizing Maps;
- Geostatistics;
- Data Mining.
Área de concentração: Geologia e Recursos Naturais
Titulação: Doutor em Ciências.
Banca examinadora: - Carlos Roberto de Souza Filho;
- Emilson Pereira Leite;
- Francisco José Fonseca Ferreira;
- Oderson Antonio de Souza Filho;
- Ricardo Perobelli Borba
Data da defesa: 06/05/2011
Programa de Pós-graduação em Geociências.

UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS
PÓS-GRADUAÇÃO EM GEOCIÊNCIAS NA
ÀREA DE GEOLOGIA E RECURSOS NATURAIS

AUTOR: Fábio Iwashita

"Modeling of Soil Weathering on Hillslopes: Coping WithNonlinearity and Coupled Process Using a Data-Driven Approach".

ORIENTADOR: Prof. Dr. Carlos Roberto de Souza Filho
Co-orientador: Prof. Dr. Michael James Friedel

Aprovada em: 06 / 05 / 2011

EXAMINADORES:

Prof. Dr. Carlos Roberto de Souza Filho ____________________ - Presidente

Prof. Dr. Ricardo Perobelli Borba ____________________

Prof. Dr. Emilson Pereira Leite ____________________

Prof. Dr. Francisco José Fonseca Ferreira ____________________

Dr. Oderson Antônio de Souza Filho ____________________

Campinas, 06 de maio de 2011

À minha família dedico…

Essentially, all models are wrong, but some are useful.

George E. P. Box

There are 10 types of people in the world;
Those who understand binary and those who don't.

AGRADECIMENTOS

Expresso minha gratidão às pessoas e instituições cuja colaboração foi indispensável para o desenvolvimento desta tese de doutorado.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico/CNPq pelo financiamento do doutorado e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior/CAPES pelo apoio financeiro ao longo do período que estive como pesquisador visitante no Serviço Geológico Americano.

Ao prof. Dr. Carlos Roberto de Souza Filho pelo apoio e dedicada orientação, sempre respondendo aos emails e me ajudando, não importando se era madrugada, domingo ou feriado.

Ao Dr. Michel James Friedel pela paciência, incentivo e fundamental orientação durante meu período junto ao USGS.

Ao Dr. Victor Labson pelo gigantesco apoio, ao gentilmente me receber e fornecer infra-estrutura para o desenvolvimento da minha pesquisa junto ao Crustal Geophysics and Geochemistry Science Center, U.S. Geological Survey/USGS, Denver, CO, e aos pesquisadores Dr. George Breit e Dr. David Smith pelos valiosos comentários e sugestões sobre geoquímica de solos durante minha estadia no USGS.

Ao Dr. Francisco José Fonseca Ferreira que gentilmente pré-processou dados geofísicos cedidos pela Petrobrás, e forneceu dados de poços para o desenvolvimento dos trabalhos.

Ao Dr. Mario Miyazawa, Dr. Otavio Augusto Boni Licht e geólogos Mario Kondo e João Horácio Pereira pelo fornecimento de geoquímica de solos e hidroquímica de poços de água subterrâneas.

À geógrafa e amiga Glaucielen Farias Ribeiro que gentilmente cedeu dados de textura do solo e condutividade hidráulica.

Às secretárias Valdirene Pinotti e Mª Gorete S. S. Bernardelli cujo valioso apoio técnico foi indispensável para elaboração deste trabalho.

À Dra. Holly Huyck e ao Dr. Daniel O'Connell por me receberem em Evergreen e por depositarem em mim enorme confiança durante minha estadia em sua propriedade.

À família Friedel, Nancy, Ben e Becca que me receberam de braços abertos durante os feriados que passei longe de minha família.

PUBLICAÇÕES RESULTANTES DO DOUTORADO

Artigos Publicados

Iwashita, F., Friedel, M. J., Souza Filho, C. R., Fraser, S. J. Hillslope chemical weathering across Paraná, Brazil: A data mining-GIS hybrid approach, **Geomorphology**.
http://dx.doi.org/10.1016/j.geomorph.2011.05.006

Artigos Submetidos

Iwashita, F., Friedel, M. J., Ribeiro, G. F., Fraser, S. J. Intelligent estimation of spatially distributed soil physical properties**, Geoderma**, 25p.

Iwashita, F. Souza Filho, C. R. Evaluating SRTM-90m interpolation using SRTM-30 from U.S. territory, **Computer & Geosciences**, 12p.

Friedel, M. J., **Iwashita**, F. Nonlinear modeling of autocorrelated random variables for application to inverse problems**, Environmental Modeling & Software,** 32p.

Artigos em revisão

Iwashita, F., Souza Filho, C.R., Ferreira, F. J. F., Fraser, S. J. Estimating physical-chemical properties in fractured aquifers using self-organizing maps imputation approach: Study of hydraulic connectivity between Serra Geral and Guarani aquifer in Paraná state, Brazil, **Journal of Hydrology,** 38p.

Trabalhos Publicados em Anais de Congressos

Iwashita, F., F., Friedel, M. J., Souza Filho, C.R., Fraser, S. J. Using self-organizing maps to analyze high-dimensional geochemistry data across Paraná, Brazil. In: **15th Simpósio Brasileiro de Sensoriamento Remoto.** Curitiba, Brazil

Iwashita, F., Friedel, M. J., Ribeiro, G. F., Fraser, S. J., Souza Filho, C. R. 2010. Using self-organizing maps to predict soil texture and hydraulic conductivity in Poços de Caldas, Minas Gerais, Brazil. In*:* **45th Congresso Brasileiro de Geologia**. Belém, Brazil

Iwashita, F., Friedel, M. J., Souza Filho, C.R. 2010. Surveying soil chemical weathering in Paraná state/Brazil: A data mining-GIS hybrid approach. In: **3rd USGS Modeling Conference**. Denver-CO.

Iwashita, F., Souza Filho, C. R. 2009. Identificação de anomalias geoquímicas de cobre e zinco presentes no solo através de algoritmos genéticos e envelopes bioclimáticos no estado do Paraná. In: **44th Congresso Brasileiro de Geologia**. Curitiba, Brazil.

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GEOCIÊNCIAS
ÁREA DE GEOLOGIA E RECURSOS NATURAIS

Modelagem de fenômenos intempéricos e erosionais em vertentes: Uma aproximação de suas componentes não-lineares com base em métodos de mineração de dados

RESUMO

Tese de Doutorado

Fabio Iwashita

Esta tese de doutorado tem como objetivo aprofundar o conhecimento sobre as relações das propriedades físico-quimicas do solo com a morfometria do relevo, buscando quantificar essas relações para a construção de modelos conceituais e preditivos. Mapas auto-organizáveis e modelos de sistemas de informação geográfica foram utilizados para investigar as relações não lineares associadas ao intemperismo químico e físico, fatores associados a fenômenos hidrológicos e à evolução dos solos. Três estudos de caso são apresentados: o intemperismo químico de solo no estado do Paraná (22 variáveis e 304 amostras), o transporte físico de sedimentos em Poços de Caldas (9 variáveis e 29 amostras), e hidroquímica de aqüíferos na Formação Serra Geral no Estado do Paraná (27 variáveis e 976 amostras). O método combinando simulação estocástica e mineração de dados permitiu explorar as relações entre relevo, granulometria e geoquímica dos solos. Regiões mais elevadas e com morfometria convexa apresentaram alta denudação de elementos móveis (e.g., Ca) e baixa de elementos pouco móveis (e.g., Al). O mesmo padrão foi observado para granulometria de solos, ou seja, alta proporção de areia em áreas altas e convexas da bacia e altos teores de argila, com baixa condutividade hidráulica, em regiões convexas próximas aos canais de drenagem. O comportamento espacial da hidroquímica das águas do aqüífero Serra Geral apontou áreas de potencial conectividade entre aqüíferos, áreas de recarga recente e de alto tempo de residência. Foram construídos modelos preditivos não tendenciosos das propriedades do solo em subsuperfície partindo da premissa de que o intemperismo e a morfometria se relacionam através de um processo duplamente dependente, onde a denudação física e química atua no delineamento do relevo e a morfometria do terreno é um fator que caracteriza as condições físico-químicas do solo.

Palavras-chave: intemperismo de solos, mapas auto-organizáveis, simulação Monte Carlo, imputação, Paraná, Poços de Caldas, aqüífero Serra Geral.

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GEOCIÊNCIAS
ÁREA DE GEOLOGIA E RECURSOS NATURAIS

Modeling of soil weathering on hillslopes: Coping with nonlinearity and coupled processes using a data-driven approach

ABSTRACT

Tese de Doutorado

Fabio Iwashita

This Doctoral thesis aims to explore the relationship between soil physical-chemical properties and relief morphometry, and quantifying these relationships to build conceptual and predictive models. Self-organizing maps and Geographic Information Systems modeling are here used to investigate nonlinear correlations associated with chemical and physical denudation; which are factors connected with hydrological phenomena and soil evolution. Three study cases are presented: soil chemical weathering within the limits of the Parana State, southern Brazil (22 variables and 304 samples), physical transport of sediments in the alkaline intrusive complex of Poços de Caldas, southeastern Brazil (9 variables and 29 samples), and hydrochemistry of Serra Geral aquifers also in the Parana State (27 variables and 976 samples). The method combining stochastic simulation and data mining allows exploring the relationships between topography, soil texture and soil geochemistry. In the Parana State, higher regions and areas with convex morphometry shows, respectively, higher and lower denudation rates of mobile (e.g., Ca) and less mobile (e.g., Al) elements. The same pattern is observed for soil particle size. In this case, high proportion of sand is found in highlands and convex areas inside the basin, and high clay content, with low hydraulic conductivity, occurs in convex regions, near drainage channels. The spatial behavior of the Serra Geral aquifer's hydrochemistry pointed out to areas with potential connectivity with the Guarani aquifer system, recent recharge areas, and long-standing waters. Predictive, unbiased models are built for soil properties on the premise that weathering and morphology are related through a two-way dependent process, where the physical and chemical denudation delineates the elevations of the land surface, and terrain morphometry is a factor that characterizes the physical-chemical conditions of the soil.

Keywords: soil weathering, self-organizing maps, Monte Carlo simulation, imputation, Parana, Poços de Caldas, Serra Geral aquifer

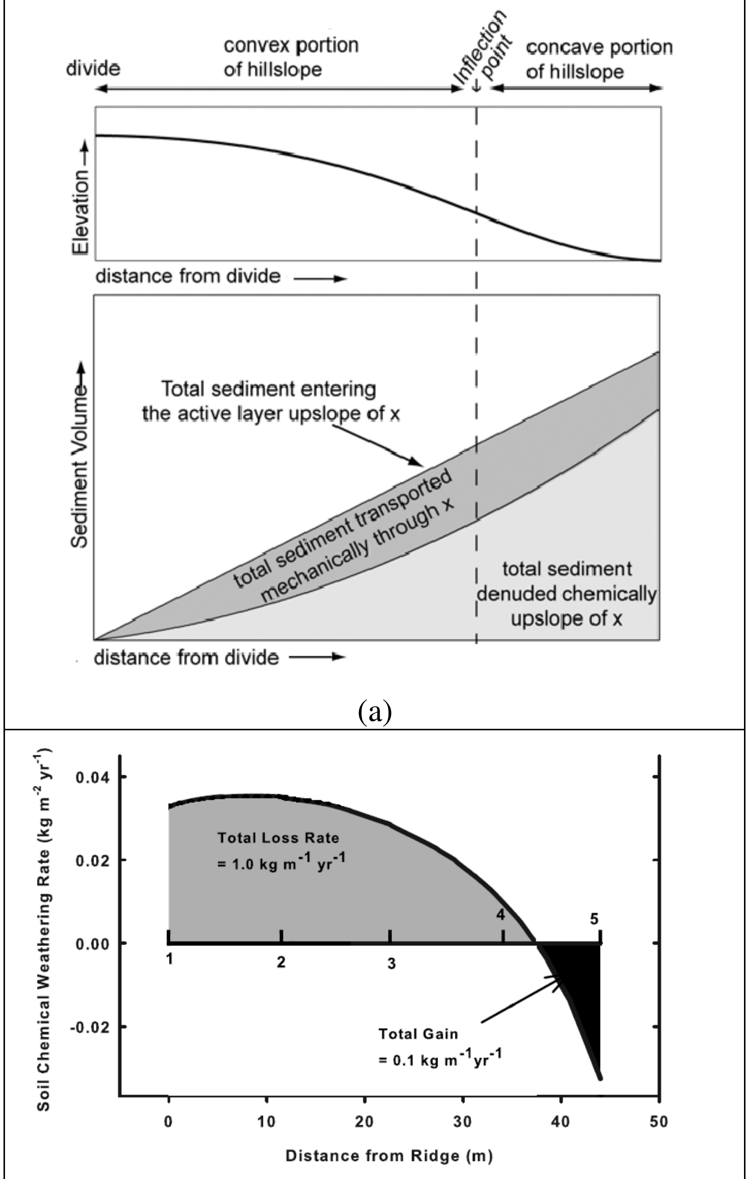
SUMÁRIO

1. INTRODUÇÃO

A morfometria do terreno reflete os processos de intemperismos físico e químico, responsáveis por sua evolução e delineamento. Entender o processo de intemperismo requer conhecimento sobre os fenômenos que influenciam a formação da paisagem. Os primeiros modelos numéricos e métodos empíricos permitiam a quantificação da perda de massa do solo apenas a partir de um ponto de vista físico. Tais modelos consideravam as encostas uniformes em toda sua extensão (i.e., retilíneas), não refletindo a heterogeneidade das taxas de transporte e deposição ao longo das encostas (Heimsath et al., 1997). Modelos recentes de balanço de massa assumem que a taxa de transporte de sedimentos não é linear, considerando que as encostas possuem morfometria convexa próxima ao topo, retilínea na porção intermediária e côncava em sua base (Roering et al., 1999).

Recentemente, o intemperismo químico passou a ser incorporado no cálculo de balanço de massa em encostas, já considerando sua natureza não linear. Mudd e Furbish (2004) formularam uma equação generalizada de balanço de massa em encostas incluindo transporte físico de sedimentos, deposição e denudação química. Estas equações foram desenvolvidas em um modelo unidimensional sob condições de estacionariedade, onde a elevação das encostas não muda no tempo considerado (i.e., sem soerguimento). Os autores constataram que a quantidade total de massa transportada por intemperismo químico aumenta de forma não linear em relação à distância do divisor, enquanto a partir do ponto de inflexão, onde a encosta é retilínea, o transporte mecânico diminui (Figura 1a). Yoo et al. (2007) combinaram um modelo numérico com medidas de geoquímica de solos coletadas ao longo de um transecto em uma encosta no sudoeste da Austrália. As taxas de intemperismo químico do solo apontaram perda de massa próxima ao divisor e uma tendência de acúmulo conforme a distância em relação ao topo aumentava (Figura 1b).

Estes estudos mostram que as propriedades morfométricas das encostas podem ser incorporadas na modelagem do intemperismo físico e químico dos solos e aplicadas para áreas extensas através de imagens de sensores remotos, modelos digitais de elevação e Sistemas de Informação Geográfica (SIG) para calcular medidas como declividade, orientação de vertentes, curvatura vertical, curvatura horizontal e fluxo hidrológico acumulado. O intemperismo físico e químico é um fenômeno complexo e um fator importante no delineamento das encostas, uma vez que a mobilidade de elementos está intimamente ligada às condições físico-quimicas dos solos (pH, umidade, temperatura, porosidade, etc.). Por exemplo, áreas consideradas de aspecto côncavo estão associadas a um fluxo hidrológico convergente e, portanto, apresentam umidade média do solo mais elevada que áreas convexas, que geralmente caracterizam topos de morros.



(a)

(b)

Figura 1. Modelo conceitual da taxa de transporte de sedimentos baseada no perfil da encosta. a) o total de sedimentos denudados quimicamente pode ser descrito por uma função não linear em relação à distância do divisor. A massa denudada quimicamente nas áreas côncavas é maior que a transportada mecanicamente (Mudd and Furbish, 2004). b) taxa total de intemperismo químico mostrando a variação de massa em função da distância do divisor (Yoo et al., 2007).

A abordagem empírica de análise de dados geoquímicos de solo é usualmente feita através de métodos de estatística multivariada, tais como regressão linear múltipla, análise de componentes principais, análise de agrupamentos e análise fatorial. Estes métodos são robustos e confiáveis, mas dependem da admissão de determinadas premissas, tais como, distribuição normal dos resíduos, homocedasticidade, e não colinearidade entre as variáveis explicativas (Netter et al, 1996). Um complicante adicional é o fato que, segundo Reimann e Filzmore (1999), dados geoquímicos, em escala regional, não possuem distribuição normal, nem lognormal. Uma alternativa para analisar dados multivariados são métodos de mineração de dados.

De acordo com Kohonen (2001), os mapas auto-organizáveis (SOM) ou vetores quantizados são adequados para lidar com dados ruidosos, não estacionários e com distribuições não Gaussianas, pois evidenciam relações não lineares através de transformações topológicas da informação original. A ausência de premissas é uma das principais vantagens de uma abordagem baseada em mineração de dados, pois grande parte dos métodos multivariados assume que as relações entre variáveis independentes e dependentes são lineares. Outra diferença é que modelos estatísticos preditivos, como a regressão linear múltipla, penalizam a inclusão de grande número de variáveis explicativas, buscando um balanço entre o número de variáveis e a quantidade de informação explicada pelo modelo, obtendo um modelo relativamente simples cujo poder preditivo é satisfatório (Netter el al., 1996). Para lidar com um grande número de variáveis potencialmente explicativas e relações não lineares entre as mesmas, diversos trabalhos têm empregado os mapas auto-organizáveis de Kohonen, para, por exemplo, explorar as relações entre a geoquímica de rocha e imagens hiperespectrais (Penn, 2005), para classificar aspectos geomorfométricos a partir de modelos digitais de elevação (Ehsani e Quiel, 2008), caracterizar a vulnerabilidade de encostas a escorregamentos (Hentati et al., 2010), e para identificar os principais processos que controlam a distribuição dos íons Fe^{3+} e Fe^{2+} nos solos e nos sedimentos (Löhr et al., 2010).

As escalas geográficas dos modelos de intemperismo em encostas são geralmente em nível de paisagem, onde a erosão física é mínima e o modelo hidrológico mais simples. Escalas em níveis mais generalizados aumentam a quantidade e a complexidade de fenômenos a serem considerados (ASCE, 2000). Estas restrições fazem da mineração de dados, especificamente, o método SOM, uma alternativa para lidar com dados de distribuição não gaussiana, ruidosos e de ordens elevadas (polinômios quadráticos ou de maior ordem) de correlação.

O objetivo principal desta tese de doutorado é contribuir para o estabelecimento de um modelo das relações entre o intemperismo físico-químico dos solos e a morfometria do relevo em regiões próximas à superfície. A metodologia de modelagem conceitual é estendida a regiões em profundidade, incluindo a caracterização hidroquímica de um aqüífero fraturado não confinado. A hipótese do trabalho é que o intemperismo físico-químico e a forma do terreno são processos acoplados e podem ser modelados de maneira conceitual e preditiva sob condições de estacionariedade.

Para quantificar as relações estatísticas não lineares entre dados de campo e os modelos de morfometria do terreno, é proposta uma metodologia combinando redes neurais artificiais, simulação estocástica e técnicas de análise espacial, onde nenhuma premissa estatística é exigida. O método foi aplicado em três áreas de estudo: na região de Poços de Caldas (Minas Gerais) e duas no Estado do Paraná, incluindo terrenos da Formação Serra Geral. O banco de dados de Poços de Caldas compreende 29 amostras com informações de textura de solo e condutividade hidráulica, descrevendo propriedades físicas do solo em escala local. O conjunto de dados do Estado do Paraná é composto de 22 variáveis e 304 amostras que descrevem a geoquímica de solos em escala regional. O banco de dados do aqüífero Serra Geral é composto de dois subconjuntos, 19 variáveis hidroquímicas coletadas em 976 poços, e parâmetros de teste de bombeamento de 156 poços. As duas primeiras áreas de estudo representam dois aspectos do intemperismo, (químico e físico respectivamente) em diferentes cenários investigados neste trabalho; a terceira trata-se de uma aplicação da metodologia para construção de um modelo espacial hidroquímico.

Os seguintes objetivos específicos foram contemplados nessa pesquisa: (1) avaliação do método de interpolação aplicado ao modelo digital de elevação, a partir do qual foram calculadas as variáveis morfométricas, utilizadas como variáveis explicativas, (2) análise das relações entre dados publicados de granulometria, geoquímica de solo e morfometria do terreno, coletados em duas áreas de estudo, utilizando redes neurais artificiais e técnicas de visualização de planos de componentes; (3) identificação de modelos conceituais do processo de intemperismo de solos baseado no método de agrupamento k-média e topografia dos SOMs, para desenvolvimento de modelos preditivos (empíricos e numéricos); (4) geração de mapas preditivos para propriedades físicas e hidroquímicas; e (5) caracterização da incerteza dos vetores quantizados sobre a classificação e previsão das variáveis de solos utilizando técnicas estocásticas de validação cruzada.

As seções seguintes introduzem brevemente os métodos empregados nos trabalhos. Os detalhes de cada estudo são descritos nos artigos (1) *Hillslope chemical weathering across*

Paraná state, Brazil: A data mining-GIS hybrid approach, (2) *Intelligent estimation of spatial distributed soil physical properties,* (3) *Estimating physical-chemical properties in fractured aquifers using self-organizing maps imputation approach: Study of hydraulic connectivity between Serra Geral and Guarani aquifer in Paraná state, Brazil* and (4) *Evaluating SRTM-90m interpolation using SRTM-30m from U.S. territory*.

2. DADOS TOPOGRÁFICOS

Para todos os estudos desenvolvidos, a caracterização do relevo empregou os modelos digitais de elevação extraídos do SRTM – Shuttle Radar Topographic Mission, e disponibilizados pelo Serviço Geológico Americano – USGS (http://edcsns17.cr.usgs.gov/NewEarthExplorer/). Os dados disponíveis possuem resolução espacial original de 90 metros (Farr e Kobrick, 2000). As variáveis morfométricas derivadas são calculadas utilizando a metodologia do projeto TOPODATA (Valeriano, 2008), conduzido pela Divisão de Sensoriamento Remoto do Instituto Nacional de Pesquisas Espaciais – INPE, que gerou dados geomorfométricos para todo o território brasileiro com resolução de 30 metros.

Os dados SRTM foram interpolados para 30 metros de resolução (Figura 2) através de krigagem ordinária, seguindo a metodologia proposta por Valeriano et al. (2006), de onde aspectos morfométricos do relevo como declividade, orientação de vertente, curvaturas vertical e horizontal e fluxo hidrológico acumulado (Jenson e Domingue, 1988) foram computados.

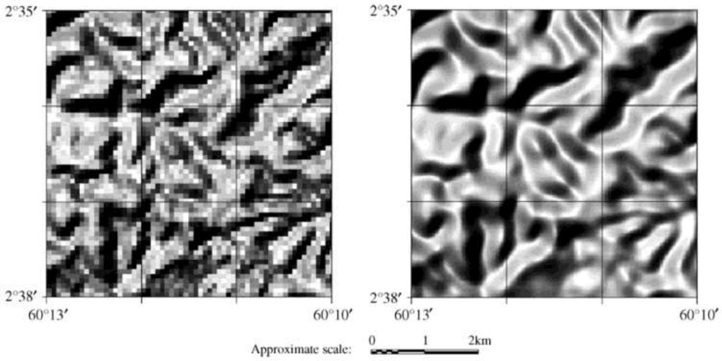


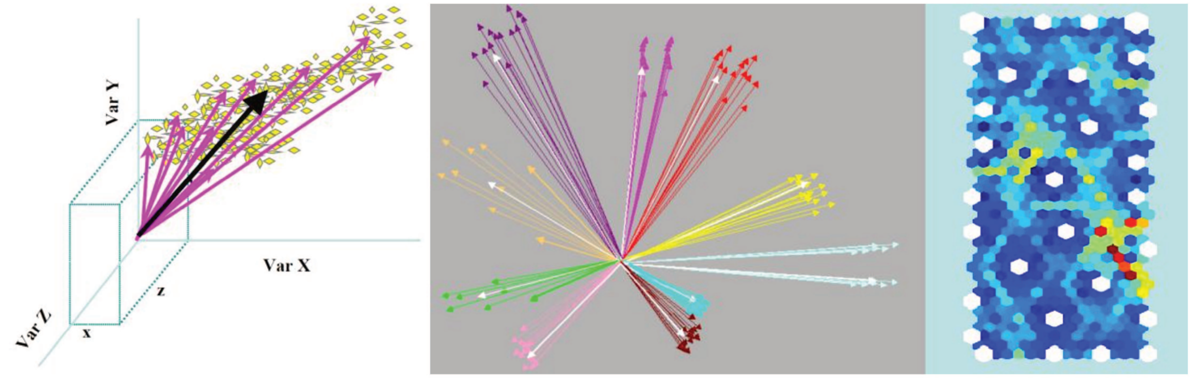
Figura 2 – Comparação dos modelos digitais de elevação com resolução espacial disponível (90m) e com superfície interpolada (30m) (Valeriano et al, 2009).

3. MAPAS AUTO-ORGANIZÁVEIS

Os SOMs pertencem a uma categoria de redes neurais artificiais (RNA) chamada redes de aprendizado por competição (ASCE, 2000). O termo 'auto-organizável' refere-se à natureza do treinamento não supervisionado do algoritmo, que possui a habilidade de organizar, ou classificar, as informações sem especificações sobre o padrão de saída. Os mapas de saída consistem em neurônios organizados em uma grade regular bidimensional, geralmente representados na forma de células hexagonais ou retangulares. Cada neurônio no mapa é representado por um vetor peso multidimensional $m=[m_1, m_2, ..., m_d]$, onde d corresponde à dimensão dos vetores de entrada, ou número de variáveis. Cada neurônio está conectado ao neurônio adjacente por uma relação de vizinhança funcional, que define a topologia ou estrutura do mapa (Vesanto et al., 2000).

A cada amostra é associado um vetor (Figura 3a), cujas propriedades refletem a contribuição da mesma em relação a outras variáveis. A partir da 'nuvem' de vetores é encontrado iterativamente o vetor BMU (Best Matching Unit) através da minimização de distâncias euclidianas para cada uma das variáveis (Kohonen, 2001; Vessanto et al, 2000): $\|x - m_c\| = \min_i \|x - m_i\|$; onde $\|\circ\|$ é a distância euclidiana medida, x é o vetor de entrada, m é o vetor peso e c é o neurônio cujo vetor peso está mais próximo do vetor de entrada x. Os mapas resultantes (Figura 3b) são organizados de tal forma que dados similares são mapeados dentro do mesmo nó ou em nós vizinhos, levando a um agrupamento espacial de padrões identificados nos dados de entrada, i.e., uma classificação dos dados com base em sua topologia no espaço n-dimensional.

Desse modo, cada variável produz um mapa componente, os quais são arranjados em um mapa com dimensões relativas a U-matrix (Ultsch, 2003). Os mapas componentes são freqüentemente utilizados para a visualização das correlações não-lineares. Células com posições e cores similares nos planos componentes descrevem contribuições similares (positivamente correlacionadas) na construção da Matriz unificada. Neste trabalho, foi aplicada a técnica k-médias para o agrupamento das células similares na topografia SOM. Este método de agrupamento particiona n observações em k grupos, onde cada observação pertence ao grupo com a média mais próxima no espaço euclidiano multidimensional, assumindo uma distribuição hiper-esférica dos dados.



(a) (b)

Figura 4 – Cada amostra representa um vetor no espaço *n*-dimensional (a). Os vetores brancos (b) representam a BMU. Estes vetores são iniciados aleatoriamente por um vetor "semente", que iterativamente muda sua posição para se ajustar à nuvem de vetores. Amostras consideradas similares são agrupadas no mesmo neurônio e projetadas em uma mapa bi-dimensional (à direita). As cores representam o grau de dissimilaridade. Cores azuis se referem a baixa dissimilaridade. Tons avermelhados correspondem a altos valores de dissimilaridade (Fraser e Dickson, 2005)

Os vetores são projetados em um toróide - uma figura topológica que dobra sobre si mesma. O toróide é uma representação simples que, quando 'desdobrada', torna-se um retângulo com bordas conectadas, eliminando possíveis problemas de borda. O toróide é explicitamente descrito como: $f(x, y, z) = R - \sqrt{x^2 + y^2})^2 + z^2 - r^2$, onde R e r representam as medidas do maior e menor raio, respectivamente (Figura 4).

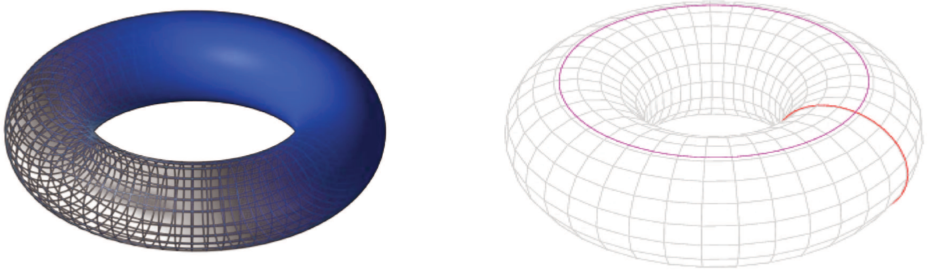


Figura 4 – Representação de um toróide

4. SIMULAÇÃO MONTE CARLO

A simulação Monte Carlo (MC) pode ser definida como um método de simulação estocástica que gera valores aleatórios fornecendo soluções numéricas aproximadas para problemas matemáticos através de experimentos de amostragem computacional (Fishman, 1996). Uma das vantagens do método é a eficiência para lidar com um alto número de parâmetros como de funções analíticas complexas ou problemas combinatórios, especialmente relevantes para o presente trabalho.

A simulação MC é baseada na produção de valores pseudo-aleatórios uniformemente distribuídos. A partir desta básica distribuição de probabilidade contínua, todas as outras distribuições são produzidas, onde os valores simulados devem ser independentes, i.e., o valor gerado em uma realização não influencia sobre o valor do próximo. O método é aplicado para calcular funções integrais definidas, encontrar soluções numéricas para equações diferenciais, para problemas de otimização, análise de incertezas e para resolução de problemas inversos. A simulação S pode ser representada por:

$$S = \int_D g(x)dx$$

onde *D* é o espaço *n*-dimensional definido, *g(x)* é a função objetivo e *x* são os valores aleatórios uniformemente distribuídos de *D*. Basicamente, os valores gerados aleatoriamente são aplicados a uma função de freqüência acumulada de uma distribuição de probabilidade. Seus parâmetros, como média e desvio padrão para distribuição normal, é representado pela função ϕ para o momento *z* por:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\{-x^2/2\}dx$$

A eficiência da simulação estocástica depende do conhecimento sobre o problema e da informação a priori que restringe a simulação, os quais são representados pelos parâmetros das funções de probabilidade calculados a partir do histograma experimental (Krajewski et al., 1991).

5. DISCUSSÃO

Para este estudo, três métodos de modelagem a partir dos resultados dos mapas auto-organizáveis foram aplicados para a (i) identificação e quantificação de correlações não lineares, (ii) estimativa multivariada e (iii) imputação de dados tabulares incompletos.

5.1. Correlações não lineares

O índice de correlação de Spearman, uma medida não paramétrica, é calculado após o rearranjo das amostras determinado pelo SOM, de maneira que correlações não lineares entre variáveis podem ser quantificadas, pois a topologia, i.e., relações de vizinhança, entre as variáveis são preservadas. A investigação de tais relações pode ser reforçada através da análise de diagramas de dispersão. Tais diagramas são compostos pelos neurônios que compõem a matriz unificada, onde os vetores associados às amostras durante o processamento dos mapas organizáveis podem ser agrupados por similaridade em um mesmo neurônio. Isso gera um diagrama cujo número de elementos representados não corresponde ao número de amostras originais.

A Figura 6 apresenta gráficos de espalhamento das amostras de geoquímica de solos coletadas no estado do Paraná, segmentadas em três grupos através do método k-médias e representados em cores distintas. Os elementos apresentaram um padrão de denudação química relacionado com características topográficas em escala local e regional. Altas elevações são mais intemperizadas (Figura 6a), apresentam baixa concentração de elementos móveis e altos teores dos mesmos elementos em baixas altitudes, indicando um mecanismo de transporte e deposição. A curvatura vertical descreve a mesma variação intempérica em escala local (Figura 6b). Em áreas convexas, próxima ao divisor local, a concentração dos elementos mais suscetíveis a mobilização são baixas, enquanto próximo ao canal de drenagem os valores são maiores.

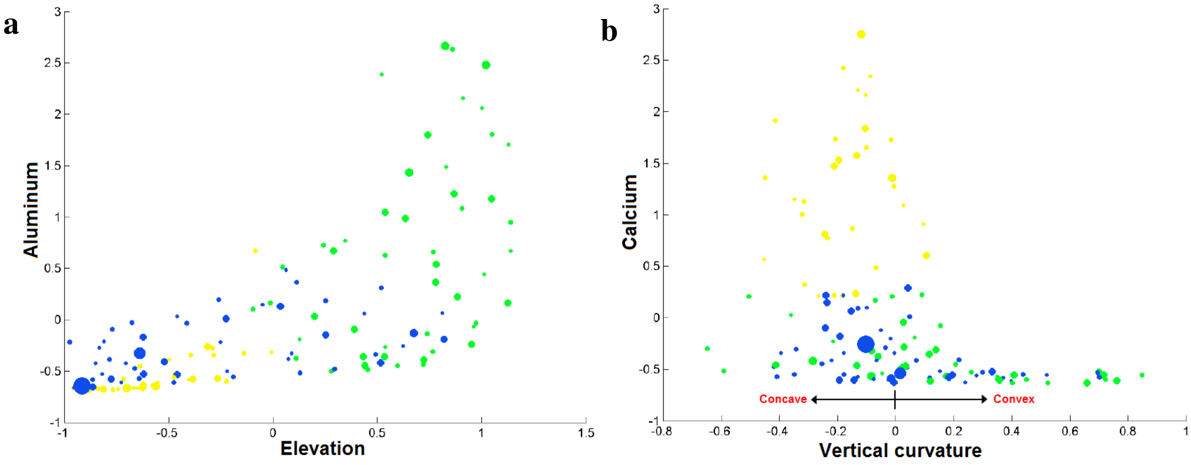
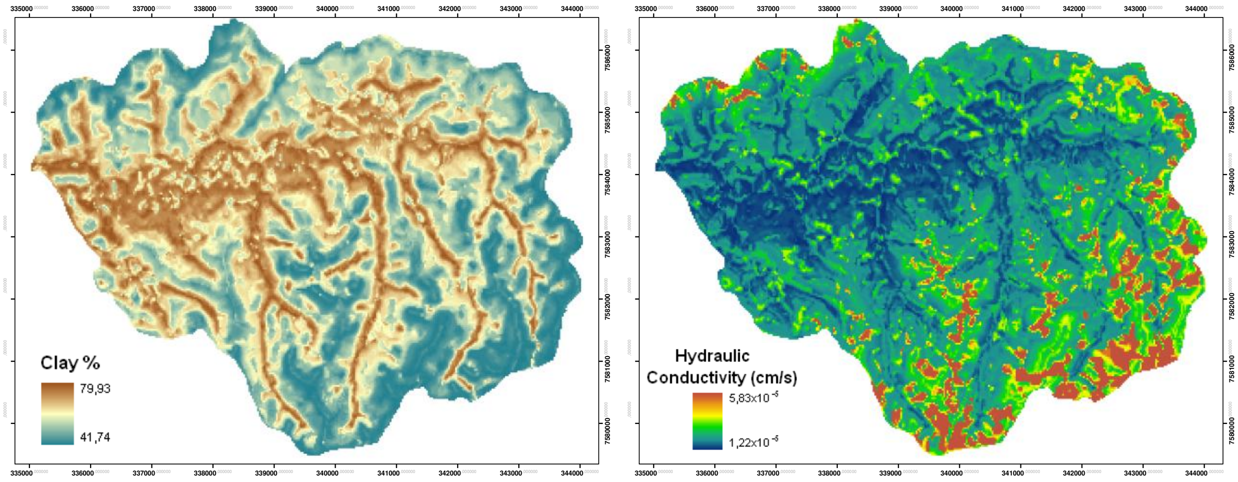


Figura 6 – Gráfico de espalhamento dos nós em valores padronizados para (a) alumínio *versus* elevação, e (b) cálcio *versus* curvatura vertical, representando o formato côncavo-convexo da vertente.

5.2. Modelos preditivos

Os mapas auto-organizáveis permitem o cálculo de estimativas utilizando um modelo de treinamento previamente construído. A simulação Monte Carlo foi utilizada para geração de valores aleatórios utilizando a matriz de correlação como restrição, onde os valores simulados mantêm a correlação existente entre as variáveis das 29 amostras originais, coletadas em campo. A simulação estocástica, combinada com mapas auto-organizáveis, atenua problemas causados pelo baixo número de amostras. Os SOMs utilizam todas as variáveis independentes ou potencialmente explicativas no processo preditivo sem penalizar a proporção de variabilidade explicada pelo modelo.

A Figura 7 apresenta os mapas resultantes das estimativas geradas pelo SOM treinado a partir dos dados gerados pela simulação Monte Carlo para dados de granulometria do solo e condutividade hidráulica coletados em Poços de Caldas. Um padrão de intemperismo análogo ao da geoquímica de solos no estado do Paraná foi identificado para propriedades físicas do solo (Figura 7), onde a proporção de areia apresenta valores altos em regiões próximas aos topos de morros e o conteúdo de argila é maior em áreas côncavas e próximas a canais de drenagem, ambas altamente correlacionadas com a condutividade hidráulica.



(a) (b)

Figura 7 – Exemplo de mapas preditivos para teores de (a) argila e (b) condutividade hidráulica do método proposto.

5.3. Imputação de dados

A imputação de dados é um método para lidar com tabelas incompletas, estimando valores com base na correlação existente entre os dados disponíveis (Malek et al., 2008). A matriz de dados completa permite a aplicação de outros modelos estatísticos multivariados, mesmo paramétricos, fundamentados no cálculo de auto-valores e auto-vetores.

A capacidade de aprendizado dos vetores quantizados, que utiliza distância Euclidiana e preserva relações topológicas, caracteriza o SOM como um método de imputação inerentemente robusto (Dickson e Giblin, 2007). A imputação de dados da hidroquímica do aqüífero Serra Geral no estado do Paraná incluiu um cenário geológico-estrutural diverso, incluindo as distintas seções hidrogeológicas descritas na literatura. A elaboração de superfícies contínuas dos valores imputados de hidroquímica do aqüífero Serra Geral permitiu identificar padrões espaciais do comportamento hidroquímico, e elaborar modelos espaciais de conexão hidrogeológica, recarga e confinamento (Figura 8).

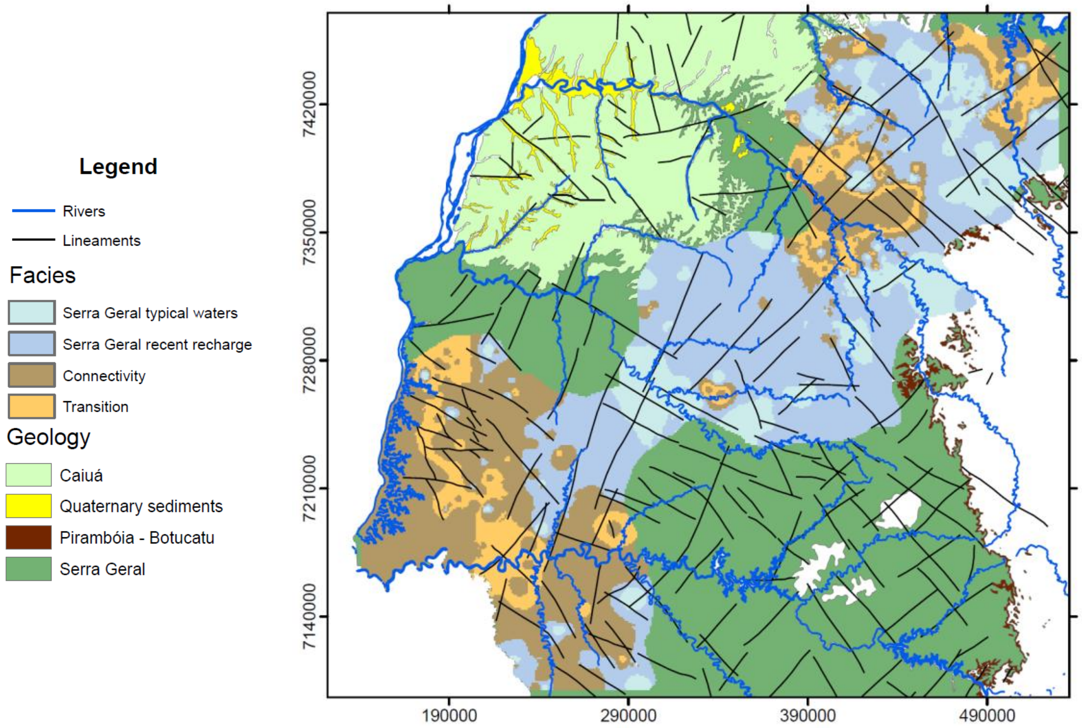


Figura 8 – Modelo espacial de seções hidrogeológica do sistema aqüífero Serra Geral no estado do Paraná.

5.4. Validação cruzada

As incertezas dos modelos produzidos foram avaliadas através dos métodos *leave-one-out* e *Bootstrap,* onde o modelo construído pelas amostras de treinamento é aplicado sobre os dados de validação. O processo consiste em retirar uma amostra, estimar a mesma com o modelo preditivo e comparar a diferença entre os valores. A validação cruzada estocástica foi aplicada em cada amostra e foram rodadas 30 realizações para cada variável analisada. A média dos resíduos é representada em um gráfico de valores observados *versus* preditos. Em um cenário ideal, deve existir uma correspondência 1:1; i.e., o resultado da validação deve se apresentar próximo de uma curva $y = x$, onde os resíduos são utilizados para verificar a existência de tendência no modelo.

6. CONSIDERAÇÕES FINAIS

Foram propostas três combinações metodológicas para a modelagem espacial do intemperismo físico e químico dos solos, onde foram exploradas as correlações de natureza não linear entre variáveis geoquímicas, elaborados modelos preditivos de propriedades físicas do solo sob a limitação de um restrito número de amostras, e construído um modelo espacial da hidroquímica do aqüífero fraturado Serra Geral a partir de tabelas incompletas.

A morfometria do terreno e suas medidas descritivas se apresentaram como variáveis de importante potencial preditivo, relação não observada através de estatística linear multivariada. As correlações não lineares entre a geoquímica de solos e as medidas derivadas do relevo permitiram a discussão sobre o papel da denudação química sobre processo global de intemperização. O modelo preditivo de propriedades físicas do solo apresentou coerência com conceitos de transporte mecânico de partículas, e exibiu estacionariedade de 2ª ordem, verificando-se a não existência de tendências. O processo de imputação permitiu o cálculo da matriz de correlação das variáveis hidroquímicas sobre tabelas incompletas. Sua espacialização e comparação com modelos conceituais existentes permitiu a visualização do comportamento de cada amostra estimada.

O uso de mapas auto-organizáveis, balizado pela simulação estocástica com subseqüente classificação de agrupamento pela técnica k-médias, permitiu a elaboração de modelos conceituais e espaciais. Espera-se que a metodologia aqui proposta, que emprega métodos estocásticos em combinação com mineração de dados, proporcione uma alternativa robusta e não tendenciosa para a análise de grandes volumes de dados a ser aplicada nos mais diversos cenários em estudos sobre evolução da paisagem.

REFERÊNCIAS

Ehsani, A. H., Quiel, F., 2008. Geomorphometric feature analysis using morphometric parametrization and artificial neural networks. Geomorphology 99, 1-12.

Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. American Geophysical Union EOS 81, 583–585.

Fraser, S. J., Dickson, B., 2005. Ordered vector quantization for the integrated analysis of geochemical and geoscientific data sets. 22nd International Geochemical Exploration Symposium, Association of Applied Geochemists, Perth, Australia.

Fishman, G. S., 1996. Monte Carlo - concepts, algorithms, and applications, Springer-Verlag, Berlin.

Heimsath, A. M., Dietrich, W. E., Nishizumi, K., Finkel, R. C., 1997. The soil production function and the landscape equilibrium. Nature 388, 358-361.

Hentati, A., Kawamura, A., Amaguchi, H., Iseri, Y., 2010. Evaluation of sedimentation vulnerability at small hillslide reservoir in the semi-arid region of Tunisia using Self-Organizing Map. Geomorphology. Accepted, in review.

Iwashita, F., Souza Filho, C. R., 2008. Avaliação da interpolação de dados SRTM-90 m através de dados SRTM-30 m do território americano. Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, INPE, Natal, Brasil, pp. 3927-3934.

Jenson S. K. and J. O. Domingue., 1988. Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis. Photogrammetric Engineering and Remote Sensing 54, 1593–1600.

Krajewski, W. F., Lakshmi, V., Georgakakos, K. P., Jain, S., 1991. A Monte Carlo study of rainfall sampling effect on a distributed catchment model. Water Resources Research 27, 119-128.

Kohonen, T., 2001. Self-organizing Maps, third edition, Springer-Verlag, Berlin.

Löhr, S. C., Grigorescu, M., Hodgkinson, J. H., Cox, M. E., Fraser, S. J. 2010. Iron Occurrence in soils and sediments of coastal catchment: A multivariate approach using self-organizing maps. Geoderma 156, 253-266.

Malek, M.A., Harun, S., Shamsuddin, S.M., Mohamad I., 2008. Imputation of time series data via Kohonen self organizing maps in the presence of missing data. Engineering and Technology 41, 501-506.

Mudd, S. M., Furbish, D. J. 2004. Influence of chemical denudation on hillslope morphology. Journal of Geophysical Research 109, F02001.

Neter, J., Kutner, M. N., Nachtssheim, C. J., Wasserman, W. 1996. Applied linear statistical models, 4th Ed. WCB/McGraw-Hill, Boston.

Penn, Br. S. 2005. Using self-organizing maps to visualize high-dimensional data. Computer & Geosciences 31, 531-544.

Reimann, C., Filzmoser, P., Garret, R. G., 2002. Factor analysis applied to regional geochemical data: problems and possibilities. Applied Geochemistry 17, 185-206.

Roering, J. J., Kirchner, J. W., Dietrich, W., 1999. Evidence for nonlinear, diffusive sediment transport on hillslopes and implications for landscape morphology. Water Resources Research 35, 853-870.

Valeriano, M. M., Kuplich, T. M., Storino, M., Amaral, B., Mendes, J. N., Lima, D. J., 2006. Modeling small watersheds in Brazilian Amazonia with shuttle radar topographic mission-90 m data. Computer & Geosciences 32, 1169-1181.

Valeriano, M. M., Rosetti, D. F., Albuquerque, P. C. G., 2009. TOPODATA: desenvolvimento da primeira versão do banco de dados geomorfométricos locais em cobertura nacional. Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, INPE, Natal, Brasil, pp. 5499-5506.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 2000. SOM Toolbox for Matlab 5. SOM toolbox team, Finland. Helsinki University of Technology, Laboratory of Computer and Information Science. http://www.cis.hut.fi/projects/somtoolbox/

Yoo, K., Amundson, R., Heimsath, A. M. Dietrich, W. E., Brimhall, G. H., 2007. Integration of geochemical mass balance with sediment transport to calculate rates of soil loss chemical weathering and transport on hillslopes. Journal of Geophysical Research 112, F02013.

Young, A., 1980. Tropical soils and soil survey. Cambridge University press, Cambridge.

HILLSLOPE CHEMICAL WEATHERING ACROSS PARANÁ, BRAZIL: A DATA MINING-GIS HYBRID APPROACH

Abstract

Self-organizing map (SOM) and geographic information system (GIS) models were used to investigate the nonlinear relationships associated with geochemical weathering processes at local (~100 km^2) and regional (~50.000 km^2) scales. The data set consisted of 19 B-horizon soil variables (P, C, pH, Al, total acidity, Ca, Mg, K, total cation exchange capacity, sum of exchangeable bases, base saturation, Cu, Zn, Fe, B, S, Mn, gammaspectrometry (total count, potassium, thorium, uranium) and magnetic susceptibility measures) and six topographic variables (elevation, slope, aspect, hydrological accumulated flux, horizontal curvature and vertical curvature) characterized at 304 locations from a quasi-regular grid spaced about 24 km across the state of Paraná. This data base was split into two subsets: one for analysis and modeling (274 samples) and another for validation (30 samples) purposes. The self-organizing map and clustering methods were used to identify and classify the relations among solid-phase chemical element concentrations and GIS derived topographic models. The correlation between elevation and k-means clusters related the relative position inside hydrologic macro basins, which was interpreted as an expression of the weathering process reaching a steady-state condition at the regional scale. Locally, the chemical element concentrations were related to the vertical curvature representing concave-convex hillslope features, where concave hillslopes with convergent flux tends to be a reducing environment and convex hillslopes with divergent flux, oxidizing environments. Stochastic crossvalidation demonstrated that the SOM produced unbiased classifications and quantified the relative amount of uncertainty in predictions. This work strengthens the hypothesis that, at B-horizon steady-state conditions, the terrain morphometry were linked with the soil geochemical weathering in a two-way dependent process, the topographic relief was a factor on environmental geochemistry while chemical weathering was for terrain feature delineation.

Keywords: self-organizing map, hillslope chemical weathering, geomorphometry, uncertainty, Brazil

1. INTRODUCTION

Terrain morphometric features reflect the physical and chemical weathering processes by which they were created (Heimsath et al., 1997). Understanding weathering therefore requires knowledge of phenomena that influence the landscape formation. Early modeling approaches were used to quantify weathering from a physical mass balance viewpoint (Roering et al., 1999). Empirical models to survey multidimensional geochemical data were developed using multivariate statistical methods that included multiple linear regression (Stewart et al., 2003), principal component analysis (Reimann et al., 2002), and cluster analysis (Hanesch et al., 2001). For these models to be reliable, however, the data had to be normally distributed, stationary, and have no co-linearity among independent (explanatory) variables (Netter et al, 1996). In addition to penalizing higher numbers of explanatory variables (Netter et al., 1996), these techniques resulted in losing important nonlinear associations. These assumptions are particularly problematic, because according to Reimann and Filzmore (1999), at the regional scale, geochemical data does not have normal or lognormal distribution. For these and other reasons, an alternative is the development and application of numerical models.

Early numerical models considered the hillslope to be uniform (rectilinear) along its extension with no provision for transport and deposition rate heterogeneity. Investigators improved on this model type by introducing a nonlinear sediment transport rate through morphometry characterized by a convex hilltop, rectilinear middle section, and concave base (Roering et al., 1999). Mudd and Furbish (2004) formulated a model that coupled physical sediment transport to chemical deposition-denudation in the hillslope weathering process. One simplifying assumption in their model was constant elevation over the time period being modeled. Application of this model revealed that the total amount of mass transported by chemical weathering increased nonlinearly with distance from the hillslope ridge, while at the rectilinear inflexion point the mechanical transport began to decrease (Fig. 1a). Yoo et al. (2007) applied a similar model to soil geochemical measurements collected along a sampling traverse in southeastern Australia. Their simulated chemical weathering rates revealed a hillslope mass loss near the divide and an accumulation near the base (Fig. 1b). Along the hillslope, three distinct geochemical environments were recognized based on the concentration of predominate dissolved ions: (1) Si, Al, and Fe at the hillslope top indicated an oxidizing environment with decreased weathering rates towards the base; (2) Ca, Mg, Na, K at the rectilinear section indicated a neutral pH environment; and (3) P and Ca at the base indicated a reducing environment in which gains in mass were comparable with losses in the upper sections. This finding, together with simulations indicating higher soil moisture content in concave areas compared with convex areas, demonstrated the direct link between element mobility and soil physical-chemical conditions, such as moisture, pH, temperature, and porosity.

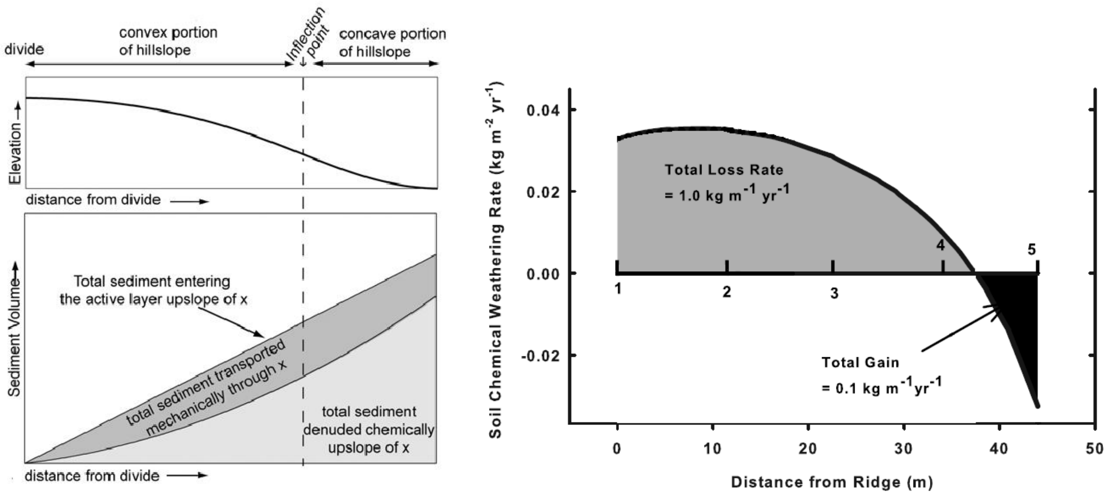


Fig. 1. A conceptual model of sediment transport rate based on the hillslope shape. a) The total sediment denuded chemically is a nonlinear function of distance from the divide. The chemical amount denudated at the concave area is larger than the mechanically transported (after Mudd and Furbish, 2004), b) Total soil weathering rate, showing the role of soil production on soil thickness and the role of soil transport on soil thickness and chemistry (after Yoo et al., 2007).

Some challenges in the construction and application of numerical hillslope models are their one-dimensionality, steady-state requirements, lack of calibration data, and nonuniqueness. Also, numerical models commonly are too rigid with respect to detecting unexpected features like the onset of trends, non-linear relations, or patterns restricted to sub-samples of a data set. These shortcomings created the need for an alternate modeling approach capable of using available data. One technique that is well-suited to noisy, sparse, nonlinear, multidimensional, and scale-dependent data is a type of unsupervised artificial neural network called the self-organizing map (Kohonen, 2001). The self-organizing map (SOM) technique has been used in related studies to explore relations among rock geochemistry and hyper-spectral images (Penn, 2005), classify geomorphometric aspect based on digital elevation models (Ehsani and Quiel, 2008), characterize hillslope landslide vulnerability (Hentati et al., 2010), identify processes controlling the distribution of iron in soil and sediment (Löhr et al., 2010), and investigate the geochemistry in shallow groundwater (Friedel et al., *in review*). The aim of this study is to understand scale-dependent relations among soil geochemical weathering and morphometric features across the state of Paraná in southeastern Brazil. The hypothesis is that a conceptual hillslope weathering model can be devised based on the statistical relations among field data and GIS metrics. To achieve the goal and to satisfy the hypothesis, the following objectives are undertaken: (1) analyze nonlinear relations among published B-horizon soil geochemical, environmental, relief morphometry, and GIS data from 304 locations using the SOM (Kohonen, 2001) and component planes visualization (Penn, 2005) techniques; (2) identify conceptual models of soil geochemical weathering processes based on k-means clustering (Vesanto and Alhoniemi, 2000) of the SOM topography for future development of predictive (empirical and numerical) models; and (3) evaluate bias and uncertainty in the quantized vector predictions and soil classifications using a stochastic cross validation technique (Rao et al, 2008).

2. STUDY AREA

Paraná is a state of Brazil, located in the South of the country. According to the Instituto Brasileiro de Geografia e Estatística-IBGE, the state covers about 199.314 km^2 and is home to about 10 million people living in 399 cities. Its gross domestic product ranks fifth in Brazil, producing about 6,2% of the national wealth. The predominant climate is characterized as subtropical with warm summers and cold winters. According to the Köppen classification, the subtropical climate has three variants: Cfa, Cfb and Af. The annual average temperature varies from 14°C to 22°C with a slightly colder climate occurring along the southern plateau, and the annual average precipitation ranges from 1500 mm to 2500 mm.

According to Licht (2001), the combination of climate and hillslope geomorphology is reflected in the dense and perennial stream network. The primary hydrologic divide, called the Serra do Mar, separates the coastal plain from other geomorphologic units in the state. This divide is associated with four regional macrobasins with tributaries to the Paraná River: Iguaçu, Ivaí, Piriqui, and Tibagi. Collectively, these regional basins host 63 protected reserves (covering about 1.187.000 hectares); 16 reserves with dense ombrophyla forest structure, 31 with mixed ombrophyla, and 16 with semideciduous seasonal forest structure (IBGE, 2010).

The geological record of Paraná is characterized by the crystalline shield, composed of Precambrian magmatic and metamorphic rocks, covered by Paleozoic and Mesozoic volcanic and sedimentary rocks comprising the Paraná basin (Fig. 2). This coverage was eroded due to uplift of the continental crust, east of the basin, exposing the basement. Tertiary and quaternary sediments partially overlay the basin and shield rocks. The crystalline basement, formed by igneous and metamorphic rocks with ages varying from Achaean to Proterozoic, is locally covered by volcano-sedimentary, sedimentary and unconsolidated sediments sequences. The crystalline shield encompasses a mega-belt formed on late Precambrian by the collision of continental and micro-continental blocks. The basin includes a second and third plateau that covers most of the state. It is a sedimentary basin, overlain by Cretaceous basalt, the Serra Geral Formation (dark green in Fig. 2), intracratonic, evolved over the South American platform and its generation began during the Silurian period and ended in the Cretaceous period (Minerais do Paraná – MINEROPAR, 1986).

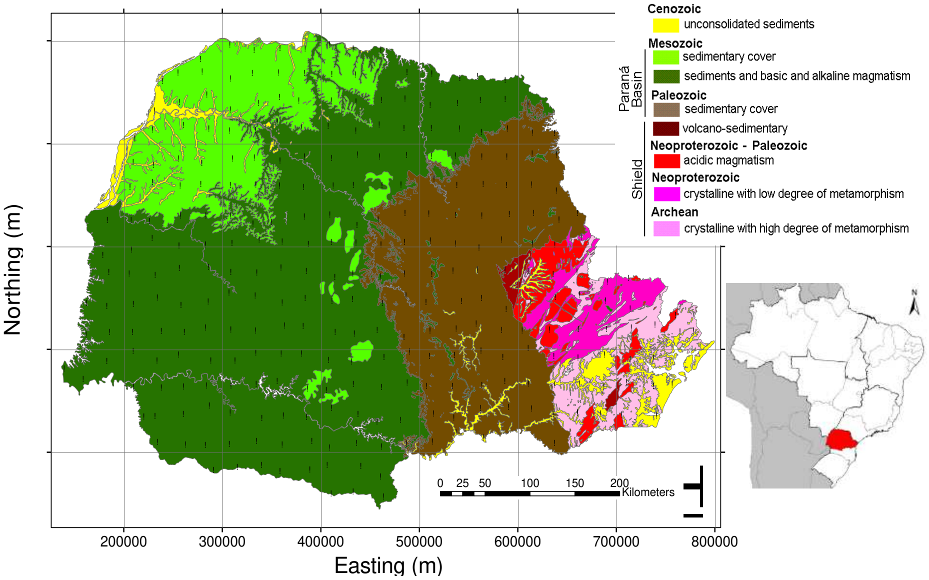


Fig. 2. Simplified geological map of Paraná state (modified from Lich, 2001) and location of the samples.

3. METHODS

Five steps were used to identify hillslope weathering relations linking the soil geochemistry to relief morphometric features. First, all data variables were standardized so that no one variable would dominate in the nonlinear modeling process (Kalteth et al., 2008). The z-score transformation is given by:

$$z_i = \frac{x_i - \bar{x}_i}{s_i} \qquad (1)$$

where *z* is the standardized value; *x* is the raw score; $\bar{x}$ is the sample average, and *s* is the sample standard deviation, i is an index for each variable. Standardizing variables in this way resulted in each having an expected value of zero and standard deviation one. Second, after the standardization data were split into two subsets: training (n = 274) and validation (n = 30). Third, the SOM (Kohonen, 2001) was used to self organize nonlinear relations among the 28 variables. Fourth, the k-means clustering technique (Forgy, 1965) was used to classify the SOM topography into statistically relevant conceptual models (Ehsani and Quiel, 2008). Finally, the geochemical concentrations were interpreted based on terrain morphometry and associated clusters.

3.1. Self organizing map

The SOM belongs to a subcategory of the artificial neural network (ANN) algorithms, called competitive learning networks, in which the computational models serve as a proxy for neurons in the human brain (ASCE, 2000). The term self-organizing is based on the unsupervised nature of the algorithm having the ability to organize information without any prior knowledge of an output pattern. In this study, the output consists of neurons organized on a two-dimensional rectangular grid having hexagonal cells (map). Each neuron in the map is represented by a multi-dimensional weight vector $m=[m_1, m_2, ..., m_d]$, where d correspond to the dimension of the input vectors. Each neuron is connected to the adjacent neuron through a functional neighborhood relation (Vesanto et al., 2000). Individual data samples are associated to a vector with properties that reflect its contributions relative to the other variables. From this cloud of data vectors, a best matching unit (BMU) is iteratively determined by minimizing the Euclidean distance measure for each variable (Kohonen, 2001; Vessanto et al, 2000):

$$\|x - m_c\| = \min_i \|x - m_i\| \qquad (2)$$

where $\|\circ\|$ is the euclidian distance, x is the input vector, m is the weight vector and c is the neuron whose vector is nearest to the input vector x. The resulting maps are organized in such way that similar data are mapped to the same or nearby nodes, and dissimilar data are mapped to nodes with greater separation distances. According to Kalteth et al. (2008) the map size, i.e. the number of nodes that will be projected into the map, plays an important role on the training process, once they determine the number of clusters where the samples will be assigned to. Vessanto et al. (2000) proposed a heuristic method to calculate the number of nodes based on a formula and the ratio between the two largest eigenvalues from the covariance matrix. However, this approach would not be practicable on a database with missing values or categorical variables. An alternative approach would be to find a suitable small topographical error, a measure (percentage) of the number of node vectors that are adjacent in n-dimensional space, but are not adjacent on the resulting self-organized map. In other words, would be the error from the rearrangement and from the 'flattening', when projecting n-dimensional data into a two-dimensional array.

The unified matrx (U-matrix) is comprised of the BMUs obtained from weight vectors linked to the input vectors; thus, each variable produces a component plane arranged in a grid that is related to the SOM matrix (Kohonen, 2001). These maps are often used to visualize correlations among variables; for example, cells with similar colors and positions inside component planes describe similar contributions (positively correlated) in the construction of the U-matrix. This information is projected onto a toroid which is unwrapped into a rectangle for viewing the SOM topography. In this study, the k-means clustering method was used to classify common cells in the SOM topography, represented in the U-matrix (Ultsch, 2003).

3.2. Topographical dataset

Characterization of the topographic relief was possible using elevation data provided by the Shuttle Radar Topographic Mission (Farr and Kobrick, 2000). The digital elevation model associated with these data was provided by the United States Geological Survey on a lattice with 90-m spatial resolution. The Topodata project, conducted by the Brazilian National Institute for Space Research-INPE (Valeriano et al., 2009), has created derived metrics data with a 30-m resolution, based on elevation data and a geographical information system (GIS) modeling techniques. The geomorphometric features provided a way to extract morphometric features, such as slope, aspect (hillslope orientation), vertical and horizontal curvature (Valeriano et al., 2006), and accumulated hydrological flux (Jenson and Domingue, 1988).

The variable slope represents the first derivative of two locations on the elevation data, while the second derivative produces the variable aspect, which indicates the position of the hillslope relative to the north. Regarding the variable aspect, it varies from 0 to 360º, with value zero pointing towards north. Since both zero and 360º represent the north, a trigonometric (cosine) transformation was applied so that these values varied from -1 (south) to 1 (north). Another derived measure, the vertical curvature depicts the hillslope profile (Fig. 3): convex, rectilinear, and concave shape, whereas the horizontal curvature is the hillslope shape when represented on the horizontal plane, describing a divergent, planar or convergent hydrological flux. These two features are highly correlated but when analyzed in combination they produce different hillslopes shapes, which could lead to a soil with distinct physical-chemical properties. The last modeled variable, hydrological accumulated flux is a measure of the number of terrain units that converge at the element being analyzed. It is used as a proxy for the distance from the ridge (Fig. 4).

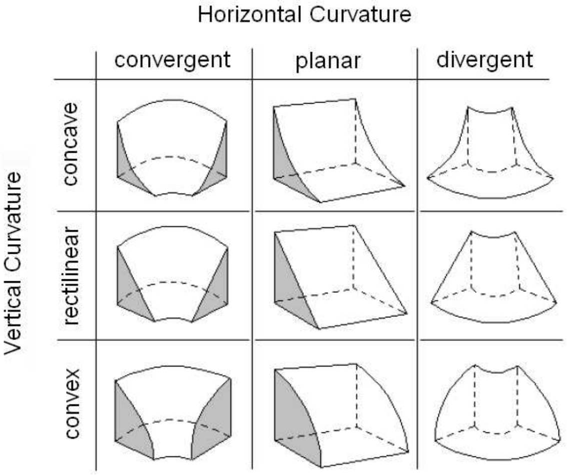


Fig. 3. Morphometric variables, where vertical curvature represents profile slopes, and horizontal curvature describes convergent or divergent fluxes (Valeriano et al., 2009).

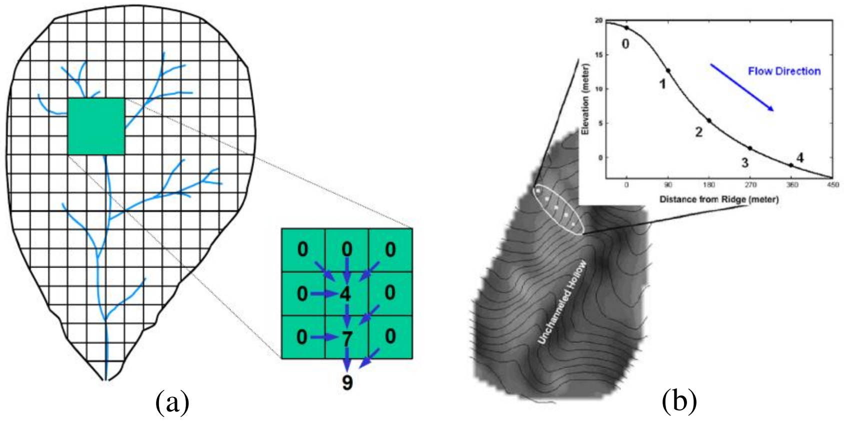


Fig.4. Schematic showing: a) the accumulated hydrologic flux (modified from Rennó, 2003); and b) the flow direction used as surrogate for the distance from the ridge (afer Yoo et al., 2007).

3.3. Data variables

To effectively capture random spatial variability of geochemical and hydrological processes, field sampling of B-horizon soil samples was conducted using a quasi-regular grid across Paraná (Park and Giesen, 2004). The Paraná Agronomic Institute performed analysis of the elements, where the analytical methods and equipment calibration procedures for geophysics measurements are described in detail on Mineropar (2005), this report also includes the descriptive statistics for each element. The following variables were analyzed: pH, Al (mg/kg), Ca (cmol$_c$/kg), Mg (cmol$_c$/kg), P (cmol$_c$/kg), K (cmol$_c$/kg), organic carbon (g/kg), H^++Al^{3+}(total acidity, in cmol$_c$/kg), Cu (mg/kg), Zn (mg/kg), Fe (mg/kg), Mn (mg/kg), S (mg/kg), B (mg/kg), V% (base saturation), cations exchange capacity (sum of exchangeable bases: Ca^{2+}+Mg^{2+}+K, in cmol$_c$/kg), T (total exchangeable cation charge: V%+CTC, in cmol$_c$/kg), gamma-spectrometry – channels total count (cps), Uranium (ppm), Potassium (%) and Thorium (ppm), and magnetic susceptibility (dimensionless). The data was assembled into a data base by the Paraná State Geological Survey and provided by project personnel (Licht, 2001).

The analytical methods have a maximum error of 5% and the detection limits are presented in Table 1, where the analyzed elements are commonly used by agricultural community to measure soil fertility. The digest solutions (Table 1) are designed to extract the soil fraction weakly bonded, i.e., not the total content, but the most susceptible portion to weathering processes, and available as nutrients for plants. The analytical method is suitable for this work, since these elements, present on this fraction of the soil, are easily released and can be associated to hydrological fluxes on subsurface, and thus to relief and hillslope morphometric features.

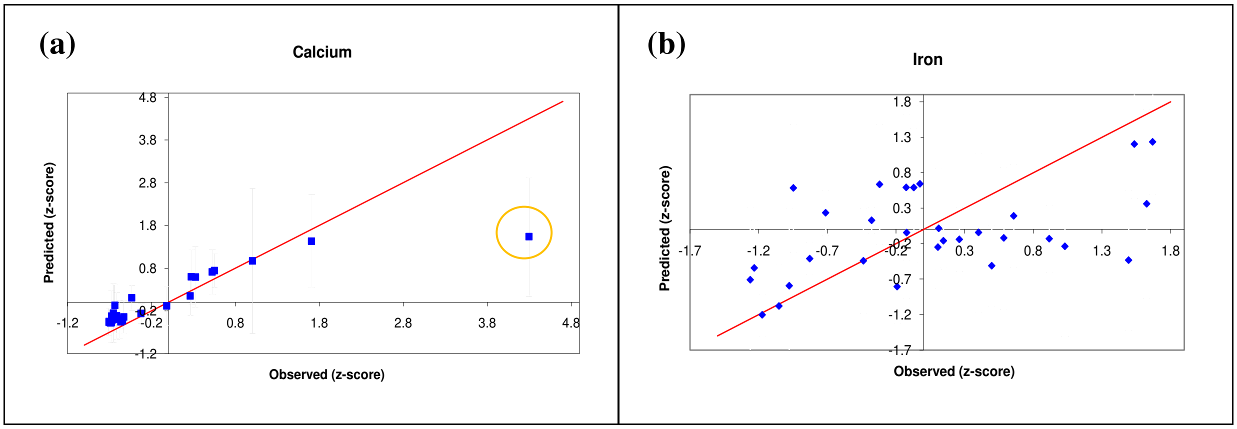
Table 1 – Soil geochemistry extraction methods

Variable	Extractor	Ratio soil to solution
pH	Calcium chloride 0.01 M	1:2.5
Al, Ca, Mg	Potassium chloride 1.0 M	1:10
P, K	Mehlich I (HCl 0.05 N + H_2SO_4 0.025 N)	1:10
Organic C	Walkley-Black ($K_2Cr_2O_7$ + H_2SO_4 conc.)	
H+Al (total acidity)	Buffer solution SMP	
Cu, Zn, Fe	Hydrochloric Acid 0.1 M	1:10
Mn	Ammonium Acetate 1.0 M, pH 7	1:10
S	Calcium Mono Phosphate + acetic acid 2M	1:2.5
B	Hydrochloric Acid 0.05 N	1:2

4. RESULTS AND DISCUSSIONS

4.1. Cross validation

The model performance was evaluated using a stochastic cross-validation approach (Rao et al, 2008). The approach consisted of five steps: leave out one sample, recreate a new SOM, estimate values, and analyze residuals. This process was applied to each variable 30 times. For each variable, the average prediction value for 30 realizations was computed and plotted against observed values to assess model bias (Fig. 5). Aside from one outlier in the Ca and Al predictions (confidence interval of 95%), the SOM model demonstrated unbiased behavior indicated by the one-to-one correspondence and constant variance for all variables. This validation process provided confidence that analysis of SOM-based hillslope relations, such as component maps, k-means clusters, and scatterplots, are meaningful.



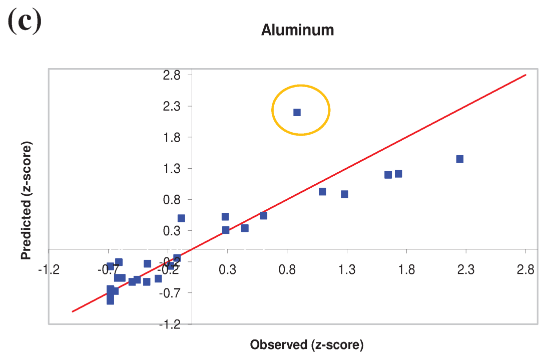


Fig. 5. Cross validation average of 30 trials with 30 samples, using leave one out strategy for a) calcium, b) iron and c) aluminum. One to one correspondence reveals that the SOM is an unbiased predictive model. Orange circles indicate outliers that are inconsistent with the nonlinear model (p-value = 0.05).

4.2. Component map analysis

A component plane can be thought of as a slice of the SOM, it actually represents one set of vector component (variable) values in all map units; that is, each component plane portrays the spread of values for the associated variable. In that regard they are similar to histograms, the difference being that the same value can be present in multiple places of the map when it belongs to different clusters, allowing an easy visualization of nonlinear correlations between variables based on the color array of the map, not fully captured by linear multivariate approaches (Astel et al., 2007). Vessanto et al., (2000) suggest a formula to calculate the optimum size for a SOM: $m = 5\sqrt{n}$, where m is the number of units (nodes or neurons) and n is the number of samples. Considering the number of training samples (n = 274), the map size would approximately have 83 nodes, and the ratio of side lengths should be based on the ratio of the two largest eigenvalues from the covariance matrix. Thus the product of the side lengths should be close to 83. The eigenvalues for factor one is 6.26 and factor two, 3.84. The ratio between them is 1.63. To find the appropriate side lengths, solving a simple system of equations is required: $\begin{cases} \frac{x}{y} = 1.63 \\ x * y = 83 \end{cases}$

Solution of this linear system leads to x = 11.5 and y = 7.2 giving or a 12 x 7 map. However, based on the number of training samples (n = 274) and considering the topographical error, a map size of 16x10 was chosen. The topographic error is a measure (percentage) of the number of node vectors that are adjacent in n-dimensional (variable) space, but are not adjacent on the resulting self-organized map. In other words, this would be an error associated with the samples rearrangement and with the 'flattening', when projecting an n-dimensional data into a two-dimensional space array, i.e., a low topographical error means that the original neighborhoods were better preserved (Dickson and Fraser, 2007).

Inspection of the component maps (Fig. 6) revealed several relations, for example, the elements B, Ca, Mg, K, Cu, P and Zn are strongly correlated, where, given the regional geology, Cu and Zn are linked to the presence of mafic rocks from the second and third plateau, and K, Mg and Ca, are associated with aluminosillicates rocks, as the basalt from Serra Geral Formation (Licht, 2001). In contrast, inverse relations exist between pH and Al, and Fe and aspect. The Al content is directly linked to the total acidity in the soil, so a high concentration of Al implies a low value of pH. Regarding Fe, the inverse correlation with respect to aspect may be due to the type of iron present. Specifically, a hillslope facing north (value of one) is more exposed to the sun and therefore subject to oxidizing conditions; oxidized Fe is less mobile than the reduced form (Gerrard, 1992).

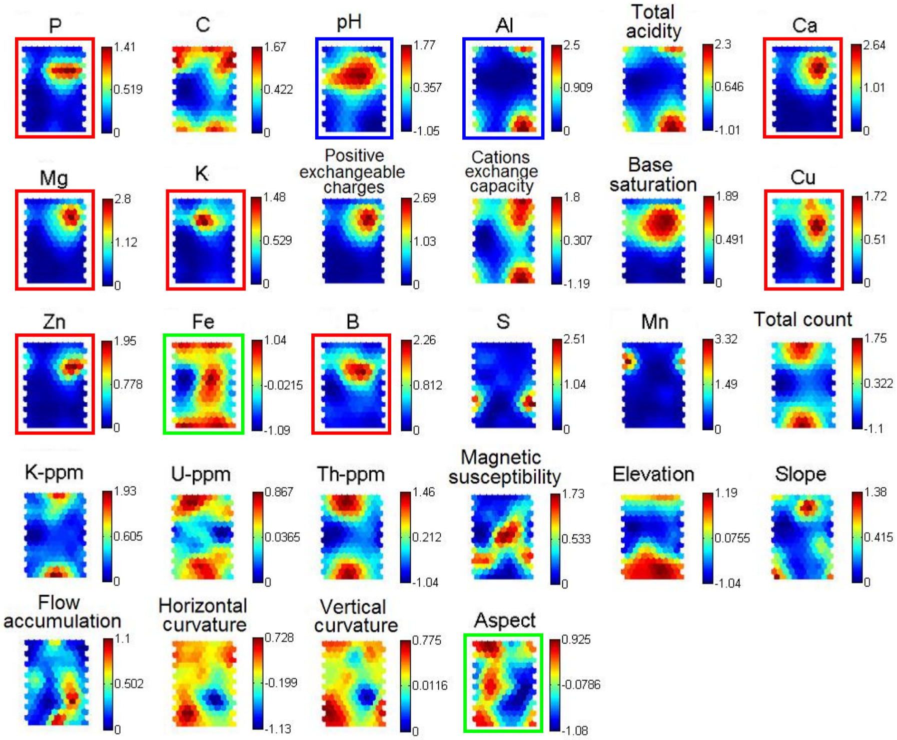


Fig. 6. Component planes used to visualize nonlinear correlation. For example, the cations highlighted by boxes in similar colors, B, Ca, Mg, K, Cu, P, Zn are correlated (similar colors), whereas pH is inversely correlated with aluminum and iron is inversely correlated with aspect (opposite colors). P is phosphorus mg/kg, C is organic carbon in g/kg, pH is the acidity, Al is aluminum cmol$_c$/kg, Total acidity is the content of H^++Al^{3+} in cmol$_c$/kg, Ca, is calcium cmol$_c$/kg,

Mg is magnesium cmol$_c$/kg, K is potassium cmol$_c$/kg, Cu is copper mg/kg, Zn is zinc mg/kg, Fe is iron mg/kg, B is borum mg/kg, S is sulfur mg/kg, Mn is manganes mg/kg, cations exchange capacity is the sum of exchangeable bases: Ca^{2+}+Mg^{2+}+K, in cmol$_c$/kg, Positive exchangeable charges are the total exchangeable cation charge: V%+CTC, in cmol$_c$/kg, Total count is the radiometric measure (cps), K % is the potassium channel in parts per million, U ppm is de uranium channel in parts per million, Th is the thorium channel in parts per million, elevation meters, slope degrees, flow accumulation is an integer, add up of the hydrological flux, horizontal curvature °/m, vertical curvature °/m, and aspect is slope orientation cos(°).

4.3. K-means clustering and scatterplot analysis

The predominant Paraná clusters one, two, and three (Fig. 7) relate to the respective hillslope regions, concave (yellow), rectilinear (green) and convex (blue). These clusters characterize relations among the relief, soil geochemistry, regional geology, and geologic structure. For example, the crystalline shield in the Ribeira Valley (Dardenne and Schobbenhaus, 2001) and Caiuá Formation (Fernandes and Coimbra, 1994) correspond to the low lying blue cluster (Table 2), whereas the yellow cluster corresponds to the SW-NE trends (Fig. 7c) of the Foz do Iguaçu graben and Cruzeiro do Oeste horst (Milani, 1997) of intermediate elevations. When plotting the Ribeira Valley region against the X-coordinate locations, the highest soil sulfur content appears nearest to the coastline. One interpretation is that the sulfur content at these locations is related to marine aerosols (Claypool et al., 1980). In this case, aerosols carried inland from the ocean are blocked by the Serra do Mar topographic high and precipitation in this region provides the sulfur that accumulates in soils near the coast.

Table 2 – Descriptive statistics for elevation (m)

	Cluster 1 - Yellow	Cluster 2 - Blue	Cluster 3 - Green
Average	504	504	817
Minimum	213	9	208
1st Quartile	388	384	758
3rd Quartile	500	511	831
Maximum	1058	1097	1288

The spatial distribution of these clusters appears related to the macrobasin (>35.000 km^2) structure (Fig. 7d). For example, the head waters of the Tibagi, Ivaí, Piriqui and Iguaçu Rivers fall into cluster three whereas the discharge areas occupy cluster two. The western part of cluster two occupies the Caiuá formation area, composed by cretaceous sediments (light green on Fig. 2). A review of maps produced by Mineropar (2005) shows that the region is enriched only in silica oxide (over 49,41% or 233.600 ppm) not the other 72 elements analyzed for the study. Considering that well developed soils reflect more intensely the effects of the weathering defined by environmental conditions, the fact that the clusters defined from B-horizon soils maintain relations with relative positions inside these basins (upper, middle and lower course) was considered evidence of a steady-state condition; that is, uplifting is not significant over the study time scale. At a local scale (<100 km^2), the vertical curvature can be used as a surrogate to the soil production rate (Heimsath et al., 1997), because the soil production rate must be greater than the erosion rate for development of a soil layer overlying the rock matrix. The hypothesis of steady-state conditions are further supported by the presence of soils with thicknesses up to 30 m reflecting a long chemical weathering period (Mineropar, 2005; Heimsath et al., 1997).

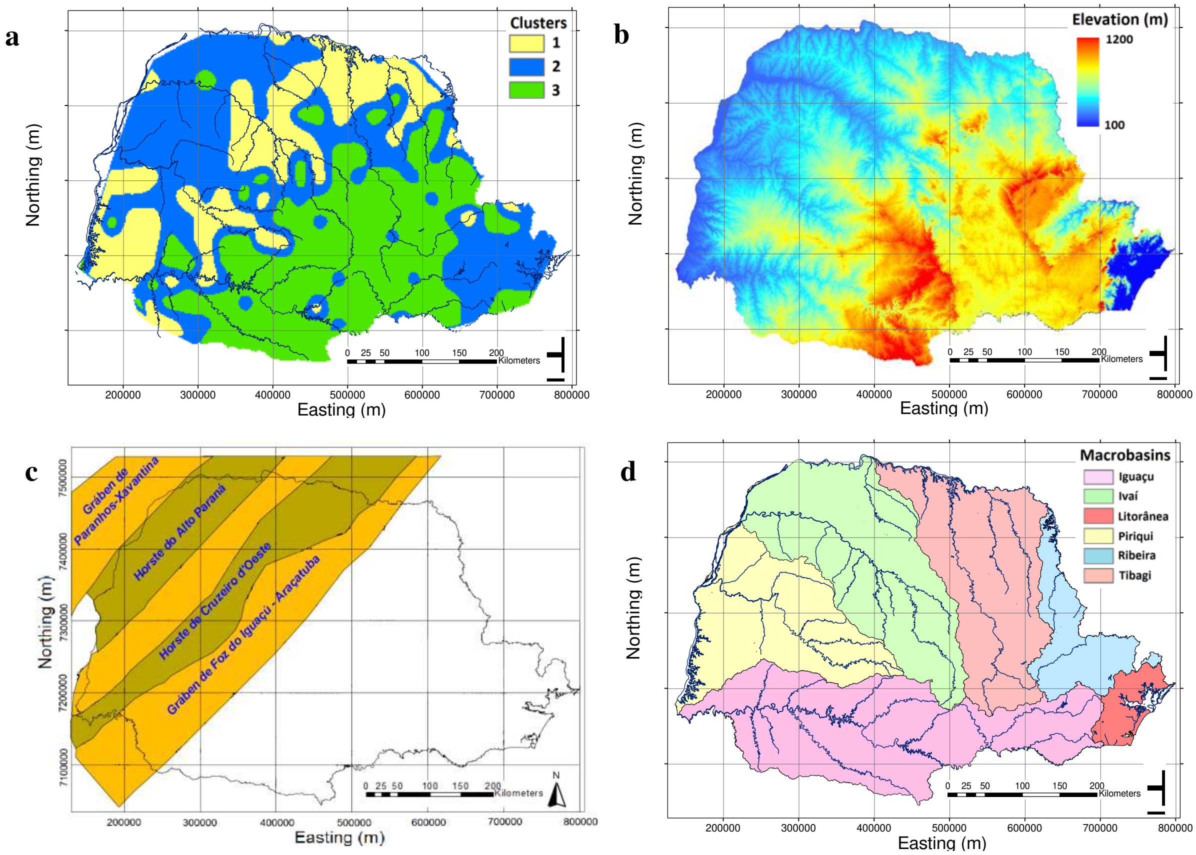


Fig. 7. Comparison between self-organizing map and relief features. a) Cluster map classified by k-means, b) elevation map, c) morph-structural map (modified from Licht, 2001), and d) macrobasins map.

In 1937, Polinov published the book ‘The Cycle Weathering’, where he proposed a decreasing sequence for chemical weathering susceptibility based on elements solubility, where the least soluble elements at far right require larger energy for their formation. The Polinov sequence was re-assessed by Hudson (1995), where the sodium was considered more soluble than the calcium.

$$Cl>SO_4>Na>Ca>Mg>K>Si>Fe>Al$$

Based on the findings of Yoo et al. (2007), the results are evaluated with respect to the representative elements [Ca], [Al] and [Fe], while representing relief morphometry, elevation was used to analyze the weathering at regional scale, whereas vertical curvature was chosen to analyze the weathering locally, given that the concave portion of the hillslope is a depositional environment, while convex areas are more susceptible to denudation (Stallard, 1988). Fig. 8 shows the scatterplots for calcium, iron and aluminum as function of elevation and vertical curvature, where the ‘samples’ are actually the nodes, results from SOM classification, each node is associated to one or more samples according to a topological similarity. The graphs can be interpreted considering the general dispersion of each element, and the pattern of each cluster individually.

The element calcium showed high concentrations in lower elevation areas and low concentrations in regions with higher altitudes (Fig. 8a), where the cluster one (yellow) relates to basalt presence (Fig. 2 and 7a), thus could indicate less weathered and therefore, less depleted feldspathic rock. Another suggestion of how weathering has been developed in the state, is the calcium concentrations related to the vertical curvature (Fig. 8d), we verified that in convex regions, usually near the divide on the hillslope, the amount of calcium is lower, whereas in concave areas there are high concentrations, additionally the samples came from a very diverse geological framework, which reinforce the chemical weathering role in these results.

Iron concentration, when analyzing cluster one (yellow), is attributed to the presence of sesquioxides, less weathered material, because according to the iron-vertical curvature plots (Fig. 8 b and e), for this cluster, iron has higher concentrations at lower elevations and on concave areas. Structural lows, as graben or regions near streams, are less subjected to physical weathering than convex areas, the very relief works as a barrier preventing the runoff favoring a relatively reducing environment (Daniels and Hammer, 1992). The iron content in cluster three is likely connected with the oxidized iron form, since it is considered less mobile and mostly present in the convex regions of the hillslope. The controlling factors of iron variation in soils and sediments were investigated by Löhr et al, (2010) using self-organizing maps to reveal nonlinear correlations between potential explanatory variables and iron content, where landform rather than land use played an important role in iron spatial distribution, supporting our findings.

Silica-alumina layers are resistant to weathering, being one of the last elements released during the pedogenesis (Embrapa, 1999). The aluminum concentrations increase with elevation and decrease toward the hillslope base, whereas hilltops are characterized by intense weathering due to hydrologic processes. Considering just the cluster one (yellow), it presented lower values for aluminum, occurring in low elevations and concave areas, relating to less weathered regions, as the iron analysis showed.

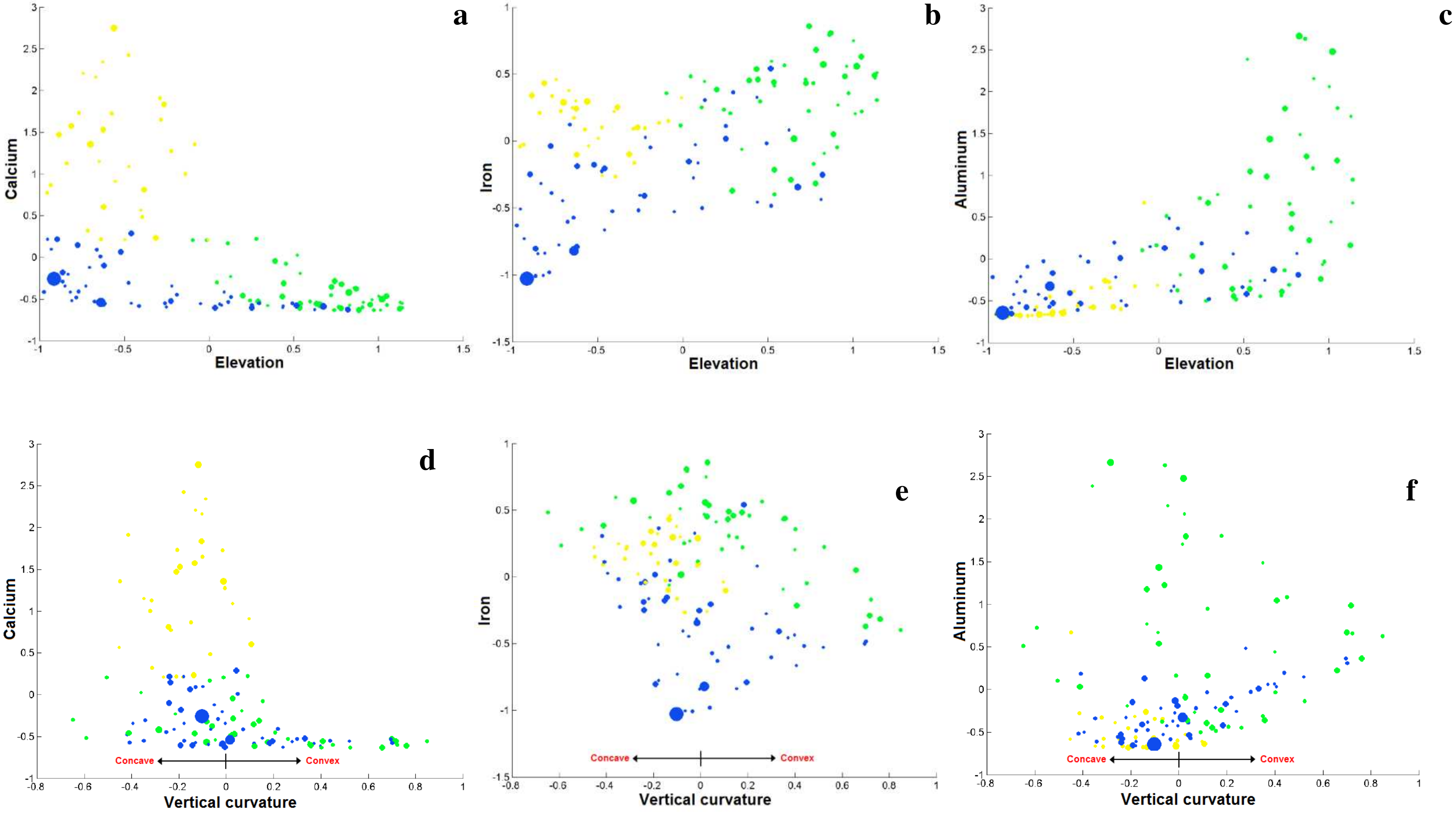


Fig. 8. Scatterplots of the major cations and oxides as a function of elevation and vertical curvature. According to the Polinov sequence (Hudson, 1995), calcium is a highly mobile therefore having higher concentrations at lower elevations that are characteristic of concave areas. By contrast, iron oxide is relatively more mobile, and aluminum were concentrated on higher elevations.

5. CONCLUSIONS

With this study, we have found that it is possible to use data mining techniques for the evaluation of multi-scale hillslope chemical weathering processes. Using a type of unsupervised artificial neural network, called the self-organizing map (SOM), multidimensional soil geochemical and geophysical variables can be projected onto a two-dimensional surface while preserving important nonlinear relations. Grouping nonlinear relations using the k-means clustering technique facilitates the development of conceptual hillslope weathering models.

At a local scale, the vertical curvature describing the convex-concave hillslope feature, relates to soil production rate and soil thickness, where SOM analysis unveiled the relations, in terms of topological array, between soil geochemistry and terrain morphometry, indicating that topographic attributes are associated with chemical denudation. The elevation depicts the weathering phenomenon at regional scale, when identifying different levels of chemical denudation along the macrobasins. The headstream showed evidence of being more weathered than the middle sections, additionally, the Caiuá formation, composed of cretaceous sediments represents the sediment transport from the macrobasins upper sections. This analysis is supported by the fact that this process has been carried out since the cretaceous, most part inside a craton, favoring our steady-state condition assumption.

Chemical weathering is an important factor for development of the terrain morphology in the state of Paraná. Chemical element concentrations depend on the hillslope morphology that constitutes a two-way process: hillslope profiles influence the weathering, and weathering influences hillslope morphology. The soil chemical composition is a result of a large number of factors including the bedrock-to-soil conversion rate, soil erosion (mass transport), and solute transport. The SOM and k-means methods made it possible to understand the nonlinear relationships associated with a large number of variables. This two-step approach can be used to support further studies to understand hillslope chemical weathering, erosion, and landscape evolution in other locations and environmental settings.

Acknowledgements

We are grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and Conselho Nacional de Desenvolvimento Científico e Tecnológico for their financial support; to Victor F. Labson, Director, Crustal Geophysics and Geochemistry Science Center (CGGSC), U.S. Geological Survey (USGS), Denver, Colorado, United States of America for providing the first author with the position of visiting scientist; to the Paraná State Geological Survey (Mineropar) represented by Otávio Licht, who kindly provided us with the geochemical dataset; to George Breit, David Smith and Philip J. Brown II, CGGSC, USGS, for their comments and suggestions regarding this study.

REFERENCES

Astel, A., Tsakoviski, S., Barbieri, P., Simeinov, V., 2007. Comparison of self- organizing maps classification approach with cluster and principal components analysis for large environmental datasets. Water Research 41, 4566 – 4578.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000. Artificial neural networks in hydrology. I: preliminary concepts. Journal of Hydrologic Engineering 5 (2), 115-123.

Claypool, G. E., Holser, W. T., Kaplan, I. R., Sakai, H., Zak, I., 1980. The age curves of sulfur and oxygen isotopes in marine sulfur and their mutual interpretation. Chemical Geology 28, 199-260.

Dardenne, M. A., Schobbenhaus, C. S., 2001. Metalogênese do Brasil. Editora Univesidade de Brasília. Brasília.

Daniels, R. B., Hammer, R. D., 1992. Soil Geomorphology, John Wiley & Sons Inc., Toronto.

Ehsani, A. H., Quiel, F., 2008. Geomorphometric feature analysis using morphometric parametrization and artificial neural networks. Geomorphology 99, 1-12.

Empresa Brasileira de Pesquisa Agropecuária–Embrapa, 1999. Sistema Brasileiro de Classificação de Solos, Embrapa Produção de Informação, Rio de Janeiro.

Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. American Geophysical Union EOS 81, 583–585.

Fernandes, L. A. Coimbra, A. M., 1994. O Grupo Caiuá (Ks): revisão estratigráfica e contexto deposicional. Revista Brasileira de Geociências 24, 164-176.

Forgy, E., 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications, Biometrics 21, 768-781.

Friedel, M.J., Souza, O.F., Yoshinaga, S. P., Silva, A. M., *in review*, Predicting well yield in northeastern Brazil from hydrogeologic and airborne geophysical measurements using self organizing maps, genetic programming, and uncertainty analysis, Journal of Hydrology, 25 p.

Gerrard, J., 1992. Soil geomorphology: an integration of pedology and geomorphology, Chapman & Hall, London.

Hanesch, M., Scholger, R., Dekkers, M. J., 2001. The application of fuzzy C-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites. Physical Chemistry Earth (A) 26, 885-891.

Hentati, A., Kawamura, A., Amaguchi, H., Iseri, Y., 2010. Evaluation of sedimentation vulnerability at small hillslide reservoir in the semi-arid region of Tunisia using Self-Organizing Map. Geomorphology 122, 56-64.

Heimsath, A. M., Dietrich, W. E., Nishizumi, K., Finkel, R. C., 1997. The soil production function and the landscape equilibrium. Nature 388, 358-361.

Hudson, B. D., 1995. Reassessment of Polynov′ s Ion Mobility Series. Soil Science Society American Journal 59, 1101-1103.

Instituto Brasileiro de Geografia e Estatística – IBGE, 2010. Banco de dados por Estados. http://www.ibge.gov.br/estadosat/perfil.php?sigla=pr. Acessed in Aug-31-2010.

Jenson S. K. and J. O. Domingue., 1988. Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis. Photogrammetric Engineering and Remote Sensing 54, 1593–1600.

Kalteth, A. M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modeling and application. Environmental Modeling and Software 23, 835-845.

Kohonen, T., 2001. Self-organizing Maps, third edition, Springer-Verlag, Berlin.

Licht, O. A. B. 2001. A geoquímica multielementar na gestão ambiental. PhD Thesis, Faculdade de Geologia, Universidade Federal do Paraná, Brazil.

Löhr, S. C., Grigorescu, M., Hodgkinson, J. H., Cox, M. E., Fraser, S. J. 2010. Iron Occurrence in soils and sediments of coastal catchment: A multivariate approach using self-organizing maps. Geoderma 156, 253-266.

Milani, E. J., 1997. Evolução tectono-estratigráfica da bacia do Paraná e seu relacionamento com a geodinâmica fanerozóica do Gondwana sul-ocidental. PhD Thesis, Instituto de Geociências, Universidade Federal do Rio Grande do Sul, Brazil.

Minerais do Paraná S.A.- MINEROPAR, 2005. Geoquímica de solo – Horizonte B: Relatório Final de projeto, Curitiba.

Minerais do Paraná S.A.- MINEROPAR, 1986. Mapa geológico do Estado do Paraná, MINEROPAR, Curitiba . Map, 60 x 80 cm. Scale 1:1.400.000.

Mudd, S. M., Furbish, D. J. 2004. Influence of chemical denudation on hillslope morphology. Journal of Geophysical Research 109, F02001.

Netter, J., Kutner, M. N., Nachtssheim, C. J., Wasserman, W. 1996. Applied linear statistical models, 4th Ed. WCB/McGraw-Hill, Boston.

Park, S. J., van de Giesen, N. 2004. Soil-landscape delineation to define spatial sampling domains for hillslope hydrology. Journal of Hydrology 295, 28-46.

Penn, Br. S. 2005. Using self-organizing maps to visualize high-dimensional data. Computer & Geosciences 31, 531-544.

Rao, C. R., Toutenburg, H., Shalabh, Heumann, C., 2008. Linear models and generalizations: least squares and alternatives, third edition, Springer-Verlag, Berlin.

Reimann, C., Filzmoser, P., 1999. Normal and lognormal data distribution in geochemistry: death of a myth. Consequeces for statistical treatment of geochemical and environmental data. Enviromental Geology 39, 1001-1014.

Reimann, C., Filzmoser, P., Garret, R. G., 2002. Factor analysis applied to regional geochemical data: problems and possibilities. Applied Geochemistry 17, 185-206.

Rennó, C. D., 2003. Construção de um sistema de análise e simulação hidrológica: aplicação a bacias hidrográficas. PhD Thesis, Remote Sensing Division – Instituto Nacional de Pesquisas Espaciais, Brazil.

Roering, J. J., Kirchner, J. W., Dietrich, W., 1999. Evidence for nonlinear, diffusive sediment transport on hillslopes and implications for landscape morphology. Water Resources Research 35, 853-870.

Stallard. R. F., 1988. Weathering and erosion in the humid tropics. In Lerman, A., Meybeck, M., North Atlantic Treaty Organization-NATO. Physical and chemical weathering in geochemical cycles, Kluwer Academic Publishers, Dorddrecht. pp. 225-246.

Stewart, M. A., Jardine, P. M., Barnett, M. O., Mehlhorn T. L., Hyder, L. K., McKay, L. D., 2003. Influence of soil geochemical an physical properties on the sorption and bioaccessibility of chromium (III). Journal of Environmental Quality 32, 129-137.

Ultsch, A., 2003. U-Matrix: a tool to visualize clusters in high dimentional data. University of Marburg, Department of Computer Science, Technical Report 36, 1-12.

Valeriano, M. M., Rosetti, D. F., Albuquerque, P. C. G., 2009. TOPODATA: desenvolvimento da primeira versão do banco de dados geomorfométricos locais em cobertura nacional. Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, INPE, Natal, Brasil, pp. 5499-5506.

Valeriano, M. M., Kuplich, T. M., Storino, M., Amaral, B., Mendes, J. N., Lima, D. J., 2006. Modeling small watersheds in Brazilian Amazonia with shuttle radar topographic mission-90 m data. Computer & Geosciences 32, 1169-1181.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 2000. SOM Toolbox for Matlab 5. SOM toolbox team, Finland. Helsinki University of Technology, Laboratory of Computer and Information Science. http://www.cis.hut.fi/projects/somtoolbox/

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Neworks 11, 586-600.

Yoo, K., Amundson, R., Heimsath, A. M. Dietrich, W. E., Brimhall, G. H., 2007. Integration of geochemical mass balance with sediment transport to calculate rates of soil loss chemical weathering and transport on hillslopes. Journal of Geophysical Research 112, F02013.

INTELLIGENT ESTIMATION OF SPATIAL DISTRIBUTED SOIL PHYSICAL PROPERTIES

Abstract

The self-organizing map (SOM) technique is used to predict soil texture and hydraulic conductivity based on relief morphometric features. The concave-convex nature of hillslopes (from hilltop to bottom of the valley) reflects a steady-state geomorphic condition. The topographic features are extracted from Shuttle Radar Topographic Mission (SRTM) elevation data; whereas soil textural (clay, silt, and sand) and hydraulic conductivity data are associated with 30 random locations (75 cm depth). In contrast to traditional principal component analysis, the SOM identifies relations among relief features, such as, slope, horizontal curvature and vertical curvature. Stochastic cross-validation indicates that the SOM is unbiased and provides a way to measure the magnitude of prediction uncertainty for all variables. The SOM cross-component plots of the soil texture reveals higher clay proportions at concave areas with convergent hydrological flux and lower proportions for convex areas with divergent flux. The sand ratio has an opposite pattern with higher values near the ridge and lower values near the valley. Silt has a trend similar to sand, although less pronounced. The relation between soil texture and concave-convex hillslope features reveals that subsurface weathering and transport is an important process that changed from loss-to-gain at the rectilinear hillslope point. These results illustrate that the SOM can be used to capture and predict nonlinear hillslope relations among relief, soil texture, and hydraulic conductivity data.

Keywords: self-organizing maps, hydraulic conductivity, soil texture, Monte-Carlo simulation, Poços de Caldas, Brazil.

1. INTRODUCTION

Knowledge of soil texture and hydraulic conductivity is important for evaluating physical and chemical processes, such as weathering, erosion, runoff, and groundwater recharge (Daniels and Hammer, 1992). Depending on the hillslope processes, subsurface material of the same mineral composition can develop into soils with different characteristics (Hugget, 1998). According to Young (1980), the primary factors influencing soil formation are climate, parent material, relief (surface shape), hydrology, organisms, time, and human activities. Because these phenomena are coupled, nonlinear, and scale dependent (ASCE, 2000b), the application of traditional methods for their estimation is challenging.

To estimate soil weathering processes, it is necessary to understand the relations among soil, relief, and hydrology (Nezat et al., 2004). Traditional studies of weathering are conducted at the hillslope (or basin) scale, where it is possible to measure sediment mass loss and transport for use in the construction of empirical or numerical models (Mudd and Furbish, 2004). Financial and time restrictions typically limit sampling at both temporal and spatial scales. The limited data availability and high variability promote increasing amounts of uncertainty in model predictions (Hornberger et al., 1998).

To overcome these limitations, remote sensing images and geographic information system (GIS) models are often used to characterize the surface and estimate field parameters for larger areas (Jensen, 2007). For example, relief morphometric features can be extracted from digital elevation models using a mobile window for gradient calculation. This facilitates the extraction of measures, such as slope, aspect, horizontal curvature and vertical curvature (Valeriano, 2008). Mudd and Furbish (2004) and Yoo et al., (2007) incorporated relief morphometric features into the hillslope physical and chemical weathering models in an effort to improve early models that consider the hillslope to be a uniform rectilinear entity.

The hillslope heterogeneity affects soil weathering characteristics. For example, the concave areas on hillslopes are frequently associated with convergent hydrological fluxes, and therefore have higher average soil moisture than convex areas. Convex regions (characterizing hilltops and ridges) are more susceptible to, erosion and mass removal, especially clays and silt more subject to transport, while concave areas constitute a more depositional environment (Heimsath et al., 1997). Therefore, it is possible to establish a relation between hillslope morphometry and soil texture, although the interactions between them often are non linear and nonunique.

Some challenges in the construction and application of numerical hillslope models are their one-dimensionality, steady-state requirements, lack of calibration data, and nonuniqueness (Loke and Barker, 1996). Also, numerical models commonly are too rigid with respect to detecting unexpected features like the onset of trends, non-linear relations, or patterns restricted to sub-samples of a data set. These shortcomings create the need for an alternate modeling approach capable of using available data. One technique that is well-suited to noisy, sparse, nonlinear, multidimensional, and scale-dependent data is a type of unsupervised artificial neural network called the self-organizing map (Kohonen, 2001). The self-organizing map (SOM) technique is used in related studies to explore relations among rock geochemistry and hyper-spectral images (Penn, 2005), classify geomorphometric aspect based on digital elevation models (Ehsani and Quiel, 2008), characterize hillslope landslide vulnerability (Hentati et al., 2010), and identify processes controlling the distribution of iron in soil and sediment (Löhr et al., 2010).

The aim of this study is to understand and predict scale-dependent relations among soil physical properties and morphometric features across Poços de Caldas municipality, southeastern Brazil. The hypothesis is that there is a relation between hillslope morphometric features and soil texture and hydraulic conductivity near the surface, additionally we assume that is possible to devise these relations based on the statistical relations among field data and GIS metrics using the self-organizing maps. To achieve the goal and satisfy this hypothesis, the following objectives are undertaken: : (1) analyze nonlinear relations among published hydraulic conductivity, soil texture and relief morphometry data from 29 locations using the SOM (Kohonen, 2001) and component planes visualization (Penn, 2005) techniques; (2) generate random correlated values for soil and relief properties using Monte Carlo stochastic simulation method (Fishman, 1996); (3) generate predicting maps of clay, silt and sand content, and hydraulic conductivity on saturated soil; (4) identify conceptual models of soil physical weathering processes based on k-means clustering (Forgy, 1965) of the SOM topography for development of predictive (empirical and numerical) models; and (5) evaluate bias and uncertainty in the quantized vector predictions using a stochastic cross-validation technique (Rao et al, 2008).

2. SETTING

The study is conducted at the Vargem de Caldas basin (42 km^2) near Poços de Caldas, Minas Gerais, Brazil (Fig. 1). The Poços de Caldas plateau is considered the largest alkaline complex in Brazil (Holmes et at., 1992) with one of the largest uranium occurrences in the world (Chapman et al., 1992). In addition to uranium, there also are important bauxite deposits and sulfuric thermal springs (Fernandes and Franklin, 2001). The climate of this area is characterized by dry winters and mild summers with precipitation of 1300 mm to 1700 mm and average temperatures of 18ºC to 22ºC (Christofoletti, 1970).

The Poços de Caldas plateau is largely comprised of Precambrian rocks that are underlain by Archean basement of rocks from the Varginha Complex (gneisses, migmatites, granulites) and by nephelines syenite (tinguaite, phonolite phoiaite) of Mezosoic-Cenozoic age. The alkaline complex had its first manifestation on superior Cretaceous (87 ma) and evolved until 60 ma (Christofoletti, 1970). The plateau has a dome shape with ridges and scarps at the outer edges facilitated by the central chimney intrusion. Collapse of the central part of the structure resulted in forming radial and ring fractures where magma ascended to the surface (Holmes et al., 1992).

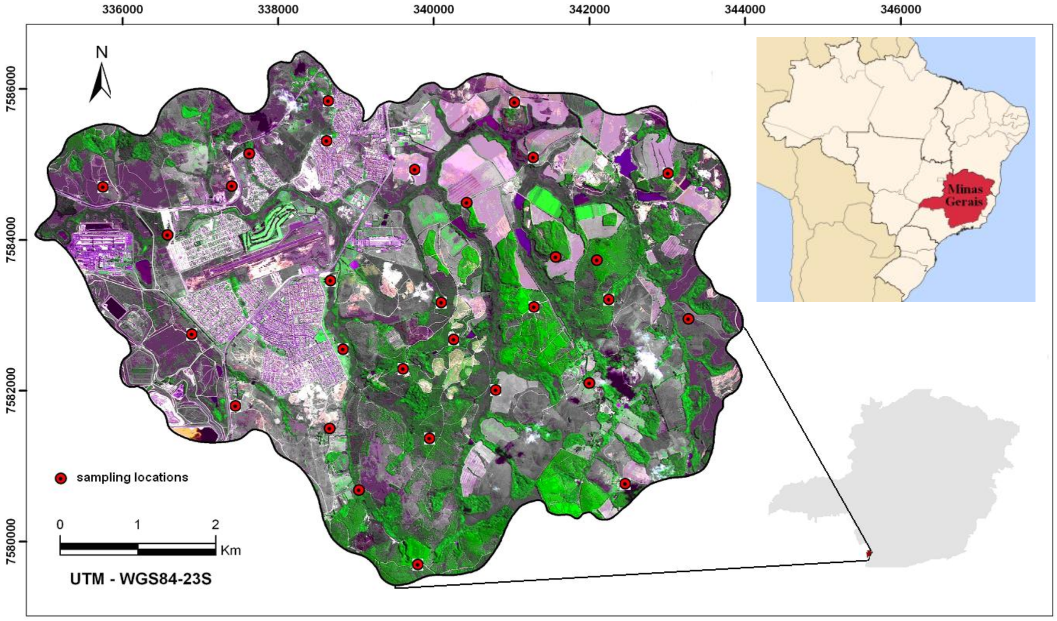


Fig. 1 – Vargem de Caldas basin, Minas Gerais, Brazil. (a) False-color composition of bands 321 (RGB) of the SPOT-4 HRVIR sensor. Permeability tests and soil samples were collected in 29 sites across the basin (red dots) (after Ribeiro, 2010)

The same mineralogy exists throughout the complex, but regional differences in weathering process resulted in soils with distinct characteristics. The soils in this study area are Regolithic Neosols and Haplic Cambisols (Embrapa, 1999). According to Moraes (2008), the Neosols are associated with colluvium and paleo-floodplains underlain by a syenite basement, being characterized by high clay proportion and terrain with weak to moderate dissection (related to drainage density and carving intensity). The Cambisols are typically present on plateaus (Fig. 2) with medium to strong dissection, defined by a higher erosive potential and a clay matrix with gravel and concretions of laterite and alkaline rocks.

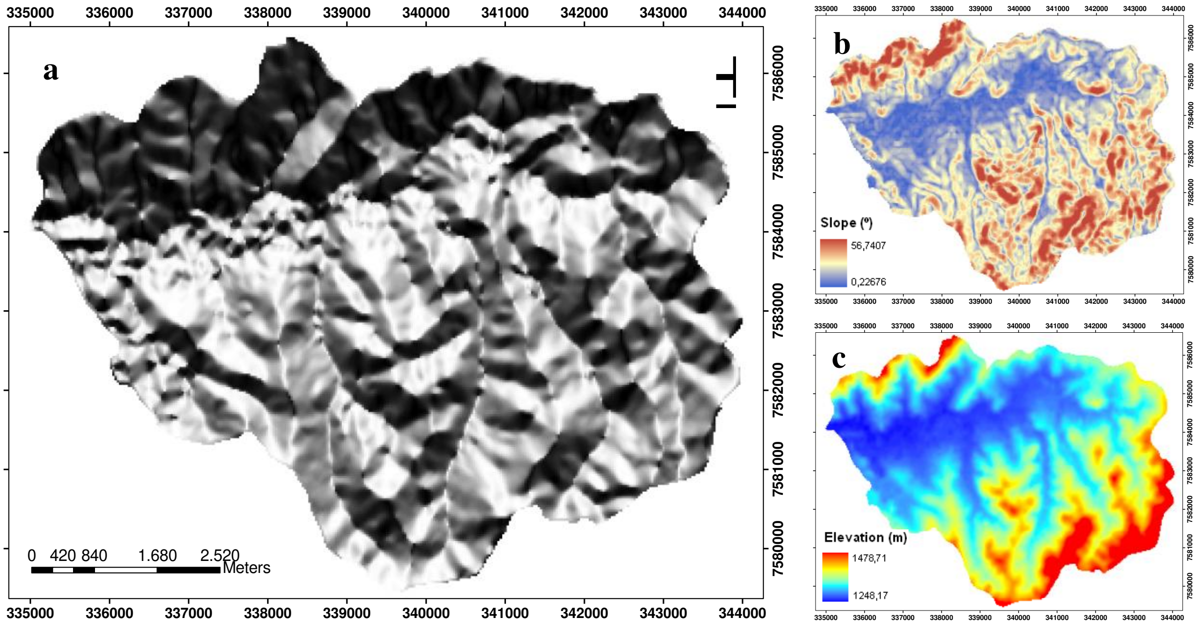


Fig. 2 – Maps characterizing the study area. a) Shaded relief map, also represents slope orientation, or aspect, clearer tones indicate north and darker, south; b) Terrain slope (°); and c) Elevation (m).

Twenty nine soil samples were collected (Ribeiro, 2010) at 75 cm depth and in random locations across Poços de Caldas region (Fig. 1). The purpose of sampling is to provide physical soil texture and hydraulic conductivity values for input to the SOM (Table 1). The SOM variable names are introduced in parentheses and summarized in Table 2. The soil texture variables, such as clay, sand, silt for each sample is determined using the method of Camargo et al. (1986) at the Soil Laboratory of the Faculdade de Engenharia Agrícola, Universidade de Campinas (FEAGRI-

UNICAMP). Soil conductivity (Conductivity) values are determined based on hydraulic field experiments and the equation proposed by Reynolds and Elrick (1989):

$$K_{fs} = \frac{C.Q}{2.\pi.H^2 + C.\pi.a^2 + \left(\frac{2.\pi.H}{\alpha *}\right)}$$

where: a = borehole radius (cm); C = shape factor, which depends on the soil type and the H/a ratio; H = hydraulic charge height on the borehole (cm); K_{fs} = hydraulic conductivity on saturated field (cm/s); Q = water yield in the soil (cm^3/s); and α = parameter related to the porous media size.

Table 1 – Descriptive statistics of collected data (Ribeiro, 2010)

	Conductivity (cm/s*10^{-5})	Sand (%)	Silt (%)	Clay (%)
Mean	0,6068	6,6	29,4	59,35
Standard deviation	1,8516	7,42	11,82	13,55
Minimum	0,0665	0	15,5	39,45
Maximum	8,5385	25,7	54,75	84,5

Table 2 – Variable notation and data descriptions used in this study.

Category	Acronym	Description
Soil	Clay	Percent clay
	Sand	Percent sand
	Silt	Percent silt
	Conductivity	Hydraulic conductivity, cm/hr
Morphometric	cosASPECT	Cosine of slope orientation
	ELEVATION	Elevation, meters
	HOR-CURV	Horizontal Curvature, °/m
	SLOPE	Slope inclination in degrees
	VERT-CURV	Vertical Curvature, °/m

Each sample location is associated with topographic morphometry data (e.g. elevation, slope, slope orientation, vertical curvature and horizontal curvature). Characterization of the topographic relief is possible using elevation data provided by the Shuttle Radar Topographic Mission (SRTM) (Farr and Kobrick, 2000). The digital elevation model associated with these data is provided by the United States Geological Survey on a lattice with 90-m spatial resolution. The Brazilian Topodata Project (Valeriano et al., 2009) used a geographical information system and related modeling tools to derive a set of geomorphometric features (30-m resolution) from the elevation data. The geomorphometric features provide suitable information to extract morphometric features, such as slope, aspect (hillslope orientation), vertical and horizontal curvature (Valeriano et al., 2006). These SRTM measures are employed by Grohmann et al. (2007) to analyze the Poços de Caldas morphotectonic and geomorphology based on drainages, 3D visualization and morphometric parameters, supporting use of these variables to investigate soil physical attributes.

The slope variable (SLOPE) represents the first derivative of two locations on the elevation data, whereas the second derivative produces the aspect variable which indicates the position of the hillslope relative to north. The aspect varies from 0 to 360º with value zero pointing towards north. Since both zero and 360º represent the north, a trigonometric (cosine) transformation (cosASPECT) was applied so that these values varied from -1 (south) to 1 (north). Other derived measures are the vertical curvature (VERT-CURV) and horizontal curvature (HORZ-CURV). The vertical curvature depicts the hillslope profile: convex, rectilinear, and concave shape, whereas the horizontal curvature depicts the hillslope shape. These two features are highly correlated but when analyzed in combination they produce different hillslopes shapes, which could lead to a soil with distinct physical-chemical properties.

3. METHODS

3. 1 The Self organizing map technique

In this paper we present a method to estimate soil hydraulic conductivity and soil texture near the surface combining SOM and Monte Carlo stochastic simulation (Fig. 3). The following steps are applied (Friedel and Iwashita, *in press*):

(1) All variables values were standardized so that no variable would dominate in the nonlinear modeling process (Kalteth et al., 2008). The z-score transformation is given by:

$$z_i = \frac{x_i - \bar{x}_i}{s_i}$$

where *z* is the standardized value; *x* is the raw score; $\bar{x}$ is the sample average, and *s* is the sample standard deviation, *i* is an index for each variable. Standardizing variables in this way resulted in each having an expected value of zero and standard deviation one.

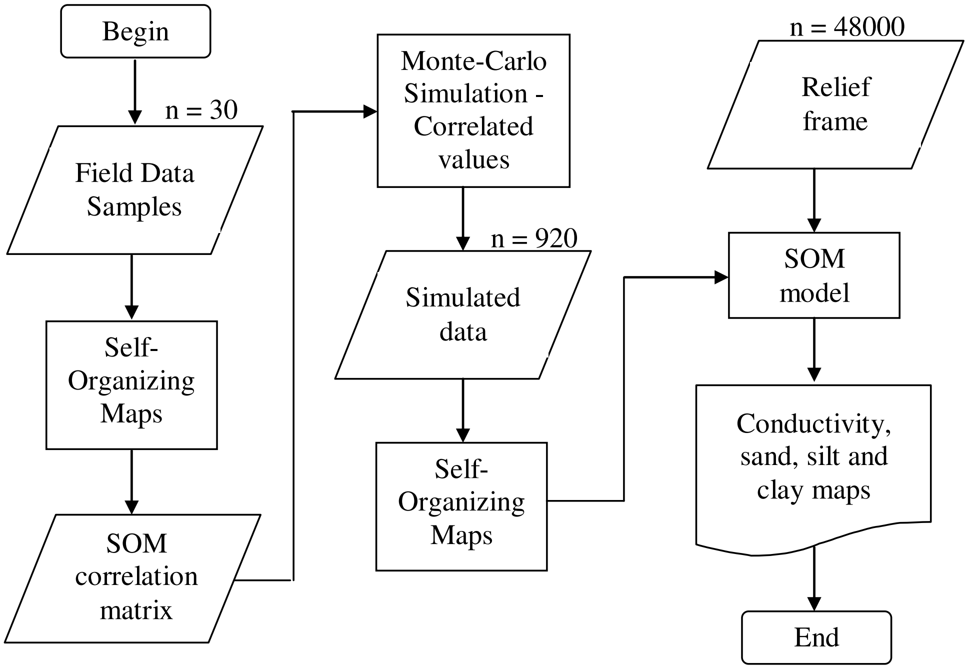


Fig. 3 – The flow chart represents the proposed method to analyze and predict soil physical properties combining data mining and stochastic simulation techniques.

(2) The SOM (Kohonen, 2001) is used to calculate the correlation coefficient between the variables. The Spearman index, a non parametric measure, are applied to the quantized vectors after the topologic rearrangement determined by the SOM (Table 3).

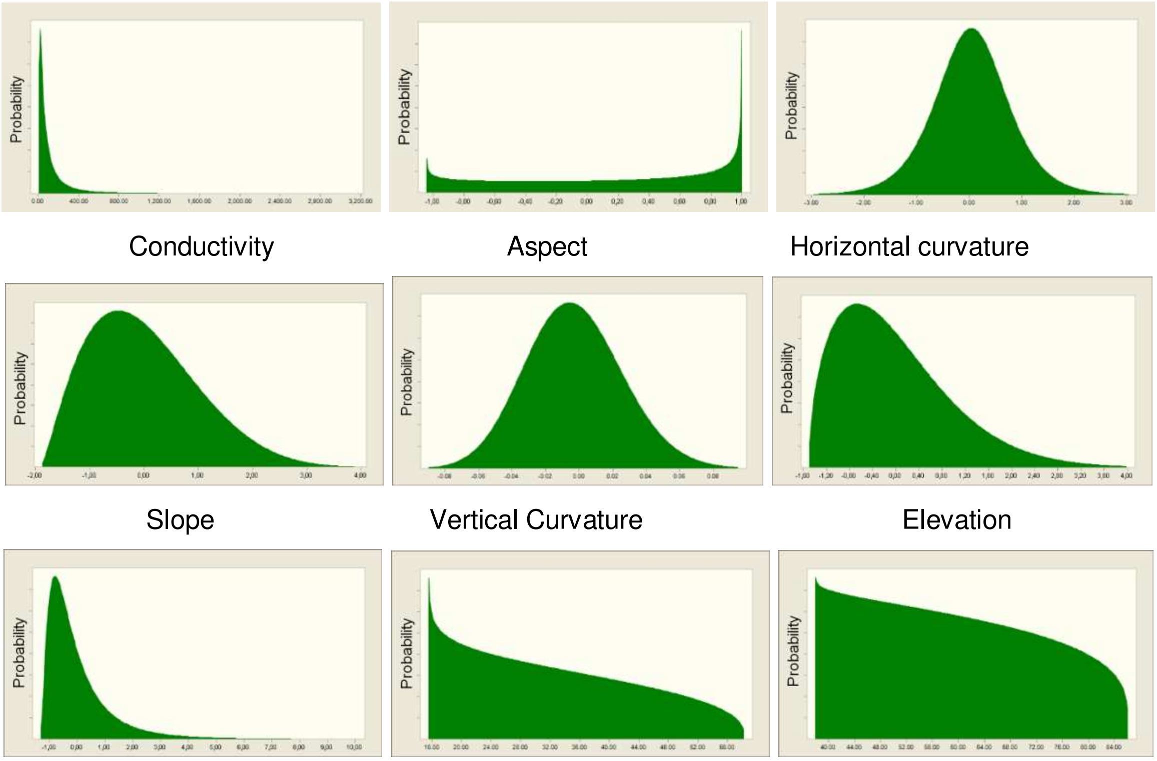
(3) For each variable a continuous probability distribution is adjusted using Kolmogorov–Smirnov (Netter et al., 1996) test with p-value = 0,05 (Fig. 4).

(4) Random correlated values are generated using Monte Carlo approach (Fishman, 1996). The stochastic simulation is conducted keeping the correlations coefficients found on step 2, that is, the correlation matrix is used as a constraint.

(5) The resulting values are used to build another SOM model. This model is applied to a continuous surface containing all the topographical features, which are used as explanatory variables to produce the conductivity and soil texture maps, the dependent variables.

Table 3 – Spearman correlation matrix after the topological reorganization.

	Conductivity	Aspect	Horizontal Curvature	Slope	Vertical Curvature	Elevation	Sand	Silt	Clay
Conductivity	1								
Aspect	-0,33	1							
Horizontal Curvature	0,12	0,42	1						
Slope	0,00	-0,81	-0,56	1					
Vertical Curvature	0,38	0,15	0,86	-0,48	1				
Elevation	0,66	-0,45	0,38	0,17	0,68	1			
Sand	0,46	0,33	0,82	-0,47	0,76	0,47	1		
Silt	0,61	-0,65	0,20	0,25	0,58	0,91	0,25	1	
Clay	-0,69	0,43	-0,47	-0,04	-0,77	-0,94	-0,58	-0,93	1



Conductivity

Aspect

Horizontal curvature

Slope

Vertical Curvature

Elevation

Sand

Silt

Clay

Fig. 4 – Distribution probability functions fitted for each variable using Kolmogorov-Smirnov test (p-value = 0,05). The functions were: conductivity, lognormal; aspect, beta; horizontal curvature, t-student; slope, beta; vertical curvature, logistic; elevation, weibull; sand, lognormal; silt, beta; clay, beta.

(6) Finally, the k-means clustering technique (Forgy, 1965) was used to classify the SOM topography into statistically relevant conceptual models (Ehsani and Quiel, 2008), where soil physical properties are interpreted based on terrain morphometry and associated clusters.

The self-organizing map (SOM) belongs to a subcategory of the artificial neural network algorithms, called competitive learning networks, in which the computational models serve as a proxy for neurons in the human brain (ASCE, 2000a). The term self-organizing is based on the unsupervised nature of the algorithm having the ability to organize information without any a priori specification of an output pattern. In this study, the output consists of neurons organized on a two-dimensional rectangular grid having hexagonal cells (map). Each neuron in the map is represented by a multi-dimensional weight vector. Each neuron is connected to the adjacent neuron through a functional neighborhood relation (Vesanto et al., 2000). Individual data samples are associated to a vector with properties that reflect its contributions relative to the other variables. From this cloud of data vectors, a best matching unit (BMU) is iteratively determined by minimizing a distance measure for each variable (Kohonen, 2001; Vessanto et al, 2000). The topology of the vectors is altered until convergence conditions are reached. The resulting maps are organized in such way that similar data are mapped to the same or nearby nodes, and dissimilar data are mapped to nodes with greater separation distances. The equations and issues regarding data gathering, normalization, and training are well-documented in Vesanto (1999), Vesanto and Alhoniemi (2000), and Kohonen (2001); therefore, no other details are given here.

3.1.1 Component planes

The component planes visualization technique (Vesanto, 1999) is used to analyze relations among the landscape variables. Whereas a component plane can be thought of as a slice of the SOM, it actually represents one set of vector component (variable) values in all map units; that is, each component plane portrays the spread of values for the associated variable. In that regard they are similar to histograms, the difference being that the same value can be present in multiple places of the map when it belongs to different clusters.

The visualization of multiple component planes allows identification of correlated variables (Vesanto, 1999). Correlations appear as similar color patterns at the same locations in differing component planes. In the case of positive correlation, the color distributions in component planes are identical among variables meaning that as one variable increases (or decreases) the others do the same. Conversely, a negative correlation among variables appears with the same pattern but opposite color distribution, meaning that as values in one variable increase those in the other variable decrease.

3.1.2 K-means clustering and scatterplots

K-means cluster analysis (Vesanto and Alhoniemi, 2000) is used to identify groupings in the SOM output neurons. As a post-processing investigation method, it has shown promising results when analyzing soil chemical weathering on hillslopes (Iwashita et al., *accepted*). This technique is considered better than hierarchical clustering because it does not depend on previously found clusters. The k–means algorithm assumes spherical clusters, and it is sensitive to the initialization process. For that reason, the algorithm is run multiple times for each k with different random initializations. The best partitioning for each number of clusters is selected based on the Euclidian distance criterion, and interesting merges are defined using the Davies–Bouldin index (Davies and Bouldin, 1977).

3.1.3 Self-organizing map estimation

The SOM estimates data values based on distances among the available vectors (Fessant and Midenet, 2002; Wang, 2003; Junninen et al., 2004; Kalteh and Berndtsson, 2007; Kalteh and Hjorth, 2009). The traditional estimation process is by replacement (called imputation), where the values are taken directly from the prototype vectors of the BMUs. Often times certain data sets will result in biased predictions (Dickson and Giblin, 2007; Malek et al., 2008) requiring a modified scheme based on bootstrapping (Breiman, 1996), ensemble average (Rallo et al., 2004), or nearest neighbor (Malek et al., 2008). In this study, we first train the SOM based on a stochastic approach (Fig. 3) and second estimate missing values in the sparse data set based on the following iterative scheme:

- An initial SOM is calculated and a first set of replacement values determined.
- SOM is recalculated and new replacements for the original missing values obtained.

- Repeat the last step until the topographic error stabilizes.

Because the replacement process for missing values is not simply replacement by a prototype vector of the BMU, it is referred to herein as estimation.

3.1.4 Self-organizing map performance

The SOM algorithm is objective, but there is subjectivity when choosing the set of data variables thought to affect prediction quality. Moreover, the data variables are spatially limited, and disparate with varying levels of uncertainty in their measurements and observations. For these reasons, the reliability of the SOM as a hillslope model is evaluated using cross-validation. The basis of cross-validation (Kohavi, 1995) is a leave-one-out strategy. This requires leaving one data value out of the training set while creating a new SOM to estimate that value based on the remaining data. Because a new SOM is created up to 30 times for each value under scrutiny, it forms the basis for the Monte Carlo framework (Rubinstein and Kroese, 2007) from which residuals are used to evaluate error statistics and model bias.

3.2 Monte Carlo Simulation

The Monte Carlo (MC) method can be defined as a stochastic simulation that generates random values providing approximate solutions for mathematical problems by performing computational sampling experiments (Fishman, 1996). One of the methods advantage is the computational efficiency for a high number of parameters such as complex analytical functions and combinatorial problems, especially relevant for the present work.

The MC simulation is based on the production of pseudo-random uniform distributed values, a basic probabilistic distribution, required to simulate all others distributions where the produced numbers must be independent, that is, the number generated in one run does not influence the value of the next one. The method is applied to calculate definite integral functions, to find numerical solutions for differential equations, for optimization problems, for uncertainty analysis and to solve inverse problems (Fishman, 1996).

The efficiency of the stochastic simulation depends on the knowledge about the problem, i.e., the prior information that constrains the simulation, thus the importance of good probability distribution fitting with reliable parameters (Krajewski et al., 1991).

4. RESULTS AND DISCUSSION

4.1 Component planes

The component planes (Fig. 5) reveal interesting aspects of the training data that include correlation, dissimilarity, and grouping. Similarity in color patterns, such as elevation and silt, indicate a strong positive correlation. In this case, high (red) ELEVATION corresponds to high amount of SILT; conversely, the low (blue) ELEVATION corresponds to low amount of SILT. The distribution of red colors into three locations indicates that they are associated with three groups of differing variables. According to Rallo et al. (2002), one of the elements necessary for accurate SOM estimation is model diversity; this includes variables with a strong negative correlation (same pattern but opposite colors, such as ELEVATION and CLAY; CLAY and SILT; and cosASPECT and SLOPE. One interpretation is that at high elevations there is low clay content, whereas at low elevations clay accumulates being manifested as high values. Similarly, as CLAY and cosASPECT increase then SILT and SLOPE decrease. Less clearly defined with more dispersion is the SAND variable. Based on a review of its component plane, the SAND is associated with four groups. In contrast, CONDUCTIVITY is associated with two groups (indicated by red color in the upper right and left corners).

Although conductivity is significantly correlated to sand, given the connection between porosity and conductivity, SOM exposed higher correlation with silt instead. The lower content of sand throughout the area could be the main responsible for the reduced influence of sand grains on the hydraulic conductivity spatial pattern. Supporting this analysis, sand content is positively correlated to vertical curvature and horizontal curvature, where higher values for these variables point to convex and divergent flux areas respectively, what usually characterized ridges and hilltops, more susceptible to erosion. Plateau areas, with higher elevations are also more susceptible to erosion and sediment transport, leading to higher values of conductivity near the surface. Slope inclination is inversely correlated to the slope orientation due to the geomorphological structure of the basin, where southward hills are steeper creating shorter drainage channels, while hillslopes close to a north orientation are smoother and associated to streams with larger length.

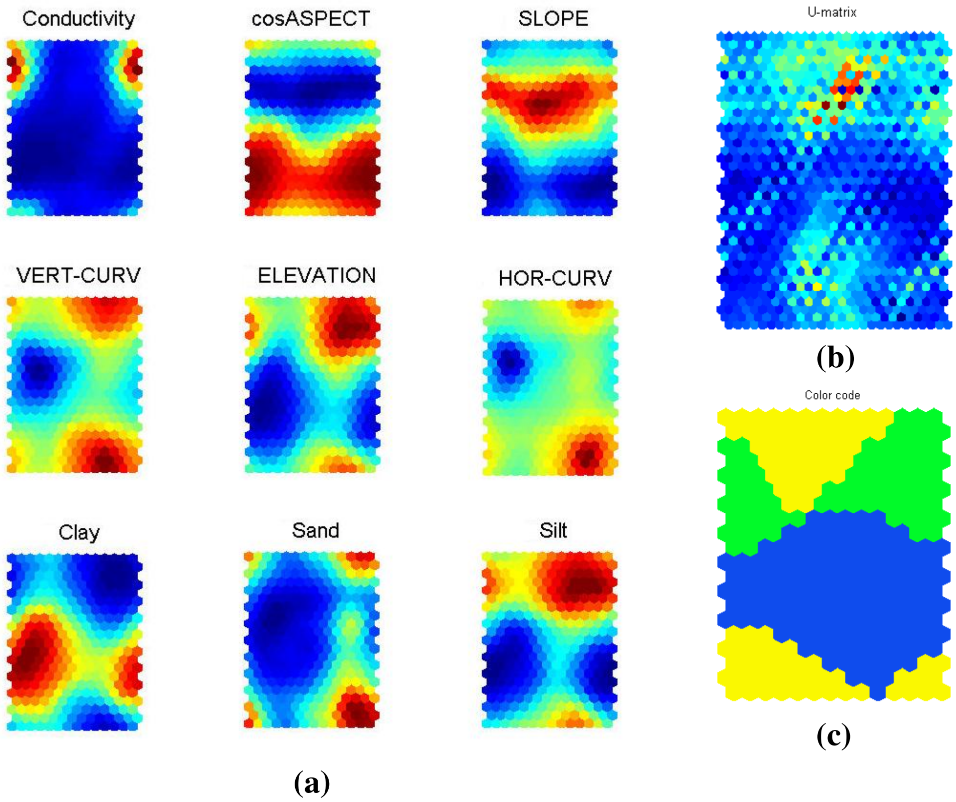


Fig. 5 – (a) Component planes used to visualize nonlinear correlation. All variables were standardized using z-score. Conductivity (m/s), sand (%), silt(%), clay (%), elevation (m), slope (°), horizontal curvature (°/m), vertical curvature (°/m), and aspect is slope orientation cos(°). (b) U-Matrix. (c) U-Matrix classified using k-means technique.

4.2 K-means clustering and scatterplot analysis

The statistical grouping of SOM nodes is done by k-means cluster analysis with natural merges identified using the Davies-Bouldin criteria. Three classes are identified based on the Davies-Bouldin criteria as being a natural number of distinct combinations for variables reflecting hillslope landscapes: convex, rectilinear and concave. A summary of median variable values (missing values are estimated beforehand using the SOM; see section 2.1.3) comprising each conceptual model is presented in Table 4. The matrix designators refer to a likelihood group in which the median values indicate their relative importance for that model; for example, low likelihood (0-33 percent), moderate likelihood (34-67 percent), or high likelihood (68-100 percent).

Regarding elevation, the relations visualized on the component maps are verified here (Fig. 6); it is positive for silt and conductivity, negative for clay, and weakly positive for sand. It is possible to associate the yellow group with low elevations, the blue with higher elevations, and the green cluster with intermediate to high elevations. Through this connection is possible to construct an integrated analysis considering the behavior of the cluster in the other graphs, like vertical curvature, for instance.

Table 4 – Conceptual models

Category	Variable	Median values for clusters			Conceptual hillslope models		
		1	2	3	1	2	3
Soil	Clay	0,79	-0,64	-0,49	High	Low	Low
	Sand	-0,60	-0,16	0,36	Low	Moderate	Moderate
	Silt	-0,71	0,65	0,32	Low	High	Moderate
	Conductivity	-0,41	-0,22	-0,32	Low	Low	Low
Morphometric	CosAspect	0,35	-0,12	0,44	Northward	Southward	Northward
	Elevation	-0,80	0,67	-0,16	Low	High	Medium
	Hor-curv	-0,21	0,02	0,60	Convergent	Parallel	Divergent
	Slope	-0,67	0,78	-0,31	Plain	Steep	Moderate
	Vert-curv	-0,39	-0,09	0,59	Concave	Rectilinear	Convex

Hillslope morphology provides an important indication of the weathering rate on the Poços de Caldas region, due to the presence of bauxites deposits, which are mostly formed on crest and interfluves, where the particle transport and drainage provoke the silica removal preventing its dissolved form to combine with alumina forming kaolinite (Holmes et al, 1992). This analysis is supported by the fact that yellow cluster occurs only on concave regions with low elevation (Fig. 6), and related to high content of clay, and low values for sand, silt and hydraulic conductivity. The blue group (Fig. 6) can be linked with the hillslope mid section and with high values of hydraulic conductivity, which can be observed when conductivity is analyzed against sand, silt and clay. The green set can be associated with hilltop areas, given the vertical curvature plots where positive values represent convex characteristics. Additionally, the green cluster is associated with larger proportions of sand, possible to visualize in all sand diagrams figures (Fig. 6).

4. 3 Conductivity and soil texture maps

The low-sand proportion predicted using the model (Fig. 7) supports the silty-clay texture typical in the Poços de Caldas region soils. The highest percentages found at 75 cm depth reached 20% of the matrix content and are more common to hilltops and divides. Sand spatial distribution express a variation at hillslope scale, given that sand highest correlations are with vertical and horizontal curvature, and that sand grains are transported for a shorter distance than silt and clay particles, since is the largest particle considered in this analysis.

The silt content is marked by strong correspondence to elevation, also revealed in the component maps analysis. Elevated silt amounts are associated to Haplic Cambisols, which according to Moraes (2008), occurs on areas of high topography, horst structures and strongly dissected regions, also associated with laterization process. The clay predicted maps are characterized by larger clay proportions on areas near rivers and drainage channels. However, according to the estimated values by the model, all the soil across the Vargem de Caldas basin is considered clayey, where the predicted percentage varies between 41 to 79%.

The conductivity map depicts low conductivity on floodplains and near drainage channels, whereas on plateau areas the values are higher. According to Jiménez-Rueda et al. (1993) low topographic areas in this region favor reducing environments leading to illite and smectite formation. Smectite is an expansive clay mineral, acting like a sponge holding the water content and decreasing the hydraulic conductivity. Predominantly occurring on the south part of the basin, the plateaus are constituted by uplifted blocks, structures intensely eroded that provide concretional material for lower elevations, characterizing the alloctonous nature of the regolith present on the soil profiles (Moraes, 2008). High erosion rates and severe weathering favor the intense sediment transport and abrupt mass movements like landslides, forming colluvial fans common in this region.

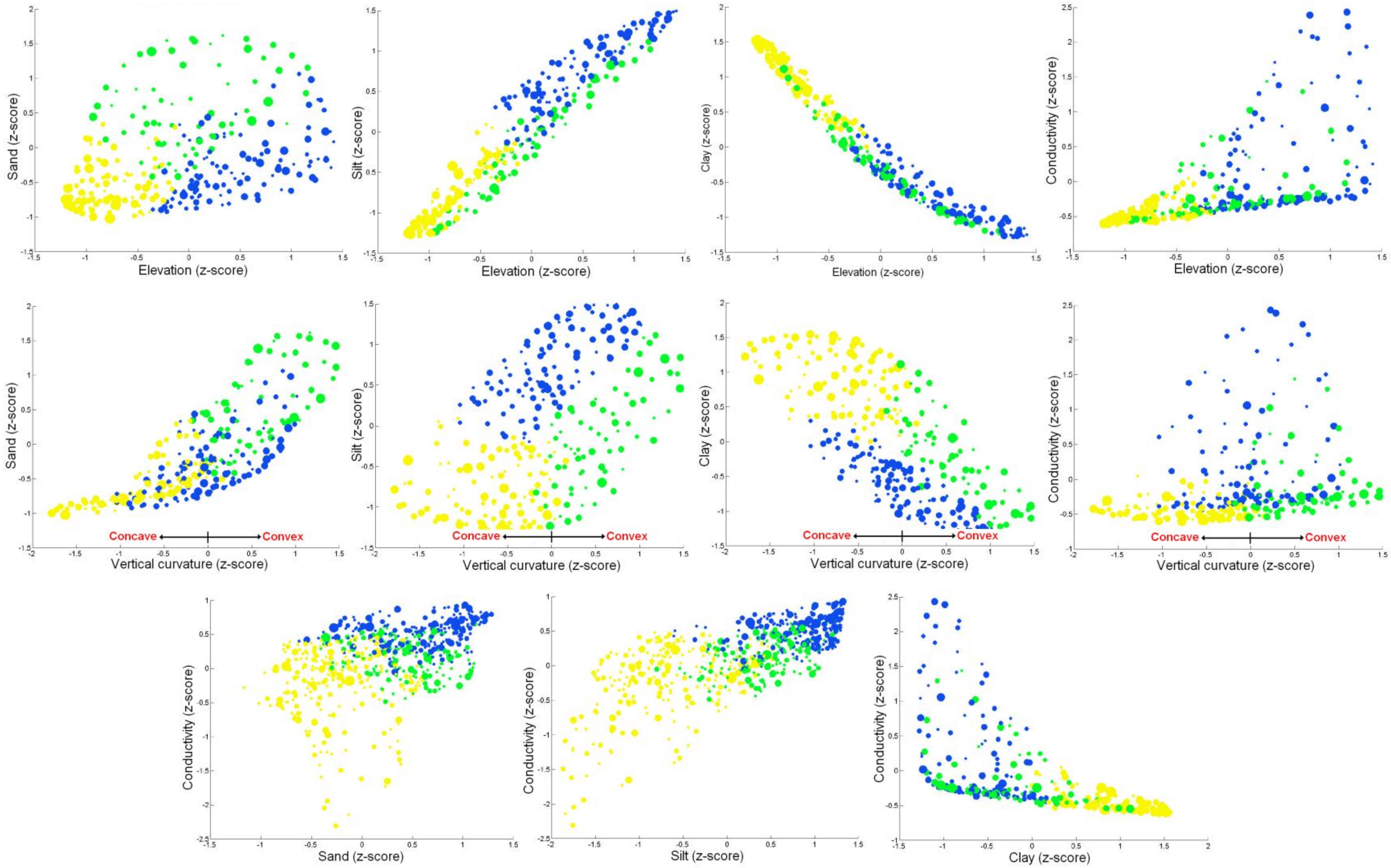


Fig. 6 – Scatterplots of the nodes, clustered by k-means technique, comparing each dependent variable against elevation, vertical curvature and between themselves. Each cluster can be analyzed to its correspondent color in all the plots.

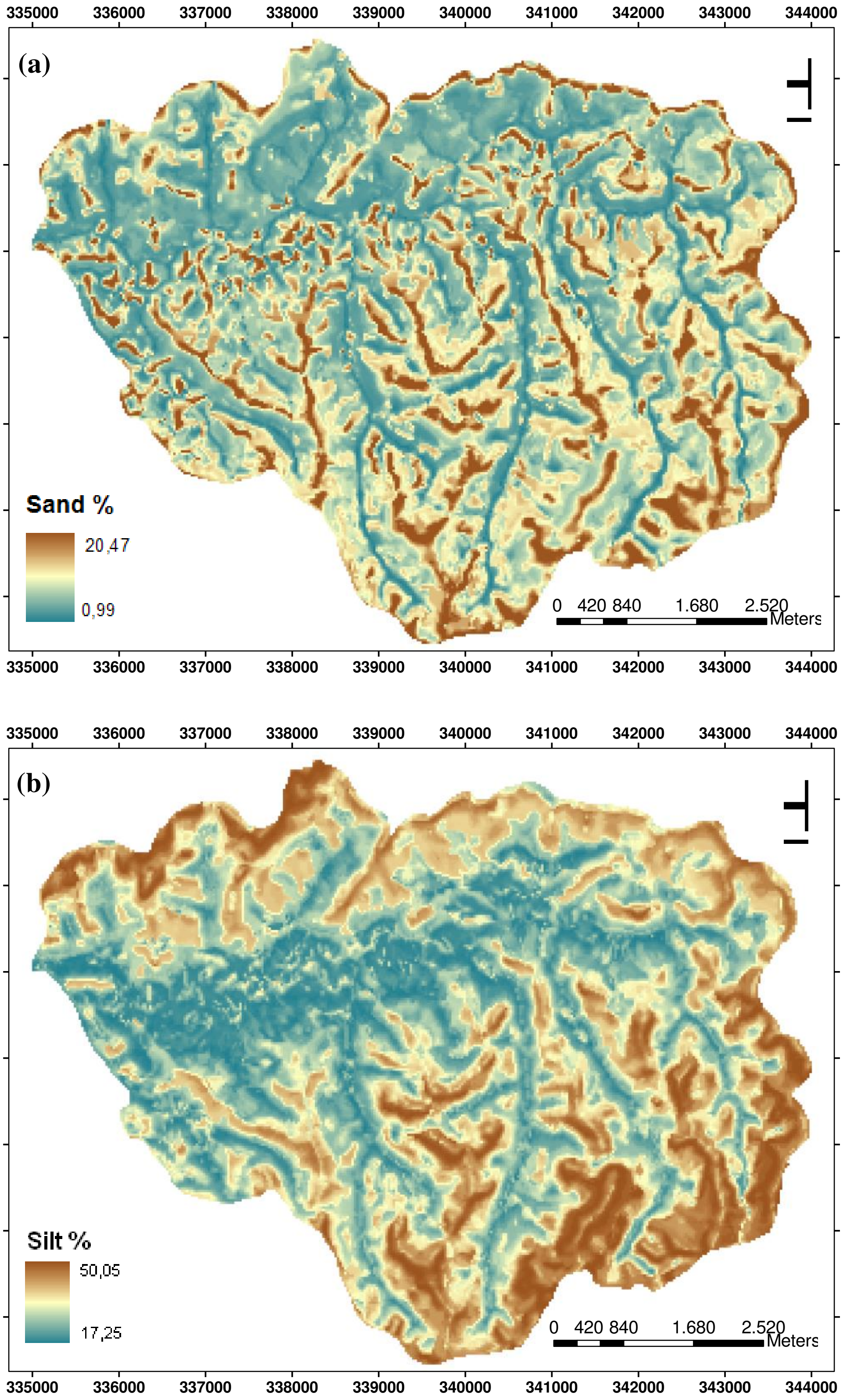


Fig. 7 – Final product of the proposed method. Predicting maps for a) Sand (%); b) Silt (%); c) Clay (%); and d) Hydraulic conductivity (cm/s). The predicted textures are in agreement with Moraes (2008) analysis, where the soils have high content of clay throughout the basin.

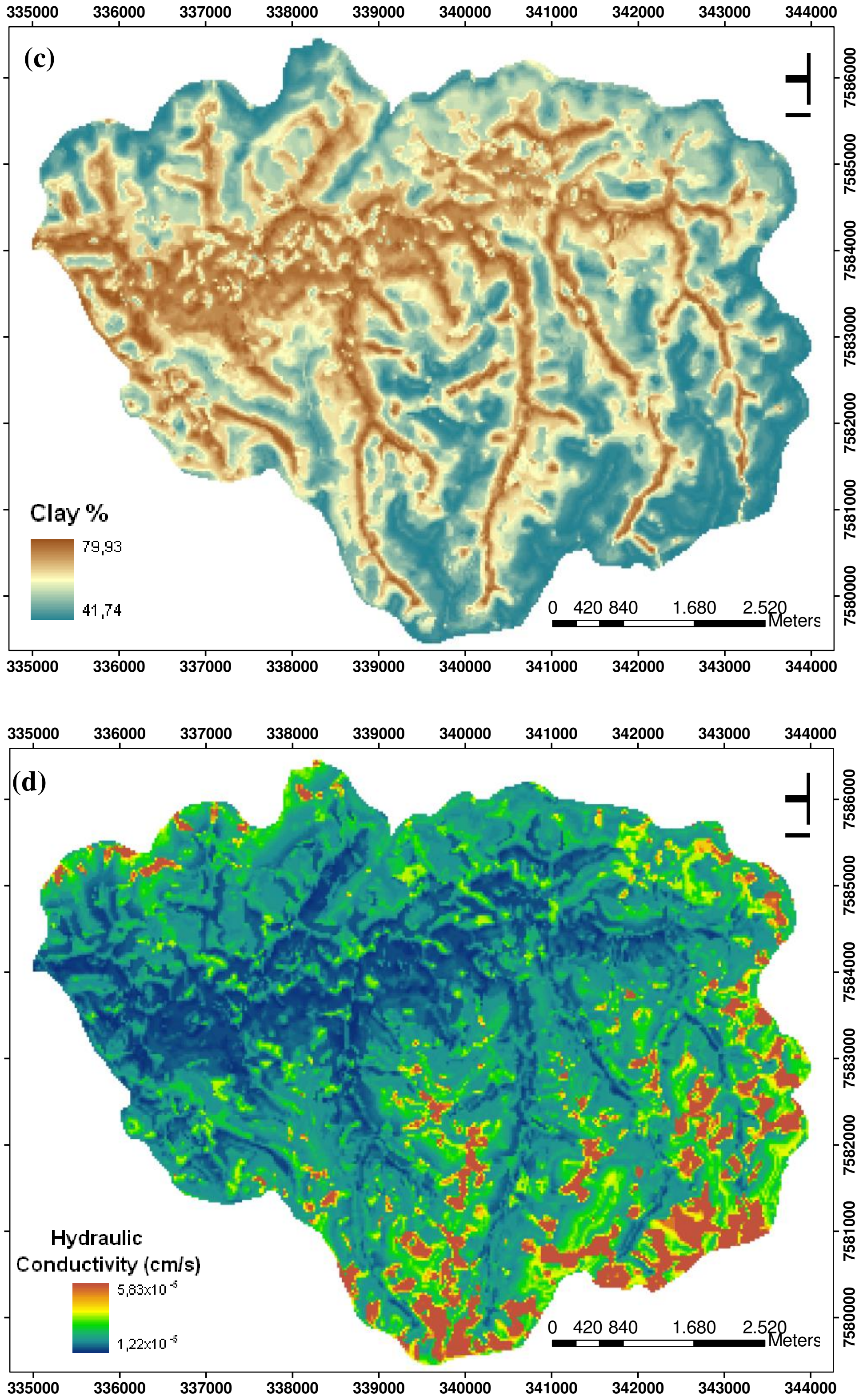
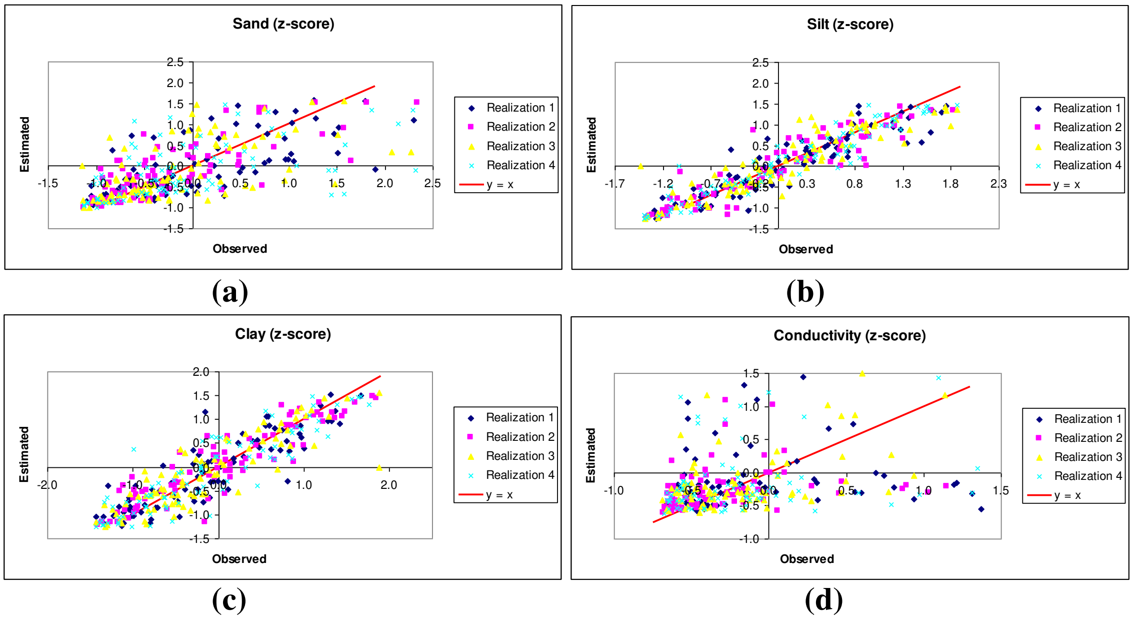


Fig. 7 cont. – Final product of the proposed method. Predicting maps for a) Sand (%); b) Silt (%); c) Clay (%); and d) Hydraulic conductivity (cm/s). The predicted textures are in agreement with Moraes (2008) analysis, where the soils have high content of clay throughout the basin.

4.4 Cross-validation

The model performance is evaluated using a stochastic cross-validation method, the Poor man's approach (Rao et al, 2008) according to the following steps: (a) the simulated data (n = 1000) is split intro training (n = 920) and validation (n = 80) sets (where the validation set is randomly drawn); (b) the training set is used to generate a SOM model where the 80 missing points are estimated; (c) next, the predicted values were compared to the observed values to confirm a one-to-one correspondence to assess model bias; (d) then from the original dataset a different subsample of 80 elements are randomly taken and the steps b and c are repeated. These procedures are carried out 4 times producing the cross-validation plots presented in Fig. 8. The SOM model demonstrates unbiased behavior indicated by the one-to-one correspondence and constant variance for clay, silt and sand. The conductivity plot reveals higher uncertainty, which we assign to the inherent field experience limitations, while the soil texture is analyzed under controlled conditions in a laboratory, decreasing the uncertainties measurements.



(a) **(b)**

(c) **(d)**

Fig. 8 – Cross validation for the predicting models generated by the proposed procedure, using Poor man's approach. a) Sand; b) Silt; c) Clay; and d) hydraulic conductivity. One to one correspondence reveals that the SOM is an unbiased predictive model.

5. CONCLUSIONS

In this study, we find that the self-organizing map (SOM) technique and Monte Carlo simulation method can be used to evaluate and build an unbiased predicting model for soil physical properties on hillslopes at the local scale using published data from 30 field locations. Some of the noteworthy conclusions are as follows:

(1) Grouping nonlinear relations using the k-means clustering technique facilitates the development of conceptual hillslope models for further understanding of the soil weathering processes. It is possible to identify relations among variables in the hillslope (SAND, CLAY, SILT, CONDUCTIVITY) at the local scale. Many variables in certain categories are highly correlated indicating an unnecessary redundancy in describing system contributions. Topographic morphometry, hydraulic conductivity near the surface and soil texture data can be projected onto a 2-dimensional surface while preserving important nonlinear relations, for building an unbiased predicting model.

(2) The convex-concave hillslope feature, relates to soil production rate, soil erosion, and therefore soil thickness, where SOM analysis unveiled the relations, in terms of topological array, between soil physical properties and terrain morphometry, decreasing the low density sampling limitation through Monte Carlo stochastic simulation approach describing the physical factors that may influence the particle transport phenomenon.

(3) The proposed method is suitable to survey soil chemical and physical properties, revealing and quantifying relationships between soil variables and terrain morphometry, not properly observed by linear multivariate statistical approaches, additionally, does not have any assumptions for the collected dataset, where all the analyses were conducted preserving all variables and keeping the associated parameters. We expect the proposed modeling method to be used as an alternative approach for further studies exploring hillslope weathering, erosion, and hydrological processes in other locations and environmental settings.

Acknowledgements

We are grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and Conselho Nacional de Desenvolvimento Científico e Tecnológico for their financial support; to Victor F.

Labson, Director, Crustal Imaging and Geochemistry Science Center (CIGSC), U.S. Geological Survey (USGS), Denver, Colorado, United States of America for providing the first author with the position of visiting scientist.

REFERENCES

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000a. Artificial neural networks in hydrology. I: preliminary concepts. Journal of Hydrologic Engineering 5, 115-123.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000b. Artificial neural networks in hydrology. II: hydrologic applications. Journal of Hydrologic Engineering 5, 124-137.

Breiman, L., 1996. Bragging Predictors. Machine Learning 24, 123-140.

Camargo, O. A., Moniz, A. C., Jorge, J. A., Valadares, J. M. A. S., 1986. Métodos de análise química, mineralógica e física de solos do Instituto Agronômico de Campinas. Report - IAC 106, Instituto Agronômico de Campinas, 94 pp.

Chapman, N. A., McKinley, I. G., Franca, E. P., Shea, M. E., Smellie, J. A. T., 1992. The Poços de Caldas project: an introduction and summary of its implications for radioactive waste disposal. Journal of Geochemical Exploration 45, 1–24.

Christofoletti, A., 1970. Análise morfométrica das bacias hidrológicas do Planalto de Poços de Caldas–MG. Livre Docente Dissertation, Faculdade de Filosofia, Ciências e Letras, Rio Claro, Brazil.

Daniels, R. B., Hammer, R. D., 1992. Soil Geomorphology, John Wiley & Sons Inc., Toronto.

Davies, D.L., Bouldin, D.W., 1977. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227.

Dickson, B.L., Giblin, A.M., 2007. An evaluation of methods for imputation of missing trace element data in groundwaters. Geochemistry: Exploration, Environment, and Analysis 7, 173-178.

Ehsani, A. H., Quiel, F., 2008. Geomorphometric feature analysis using morphometric parametrization and artificial neural networks. Geomorphology 99, 1-12.

Empresa Brasileira de Pesquisa Agropecuária–Embrapa, 1999. Sistema Brasileiro de Classificação de Solos, Embrapa Produção de Informação, Rio de Janeiro.

Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. American Geophysical Union EOS 81, 583–585.

Fernandes, H. M., Franklin, M. R., 2001. Assessment of acid rock drainage pollutantsrelease in the uranium mining site of Poços de Caldas – Brazil. Journal of Environmental Radioactive 54, 5–25.

Fessant, F., Midenet, S., 2002. Self-organizing map for data imputation and correction in surveys. Neural Computing Applications 10, 300-310.

Fishman, G. S., 1996. Monte Carlo - concepts, algorithms, and applications, Springer-Verlag, Berlin.

Forgy, E., 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications, Biometrics 21, 768-781.

Friedel, M.J., Iwashita, F., *In press*, Nonparametric modeling of autocorrelated random variables for application to inverse problems, Environmental Modelling & Software.

Grohmann, C. H., Riccomini, C., Alves, F. M., 2007. SRTM-based morphotectonic analysis of the Poços de Caldas Alkaline Massif, southeast Brazil. Computer & Geosciences 33, 10-19.

Hentati, A., Kawamura, A., Amaguchi, H., Iseri, Y., 2010. Evaluation of sedimentation vulnerability at small hillslide reservoir in the semi-arid region of Tunisia using Self-Organizing Map. Geomorphology 122, 56-64.

Holmes, D. C., Pitty, A. E., Noy, D. J., 1992. Geomorphological and hydrogeological features of the Poços de Caldas caldera analogue study sites. Journal of Geochemical Exploration 45, 215–247.

Heimsath, A. M., Dietrich, W. E., Nishizumi, K., Finkel, R. C. 1997. The soil production function and the landscape equilibrium. Nature 388, 358-361.

Hornberger, G. M., Raffensperger, J. P., Wiberg, P. L., Eshleman, K. N., 1998. Element of physical hydrology, The John Hopkins University Press, Baltimore.

Hugget, R. J., 1998. Soil chronosequences, soil development, and soil evolution: a critical review. Catena 32, 155–172.

Iwashita, F., Friedel, M. J., Souza Filho, C. R., Fraser, S. J., *in press*. Hillslope chemical weathering across Paraná state, Brazil: A data mining–GIS hybrid approach, Geomorphology. http://dx.doi.org/10.1016/j.geomorph.2011.05.006

Jensen, J. R., 2007. Remote sensing of environment: an earth resource perspective, second edition. Pearson Prentice Hall, New Jersey.

Jiménez-Rueda, J. R., Nunes, E., Mattos, J. T., 1993. Caracterização fisiográfica e morfoestrutural da Folha São José de Mipibu, RN. Geociências 12, 481–491.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen J., Kolehmainen, M., 2004. Methods for imputation for missing values in air quality data sets. Atmospheric Environment 38, 2895-2907.

Kalteh, A.M., Berndtsson, R., 2007. Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). Hydrological Science Journal 52, 305-317.

Kalteh, A. M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software 23, 835-845.

Kalteh, A.M., Hjorth, P., 2009. Imputation of missing values in precipitation-runoff process database. Hydrology Research 40, 420-432.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence, Montreal, Canada, pp.1137-1143.

Kohonen, T., 2001. Self-organizing Maps, third edition, Springer-Verlag, Berlin.

Krajewski, W. F., Lakshmi, V., Georgakakos, K. P., Jain, S., 1991. A Monte Carlo study of rainfall sampling effect on a distributed catchment model. Water Resources Research 27, 119-128.

Löhr, S. C., Grigorescu, M., Hodgkinson, J. H., Cox, M. E., Fraser, S. J. 2010. Iron Occurrence in soils and sediments of coastal catchment: A multivariate approach using self-organizing maps. Geoderma 156, 253-266.

Loke, M. H., Barker, R. D., 1996. Practical techniques for 3D resistivity survey and data inversion. Geophysical Prospecting 44, 499-523.

Malek, M.A., Harun, S., Shamsuddin, S.M., Mohamad I., 2008. Imputation of time series data via Kohonen self organizing maps in the presence of missing data. Engineering and Technology 41, 501-506.

Moraes, F. T., 2008. Zoneamento geoambiental do planalto de Poços de Caldas, MG/SP a partir de análise fisiográfica e pedoestratigráfica. Ph.D. Dissertation. Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista, Brazil.

Mudd, S. M., Furbish, D. J. 2004. Influence of chemical denudation on hillslope morphology. Journal of Geophysical Research 109, F02001.

Neter, J., Kutner, M. N., Nachtssheim, C. J., Wasserman, W. 1996. Applied linear statistical models, 4th Ed. WCB/McGraw-Hill, Boston.

Nezat, C. A., Blum, J. D., Klaue, A., Johnson, C. E., Siccama, T. G., 2004. Influence of landscape position and vegetation on long-term weathering rates at the Hubbard Brook Experimental Forest, New Hampshire, USA. Geochimica et Cosmochimica Acta 68, 3065-3078.

Penn, B. S. 2005. Using self-organizing maps to visualize high-dimensional data. Computer & Geosciences 31, 531-544.

Rao, C. R., Toutenburg, H., Shalabh, Heumann, C., 2008. Linear models and generalizations: least squares and alternatives, third edition, Springer-Verlag, Berlin.

Rallo, R., Ferre-Gine, J., Arenas, A., Giralt, F., 2002. Neural virtual sensor for the inferential prediction of product quality form process variables. Computers and Chemical Engineering 26, 1735-54.

Reynolds, W. D.; Elrick, D. E., 1989. Hydraulic conductivity measurement in the unsaturated Zone using improved well analyses. Ground Water Monitoring & Remediation 9, 184-193.

Ribeiro, G. F., 2010. Estudo do meio físico que influenciam a capacidade de infiltração das águas da bacia hidrográfica Vargem de Caldas, Minas Gerais. Master Thesis, Instituto de Geociências, Universidade de Campinas, Brazil.

Rubinstein, R.Y., Kroese, D.P., 2007. Simulation and the Monte Carlo Method. 2nd ed. New York: John Wiley & Sons.

Valeriano, M. M., 2008. Topodata: Guia para utilização de dados geomofométricos locais. Report INPE-15318-RPQ/818, Divisão de Sensoriamento Remoto, Instituto Nacional de Pesquisas Espaciais, 75p.

Valeriano, M. M., Kuplich, T. M., Storino, M., Amaral, B., Mendes, J. N., Lima, D. J., 2006. Modeling small watersheds in Brazilian Amazonia with shuttle radar topographic mission-90 m data. Computer & Geosciences 32, 1169-1181.

Valeriano, M. M., Rosetti, D. F., Albuquerque, P. C. G., 2009. TOPODATA: desenvolvimento da primeira versão do banco de dados geomorfométricos locais em cobertura nacional. XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, pp. 5499-5506.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 2000. SOM Toolbox for Matlab 5. SOM toolbox team, Finland. Helsinki University of Technology, Laboratory of Computer and Information Science. http://www.cis.hut.fi/projects/somtoolbox/

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Neworks 11, 586-600.

Wang, S., 2003. Application of Self-organising maps for data mining with incomplete data sets. Neural Computing and Applications 12, 42-48.

Yoo, K., Amundson, R., Heimsath, A. M. Dietrich, W. E., Brimhall, G. H., 2007. Integration of geochemical mass balance with sediment transport to calculate rates of soil loss chemical weathering and transport on hillslopes. Journal of Geophysical Research 112, F02013.

Young, A., 1980. Tropical soils and soil survey. Cambridge University press, Cambridge.

ESTIMATING PHYSICAL-CHEMICAL PROPERTIES IN FRACTURED AQUIFERS USING SELF-ORGANIZING MAPS IMPUTATION APPROACH: STUDY OF HYDRAULIC CONNECTIVITY BETWEEN SERRA GERAL AND GUARANI AQUIFER IN PARANA STATE, BRAZIL

Abstract

Self-organizing maps (SOM) is here used to imputate missing values of hydraulic transmissivity and hydrochemistry and to evaluate hydraulic connections between the Serra Geral (SGAS) and Guarani (GAS) aquifer systems in State of Parana, Brazil. A spatial model for the aquifer is built and elements of spatial variability compared with -current conceptual and hydrochemical models. SOM is employed to calculate nonlinear correlations between 27 variables from 976 wells in the Serra Geral aquifer. These include hydrochemical (19), geophysical (1) and morphometric (5) variables, plus others derived from digital elevation model geoprocessing (2). Using a second dataset with 156 samples, transmissivity is estimated from well pump tests. Results and related parameters are used to train the SOM for imputation process. The K-means technique is used to find relevant clusters, whereas Davies-Bouldin index indicates the optimal numbers of groups. SGAS typical waters are carbonate-calcium and carbonate-magnesium, whereas GAS has sodium, chloride, fluoride and sulfate as characteristic elements. The analysis of flux connections between the two systems is based on anomalous hydrochemistry spatial behaviors. SOM predicted points are consistent with current connectivity models, where vertical fluxes from GAS are possibly strongly influenced by geological structures. SOM imputation and pos-processing by the K-means approach revealed different hydrochemical facies for the SGAS and several new areas with possible connections between the two aquifer systems, constituting a feasible alternative to deal with missing values in a multivariate dataset.

Keywords: hydrochemistry, MATLAB, data minig, artificial neural networks

1. INTRODUCTION

The Serra Geral aquifer is one of the largest and most important in Brazil. It is a transboundary, unconfined and fractured aquifer formed within a sequence of lower Cretaceous Parana flood basalts (Thiede and Vasconcelos, 2010). Its outcropping surface reaches 1.2 million km^2. The Serra Geral aquifer system (SGAS) covers four Brazilian states and three countries, Argentina, Uruguay and Paraguay (Fig. 1a). It is responsible for the confinement of the Guarani Aquifer System (GAS) in the middle sections of the Parana basin - i.e., a porous aquifer composed primarily by sandstone from the Botucatu and Pirambóia Formations (Sracek and Hirata, 2002).

Despite the apparent confinement of the GAS, many studies have shown a hydraulic communication between the two systems based on hydrochemistry (Fraga, 1986). In this scenario, favorable hydraulic conditions associated with geological discontinuities enables the rise of stored water from the GAS to the SGAS (Fig. 2). The common approach to point out potential hydraulic connectivity areas between the two aquifers is to analyze characteristic elements from GAS, comparing their relative concentrations and spatial distribution with typical SGAS waters (Mocellin, 2009; Nanni et al., 2009; Silva, 2007; Rosa Filho et al., 2006; Ferreira et al., 2005, Bittencourt et al., 2003; Portela Filho, 2003, Sracek and Hirata, 2002).

The integrated analysis of geological structures and variation of chemical elements is the most frequently employed methodology to evaluate possible connections between the Serra Geral and the Guarani aquifers (Ferreira et al., 2005). Remote sensing and geophysical data have been used to support the study of aquifers when characterizing geological structures (Nanni et al., 2009), or as a predictive variable in the absence of monitoring wells (Souza Filho et al., 2010). Additionally, there is a possible influence from relief (morphological) features over water chemistry, particularly where a thick layer of soil combined with high clay content may prevent the recharge of the SGAS unconfined section (Nanni et al., 2009). When these conditions are associated with vertical faults, the relative concentration of Guarani trait elements increase and the connection between the two systems becomes more evident.

The characterization of the hydrochemistry of aquifers can be done through the Piper diagram - a ternary graph and a linear measure representing the relative variations of sulfate, chloride, carbonate + bicarbonate (anions axis) and calcium, magnesium and sodium (cations axis). Data for pH, major cations and major anions are widely available worldwide since they are commonly used to depict the groundwater hydrochemistry; nevertheless in most cases such databases are still incomplete or inconsistent.

Most parametric statistical methods, such as analysis of variance (Winter et al., 2006), requires a complete data matrix to calculate multivariate exploratory measures, as the covariance matrix. Other methods, like cluster analysis (Suk and Lee, 1999), principal component analysis (Astel, et al., 2007) and factor analysis, require computation of eigenvalues and eigenvectors (Netter et al., 1996). An alternative to deal with missing values are imputation methods (Malek et al., 2008), which comprise statistical and mathematical approaches to estimate missing values in datasets based on a combination of the available data (Dickson and Giblin, 2007). The self-organizing maps - SOMs (Kohonen, 2001), a type of unsupervised neural networks have been used to characterize and survey groundwater chemistry (Lu and Lo, 2002; Sánchez-Martos et al., 2002; Hong and Rosen, 2001), and also have been employed as an imputation method (Wang, 2003) for precipitation and run-off processes (Kalteth and Hjorth, 2009), air quality datasets (Junninen et al. 2004), data survey (Fessant and Midenet, 2002) and detection of unexploded ordnances (Benavides et al. 2009).

The SOM is considered a technique to visualize high dimensional data sets, representing them in two or three dimensions, projected onto maps composed by code vectors (ASCE, 2000a; 2000b). Each code vector has the same dimension as the input data array. Through an iterative process, the SOMs are trained to fit the input data set, whereas each sample has an associated n-dimensional vector (Kohonen, 2001). The SOM's attribute of learning vector quantization using Euclidean distance while preserving topological relations between the samples, makes it an inherently robust imputation method (Dickson and Giblin, 2007).

The main objective of this work is to build a spatial model for hydrochemistry and hydraulic transmissivity of the Serra Geral fractured aquifer in the State of Parana, Brazil, using the self-organizing map imputation method, and comparing estimated values with conceptual models established in the literature. To achieve this goal, the following specific objectives will be met: (1) use SOM to determine nonlinear correlations between hydrochemical elements, relief morphometry, and aeromagnetic data; (2) use SOM to imputate missing values in groundwater database; (3) apply the k-means method to find the relevant clusters (Davies and Bouldin, 1977);

(4) find possible flux connections between the Serra Geral e the Guarani aquifers, based on hydrochemistry anomalous spatial behavior.

2. SELF ORGANIZING MAPS

In a SOM analysis each sample is treated as a vector in a data space determined by its variables. Measures of vector similarity are then used to order and segment the input data into meaningful natural patterns (Fraser and Dickson, 2007). In an iterative process, "seedvectors" are modified to represent the distribution of the input data in the data space and, once trained, become known as "best-matching units" (BMUs). The SOM output usually consists of an ordered array of nodes (the BMUs) arranged in a regular, two-dimensional grid (the map). Each input sample is represented by the closest BMU and the BMUs are arranged on the map in such a way, as to preserve topology (Kohonen, 2001). Thus, data points lying close to each other in the n-dimensional input space are mapped onto proximal BMUs. This characteristic is important because it allows the analysis to preserve the input space topology, a feature which, together with the ability to learn and organize information without being given the associated dependent output values for the input pattern, makes it possible to use the SOM as a dimensionality reduction tool (Fraser and Dickson, 2007).

The U-matrix representation of the map indicates the closeness between adjacent nodes on the map in terms of Euclidean distance (Bierlein et al., 2008). A color scale is used so that cool colors (blues) separate adjacent nodes that are closer or similar, while warm colors (reds) indicate larger distances and greater differences between the nodes. To assist in this display, alternate dummy nodes are added to the U-matrix. These are colored according to the distance between adjacent nodes, whereas the nodes that represent actual vectors are colored according to the average of the distances to their neighbors. A group of nodes with small distances between them form a cluster. Zones of nodes with large distances between them separate clusters.

Component plots show the variation of a particular variable across the map using a color-temperature scale. The highest values correspond to red regions and the lowest values to blue zones. The component plots are used to determine the zones (units on the map) where the variable value is high or low, and to observe any correlation or relationship among the variables (López García and Machón González, 2004). These correlations can be detected by means of the color gradient on each component plane. Two variables with parallel gradients show a direct correlation, whereas inverse gradients show a negative correlation.

The similarity index plot is an additional presentation of the component plots that aids their visual comparison. The principal components of the SOM output data are extracted and the component plots are then plotted on the similarity index, with the x and y axes representing principal components 1 and 2, respectively (López García and Machón González, 2004). Thus, components that plot closer together on the similarity index are more similar to each other than they are to distant components.

An initial idea of the number of clusters in the SOM, as well as their spatial relationships, can often be acquired by visual inspection. However, relationships can be quite complex and it can be difficult to make a useful interpretation based solely on the U-matrix and component plots. To assist in this process, the SOM may be divided into similar regions, by applying the non-hierarchical K-means clustering algorithm to the BMUs. An internally derived, data driven estimate of the optimum number of clusters can be selected using a Davies–Bouldin analysis (Davies and Bouldin, 1977), which is implemented in the SiroSOM software. Together with the component plots (Kohonen, 2001) and analysis of the SOM output data, clusters derived in this manner allow a sophisticated analysis of the relationships amongst the variables and between samples.

3. STUDY AREA

The Serra Geral aquifer system (SGAS) outcropping area is comprised in the State of Parana state and corresponds to 109,000 km^2 (Fig. 1a). The SGAS is classified as anisotropic, fractured and hosted by crystalline rocks. The water flows through fractures, cracks and gaps opened by tectonic displacements and weathering. The storage capacity of such fissured aquifers depends on fracture density, how large are fracture gaps and how significant are the communication between these structures (Fraga, 1986). Therefore, yield from wells drilled into fractured aquifers depends essentially on the number and density of fractures.

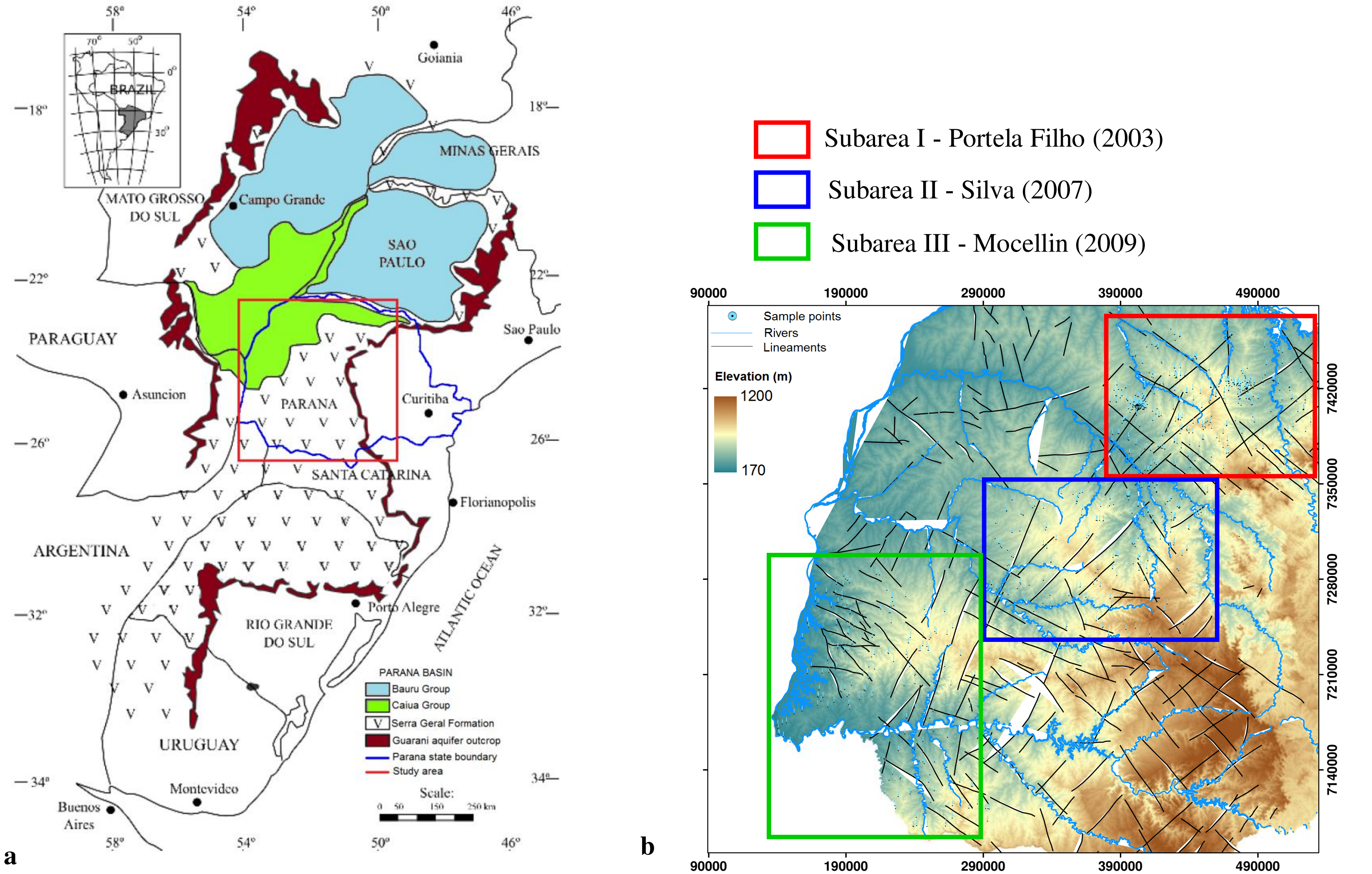


Fig. 1 – Geological Setting. (a) The Serra Geral Formation limits (note that it is comprised in four different South American countries (modified from Nanni et al., 2009). (b) Elevation map with lineaments and rivers draped over. Three subsets of data employed in the study area gathered by (I) Portela Filho (2003), (II) Silva (2007) and (III) Mocellin (2009).

The SGAS is hosted mainly by volcanic rocks as tholeiitic basalts, andesites, rhyolites and rhyodacites (Harris and Milner, 1997). The thickness of the Serra Geral Formation volcanic rocks increases from east to west, reaching 1500 meters in the central sector of the Parana Basin (Peate et al., 1988). The outcropping rocks generally present aphanitic and other micro-crystalline textures with massive or vesicular-amygdalous structure. Dikes and sills of tholeiitic and rhyodacitic composition are widespread (Turner et al., 1999).

The main processes conditioning the water chemistry of the Serra Geral aquifer are the weathering of the basaltic rocks and the associated equilibrium with secondary minerals. The geochemical interaction between percolating water and aquifer rocks along the recharge and discharge zones are crucial to define hydrochemical characteristics (Bittencourt et al. 2003). Given the lithologic characteristics of Serra Geral Formation, the SGAS water is classified as calcium and magnesium bicarbonates-rich (Fraga, 1986). However, the type of water can also be affected by mixing water from different aquifers (Fig. 2). Besides, the time interval the water stays in contact with soluble materials that constitute the aquifer is positively correlated with the total dissolved solids content.

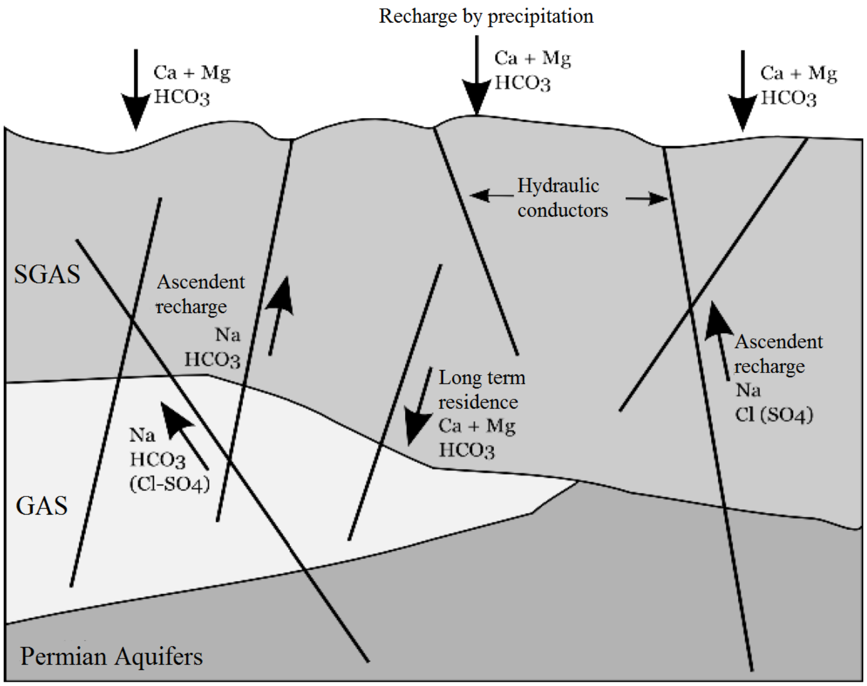


Fig. 2 – Conceptual model of the connectivity between the Guarani (GAS) and the Serra Geral (SGAS) aquifer system, representing the influence of geological structures and hydrochemistry characteristics (Nanni et al., 2009).

The hydrochemistry of the Guarani aquifer system (GAS) waters is highly variable, especially in confined areas, caused by faciologic variations, or by mixture, associated with sandstone fractures (Gastmans et al., 2010). In deep confined areas, the GAS waters are not suitable for water supply given the high content of total dissolved solids, high concentration of sulphates and fluorides above the recommended limits for human consumption (Nanni et al., 2009).

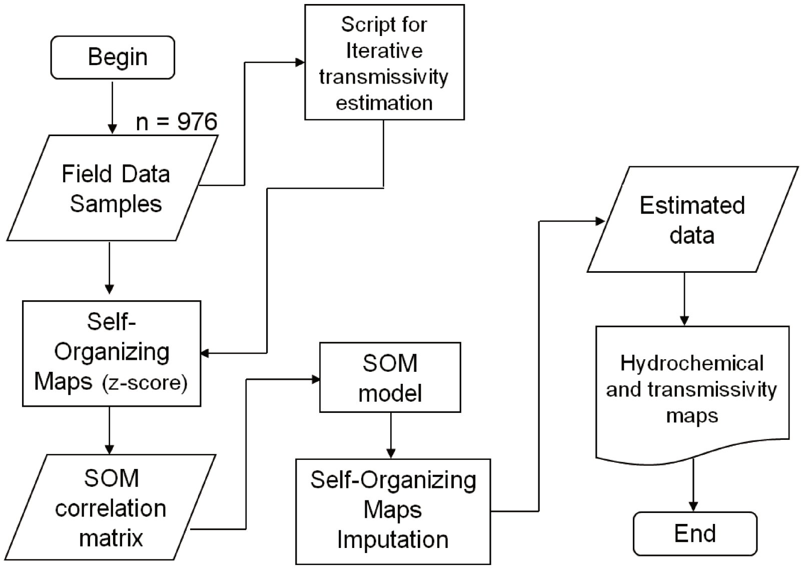
The waters in the GAS outcropping region is calcium bicarbonate-rich in composition, changing to sodium bicarbonate with increasing concentrations of chloride and sulfate towards deeper confined areas (Fig. 2) (Rabelo and Wendland, 2009). The variation is caused by a decrease in calcium content, exchanged with sodium, and caused by carbonate dissolution leading to a sodium bicarbonate groundwater type. At least part of the sodium, chloride and sulfate are likely a contribution from the Pirambóia Formation, originated from evaporite dissolution (Gastmans et al., 2010).

4. METHODS

When considering free aquifers such as the SGAS, surface hydrology and groundwater chemistry have to be investigated as coupled processes. Groundwater chemistry may be correlated to surface and near surface hydrologic phenomena, where variables such as relief features, soil texture and aeromagnetic data can be used as predicting variables. The use of these variables connect near surface phenomena like soil chemical weathering (Iwashita et al. 2011) and sediment transport to hydrogeological processes, making possible to predict groundwater quality (Souza Filho et al., 2010), and aquifer recharge (James et al., 2010). The dataset employed in this work is composed of 27 variables, 976 samples, including relief features (5), variables derived from digital elevation model geoprocessing (2), geophysic measure (1) and hydrochemical and well parameters (19). Six steps were used to model the spatial distribution of hydrochemical elements and hydraulic transmissivity (Friedel and Iwashita, 2011) (Fig. 3). First, well parameters were used to calculate hydraulic transmissivity using McLin's (2005) method. Second, all data variables were standardized so that not single variable would dominate in the nonlinear modeling process (Kalteth et al., 2008). The z-score transformation is given by:

$$z_i = \frac{x_i - \bar{x}_i}{s_i}$$

where z is the standardized value; x is the raw score; $\bar{x}$ is the sample average, and s is the sample standard deviation, i is an index for each variable. Standardizing variables in this way resulted in each having an expected value of zero and standard deviation one.



Begin
n = 976
Field Data Samples
Script for Iterative transmissivity estimation
Self-Organizing Maps (z-score)
SOM correlation matrix
SOM model
Self-Organizing Maps Imputation
Estimated data
Hydrochemical and transmissivity maps
End

Fig. 3 – Flowchart of the proposed method to imputate missing values for hydrochemistry and transmissivity data.

Third, the SOM (Kohonen, 2001) was used to self organize nonlinear relations among the 27 variables. The correlation matrix was extracted from the generated model and so was the imputation for the missing values. Fourth, the k-means clustering technique was then used to classify the SOM topography into statistically relevant groups. Fifth, The hydrochemical elements were projected into a continuous surface to evaluate their spatial distribution. In the final step, the spatial pattern of chemical elements was compared to previous conceptual models established in the literature in order to indentify sectors in the study area where potential connections between the aquifer systems may exist.

4. 1. Hydrochemical data

The hydrochemical data were provided by public institutions responsible for water monitoring and analysis in the State of Parana, including the Waters Institute (Aguas Parana) and the Sanitation Company (SANEPAR). The dataset is a composed by three subsets (Fig. 1b), also employed by Portela Filho (2003), Silva (2007) and Mocellin (2009). It has 19 variables and 976 samples (Table 1), containing major cations and anions and characterization parameters of wells.

However, given the constant advancements in chemical analytical instruments and development of new protocols, such databases often differ in which set of elements was analyzed or in accuracy of such analysis. This leads to a fairly inconsistent dataset. From the 976 samples, only 93 contain fluoride measurements, making the imputation, in this particular case, essential to study the spatial behavior of the variable. The SOM is an imputation method without any statistical assumptions that can achieve high performance even with a low number of samples for its training.

Table 1 – Acronyms for employed variables and their respective unit.

Category	Variable		
	Description	Acronym	Unit
Morphometric	Aspect	cosASPECT	dimesionless
	Elevation	ELEVATION	m
	Horizontal curvature	HOR-CURV	°/m
	Slope	SLOPE	°
	Vertical curvature	VERT-CURV	°/m
GIS	Flow accumulation	FLOWACC	integer
	Distance from lineaments	LINEASDIST	m
Geophysical	aeromagnetic	AEROMAG	*n* T
Well information	Potenciometric Level	POTENC	m
	Depth	DEPTH	m
	drawdown	DRAWDOWN	m
	Yield	YIELD	m3/h
	Specific Capacity	CAPAC	m3/h.m
Hydrochemical	Calcium	Ca	mg/L
	Magnesium	Mg	mg/L
	Sodium	Na	mg/L
	Potassium	K	mg/L
	chloride	Cl	mg/L
	Sulfate	SO4	mg/L
	Carbonate	CO3	mg/L
	Bicarbonate	HCO3	mg/L
	pH	pH	
	total dissolved solids	TDS	mg/L
	Free CO2	CO2	mg/L
	Nitrate	NO3	mg/L
	Carbonate bicarbonate	CO3-HCO3	
	Fluoride	F	mg/L

4. 2. Hydraulic transmissivity data

The Serra Geral aquifer transmissivity was estimated using specific capacity data calculated from pumping tests conducted in 154 wells by the Águas do Parana Institute, based on a modified version of the Bradbury-Rothschild iterative solution technique (Bradbury and Rothschild, 1985), here adapted for MATLAB:

$$T = \frac{Q}{4\pi(s_t - s_w)}\left[\ln\left(\frac{2.25Tt}{r_w^2 s}\right) + 2s_p\right]$$

where T = aquifer transmissivity (cm/min), Q = well discharge (m^3/min), s_t = is total drawdown observed in a production well, s_w = drawdown due to well loss S = aquifer storage coefficient (dimensionless), t = time since pumping began (min), r_w = effective wellbore radius (cm), and s_p = a partial penetration factor (dimensionless).

Since T appears on both sides of the equation, an iterative solution is required (Bradbury and Rothschild, 1985). Initially, a guess is made for T (T_{guess} in the program) on the right-hand side of the equation, and an updated solution for T (T_{calc} in the program) is obtained from the left-hand side. This updated solution is again used on the right-hand side of the equation, and a new T is again computed. This iterative process continues until some suitable tolerance criterion for error is reached (McLin, 2005).

The calculated transmissivity was used as a training set for the SOMs. The computed model estimated the missing values at places where the pumping tests information was incomplete (Fig. 3).

4. 3. Aeromagnetic data

The aeromagnetic data were provided by PETROBRAS and pre-processed and by the Applied Geophysics Laboratory of Research (LPGA) from Parana Federal University. The raw data comprise a series of aerial surveys conducted in the Parana Basin during the 80s. This survey was conducted by PETROBRAS in 1981, with NS-trending lines spaced of 2 km, a flight height of 500 m and sampling intervals of approximately 100 m. Control lines spaced of 20 km and perpendicular to the acquisition lines were also acquired during the survey. The pre-processing comprised the generation of regular grids (500x500 m) by the minimum curvature method (Briggs, 1974). The residual magnetic field was then calculated, and artifacts (noise) along the flight lines were eliminated using the micro-leveling technique (Minty, 1991). Thus, the magnetic data represent the micro-leveled anomalous magnetic field (*n*T).

4. 4. Topographical dataset

Characterization of the topographic relief was possible using elevation data provided by the Shuttle Radar Topographic Mission (http://edcsns17.cr.usgs.gov/NewEarthExplorer/). The digital elevation model associated with these data was provided by the United States Geological Survey on a lattice with 90-m spatial resolution (Farr and Kobrick, 2000). The Topodata project, conducted by the Brazilian National Institute for Space Research-INPE (Valeriano et al., 2009), has created derived metrics data with a 30-m resolution, based on elevation data and geographical information system (GIS) modeling techniques. The geomorphometric features provided a way to extract morphometric features, such as slope, aspect (hillslope orientation), vertical and horizontal curvature (Valeriano et al., 2006), and accumulated hydrological flux (Jenson and Domingue, 1988).

5. Results and discussions

To identify potential areas of hydraulic connectivity between SGAS and GAS, we considered the spatial distribution of hydrochemical variables in comparison to information is described in literature. Under potentiometric favorable conditions, the waters from the GAS ascend through geological structures (open fault planes) to the SGAS modifying the typical hydrochemical signature of the aquifer (Nanni et al., 2009).

The component planes (Fig. 4) reveal interesting aspects of the training data that include correlation, dissimilarity, and grouping. Similarity in color patterns, such as Ca and Mg, indicate a strong positive correlation. This is an interesting asset for exploratory analysis, especially when supported by the correlation matrix (Table 2) calculated after the topological evaluation. Calcium and magnesium have a correlation of 0,949. Sodium, chloride and sulfate are all positively correlated according to the component plots. Sufate correlates 0,883 and 0,478 with sodium and chloride, respectively.

The high correlation between calcium and magnesium is justified by the fact that both are products from dissolution of basaltic rocks forming minerals. Typical waters from SGAS, with longer residence time, are calcic-bicarbonate or calcic-magnesium bicarbonate (Fraga, 1986). Sodium bicarbonate-rich waters differ in the composition to solutions formed due to leaching of the Serra Geral Formation basaltic rocks. The sodium content can be attributed to several sources, like the alteration of albite, input by diffusion loading of halite and mirabilite weathering from Guarani aquifer system (Sracek and Hirata, 2002). Therefore, anomalous concentrations of sodium bicarbonate on the SGAS waters may be related to GAS, indicating a connection between the two systems. The bicarbonate anion is the most abundant in both the SGAS and GAS. It is usually originated from the dissolution of carbon dioxide mutually present in the atmosphere and soil, reacting with percolating waters or from basalt silicates hydrolysis. Thus, low bicarbonate concentrations are usually linked to recent recharged waters or with water with a long time of residence (from silicate weathering) (Bittencourt, 1978).

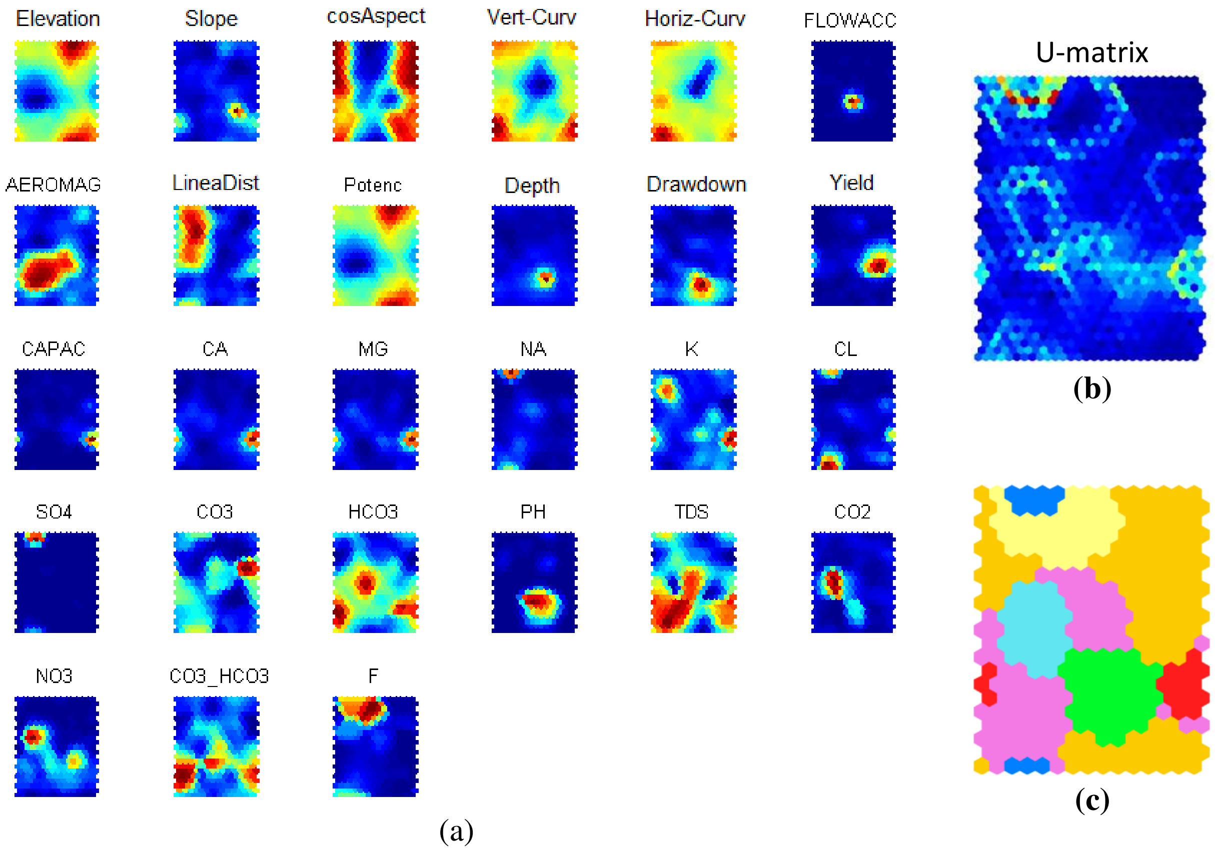


Fig. 4 – Component planes from SOM used to visualize nonlinear correlation. All variables were standardized using z-score, (b) U-Matrix. (c) U-Matrix classified using k-means technique.

Component planes for pH, carbonate and bicarbonate show partial inter-correlation (Fig. 4). The carbonate-bicarbonate content is related to the pH solution. In neutral and weakly alkaline conditions, the presence of bicarbonate is higher than the carbonate. From a pH = 8.30, the concentration of carbonate increases gradually until it replaces the presence of bicarbonate (Mocellin, 2009). For the SGAS, pH alkaline values can be attributed to the influence of the associated GAS groundwater. This is because with increasing alkalinity and pH values there is a carbonate imbalance leading to calcium depletion, causing an increase in sodium concentration (Silva, 2007). Table 2 shows the correlation matrix among 27 variables. Significant correlations (p-value = 0.01) are shown in bold. There is a noteworthy negative correlation between sodium, sulfate and fluoride concentrations and distance from the closer lineament. Areas near lineaments are associated with higher concentrations of these elements, strengthening the hypothesis of the structural conditions role in the rise of the waters from GAS. The same table shows a positive correlation between fluoride, sodium, chloride and sulfate; all considered typical elements of the GAS.

Table 2 – Correlation matrix calculated from SOM after the imputation process.

	Elevation	Potenc	Lineadist	Aeromag	Capac	Ca	Mg	Na	K	Cl	SO4	CO3	HCO3	pH	TDS	NO3	F
Elevation	1																
Potenc	**0.97**	1															
Lineadist	0.85	0.08	1														
Aeromag	**0.41**	**-0.43**	**0.12**	1													
Capac	0.03	-0.05	0.04	-0.04	1												
Ca	-0.05	**-0.11**	-0.10	-0.10	**0.37**	1											
Mg	-0.06	**-0.12**	-0.05	-0.07	**0.42**	**0.95**	1										
Na	**-0.15**	**-0.18**	**-0.30**	0.04	-0.05	0.03	-0.08	1									
K	0.11	**0.11**	**0.28**	**-0.18**	**0.15**	**0.45**	**0.45**	**-0.20**	1								
Cl	0.06	0.00	**-0.24**	**-0.12**	0.12	**0.51**	**0.40**	**0.47**	0.11	1							
SO4	-0.03	0.04	**-0.23**	**-0.13**	-0.04	0.07	-0.03	**0.88**	-0.07	**0.48**	1						
CO3	**-0.21**	-0.22	**-0.43**	0.05	0.05	**0.16**	0.08	**0.29**	-0.32	**0.34**	**0.20**	1					
HCO3	**-0.37**	**-0.43**	**-0.28**	**0.31**	0.02	**0.44**	**0.43**	**0.31**	-0.08	0.19	0.10	0.13	1				
pH	-0.04	-0.18	0.12	-0.02	**0.26**	0.10	**0.23**	-0.10	-0.03	-0.05	-0.06	-0.20	**0.22**	1			
TDS	**-0.18**	-0.27	-0.33	**0.26**	**0.13**	**0.44**	**0.35**	**0.55**	-0.07	**0.57**	**0.35**	0.36	**0.64**	0.02	1		
NO3	**-0.31**	-0.36	-0.01	**0.14**	**0.26**	**0.14**	**0.20**	0.06	-0.10	-0.04	0.01	-0.14	**0.38**	**0.48**	0.10	1	
F	**-0.13**	-0.13	**-0.38**	**-0.35**	-0.02	0.10	0.00	**0.43**	-0.12	**0.25**	**0.39**	**0.50**	0.09	-0.02	**0.16**	0.09	1

The color scales for the hydrochemical legends (Fig. 5) do not represent the maximum and minimum estimated values, but the histogram stretch, equivalent to two standard deviations from the mean. These descriptive statistical parameters are found in Table 3, for the original dataset and for the data imputated from the SOM. The purpose of generating continuous surfaces by a simple interpolation method is to analyze the spatial behavior of the imputation method results. Therefore, the interpolated surface should not be interpreted as estimation. In addition, the choice of a simple interpolation method like inverse distance weighting (IDW) allows a visualization of the influence from each sample on the continuous surface, especially of those anomalous values that may cause the bull's eye effect. Another assistance for a critical spatial analysis is different representations for samples with measured points and points where the values were imputated, illustrating how representatives the training points are.

Table 3 – Descriptive statistics for imputed values from SOM and for the JOINT original database composed by data gathered from Portela Filho (2003), Silva (2007) and Mocellin (2009).

SOM	Drawdown	Yield	Capac	Ca	Mg	Na	K	Cl	SO4	CO3	HCO3	pH	TDS	NO3	F
Average	32.27	23.86	3.96	16.17	4.43	18.53	0.77	3.40	5.76	19.96	74.20	7.74	150.96	1.05	0.15
Standard deviation	23.66	25.34	16.63	16.35	5.22	22.21	0.57	4.03	28.67	18.76	33.44	0.79	63.65	1.50	0.21
Min	0.00	5.39	0.32	1.47	0.17	0.59	0.19	0.47	0.48	2.14	13.79	6.04	32.66	0.03	0.01
Median	24.68	15.40	1.69	12.77	3.24	11.61	0.61	2.08	1.61	11.93	69.23	7.60	147.93	0.40	0.07
Max	209.94	185.11	316.84	139.50	46.25	293.83	4.37	29.21	546.40	87.91	176.27	9.89	386.46	8.01	1.54

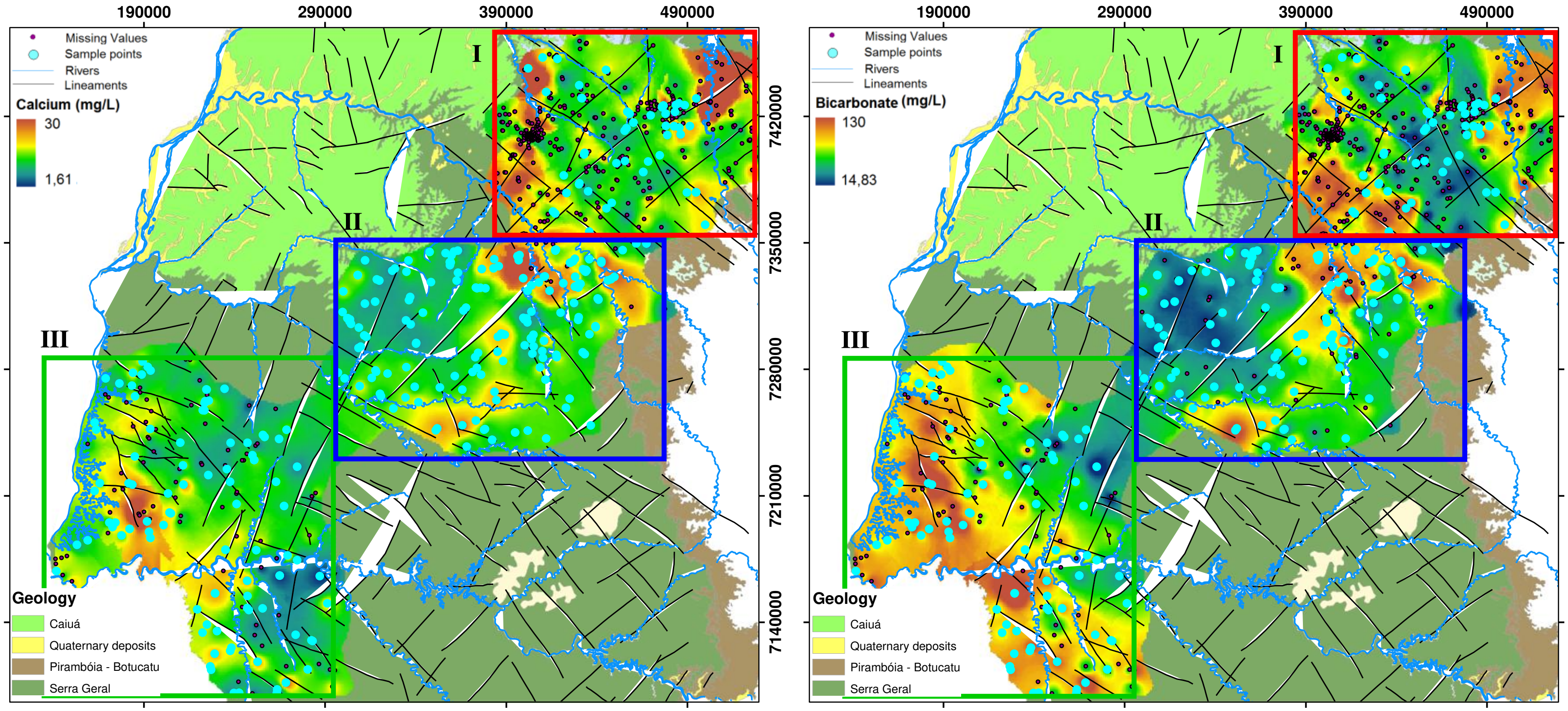
JOINT	Drawdown	Yield	Capac	Ca	Mg	Na	K	Cl	SO4	CO3	HCO3	pH	TDS	NO3	F
Average	28.64	21.37	3.44	16.12	4.55	18.29	0.74	3.24	6.14	28.23	76.24	7.64	137.02	1.14	0.19
Standard deviation	30.06	29.00	19.66	17.44	6.02	26.94	0.64	4.38	41.80	20.64	38.96	0.84	69.08	1.72	0.30
Min	0.00	1.00	-0.15	0.61	0.02	0.28	0.01	0.03	0.20	2.11	4.00	5.80	9.00	0.01	0.01
Median	20.00	10.00	0.69	13.00	3.48	9.90	0.60	1.87	1.00	26.73	71.65	7.50	121.50	0.26	0.10
Max	421.00	250.00	444.44	140.00	46.55	320.00	5.10	36.00	580.00	88.00	221.82	10.04	399.00	8.84	2.00

The calcium map (Fig. 5a) shows low concentrations in most part of subarea II and in the northeast and southeast sectors of subarea III (Fig. 1b). Regions with low levels of calcium may be related to a possible connection with GAS, due to the proportional decline of this element, or to recharge areas, especially if regions with low calcium content coincide with low levels of bicarbonate (Fig. 5b), which according to Fraga (1986), are associated with recent recharge areas. In contrast, high concentrations of calcium could indicate a long period of residence in exploited waters. Likewise, high levels of bicarbonate could either indicate a connection with the GAS or a confined section of the SGAS.

When considered together these two variables, the northeast of subarea II shows high values for calcium and bicarbonate, suggesting water confinement. The southern part of subarea III shows low values for calcium and high for bicarbonate, indicating an upward flow originated from the GAS to SGAS. To the southeast of subarea III, evidences for a hydraulic connection between the two aquifer systems are strengthened by the observation of high sodium and sulfate content (Fig. 6 a and b), taken as characteristic of GAS waters. The effusive rocks of the Parana basin are sulfide-poor, as in other forms of sulfur. Higher sulfur contents are attributed to contamination from underlying aquifers or mineralized, pyrite-rich intrusions (Bittencourt et al., 2003)

A typical feature of confined waters is high content of total dissolved solids (TDS), caused by a long period of residence, where in SGAS waters may be another sign of connection between the two systems (Fraga, 1986). The total dissolved solids comprises the sum of all the present mineral constituents in solution, having a direct relationship with the mineralogical rock composition and the time of groundwater percolation/residence within the system, thus reflecting the chemical weathering of rock forming minerals. The northern section of subarea II and the great majority of subarea III but its northeastern portion (Fig. 7a) display high levels of TDS, as previously mentioned, potentially indicating aquifer's connectivity or confinement. The low TDS content, plus the low levels of calcium and bicarbonate observed at the central sector of subarea II reinforces the hypothesis of areas with recent recharge. These same areas have lower concentrations of fluoride, which has a positive correlation with sodium, sulfate and chloride, characteristic elements found in the GAS.

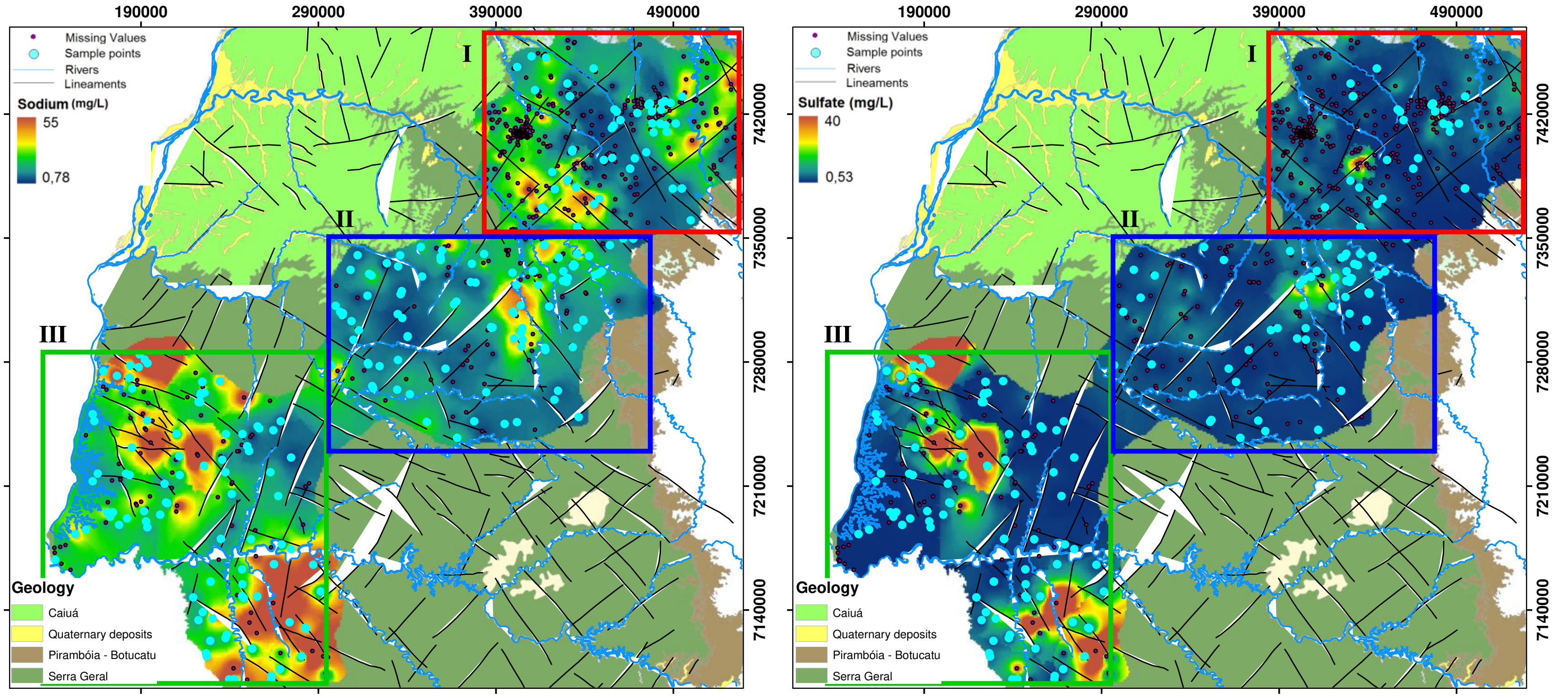
For Fraga (1986), the presence of fluoride in the SGAS is associated with upward flow of alkaline waters from the GAS. In contrast, Nanni et al. (2009) argued that the origin of fluorine yet needs further investigation because they could be the result of the SGAS secondary mineral weathering. The high concentrations of fluoride in the northern portion of subarea I is spatially coincident with high anomalous values found in surface waters (Licht, 2001), while the fluoride content in groundwater is attributed to deep geological structures, features that can be captured by aeromagnetic survey. Fluoride transported by upward flows from the GAS is connected to fracture density, thereby facilitating the flow from the aquifer. High fluoride values could be enhanced by superficial processes such as, the presence of a thick layer of soil and a high proportion of clay, factors that could prevent the recharge of the SGAS (Nanni, et al. 2009).



(a) Calcium

(b) Bicarbonate

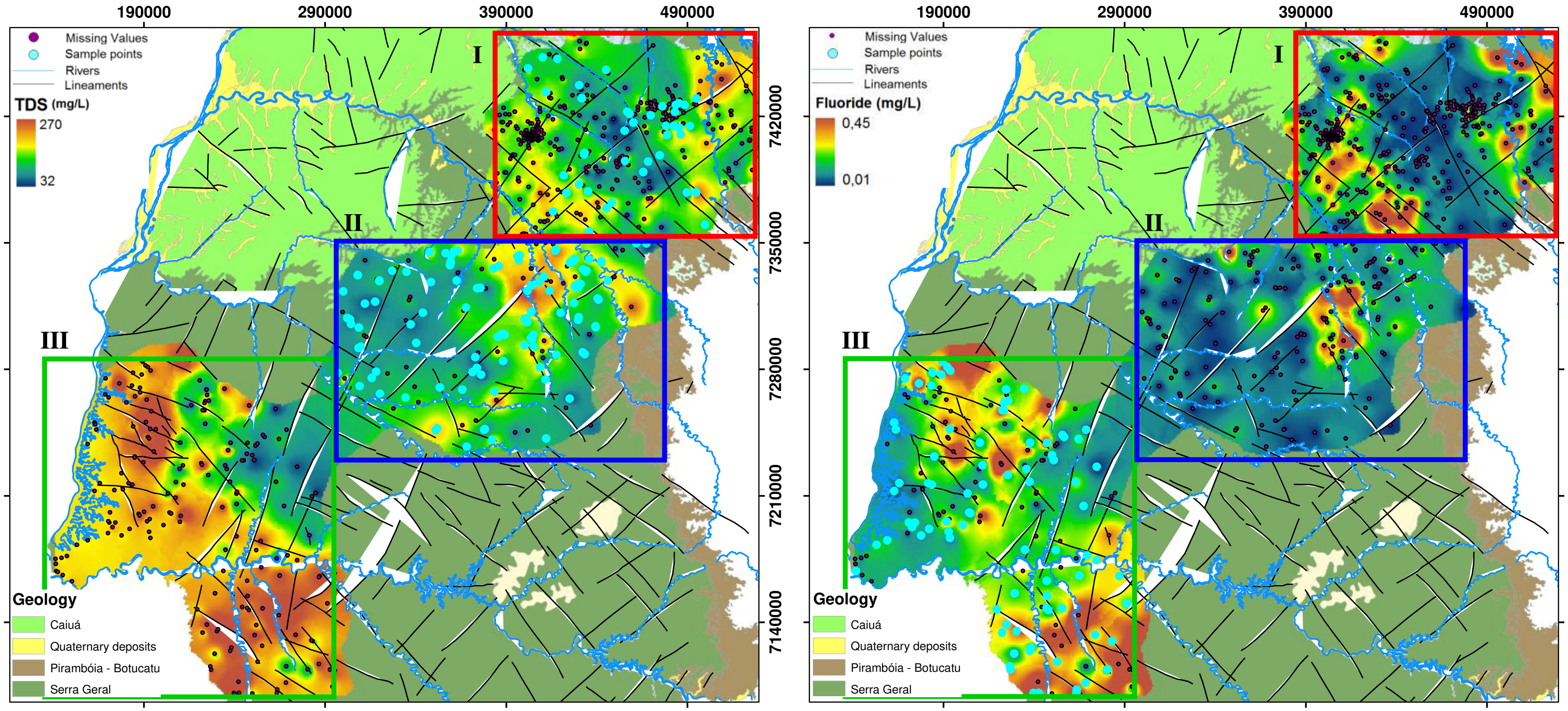
Fig. 5 – Imputed values and continuous surfaces calculated by the inverse distance weighting (IDW) method. (a) Calcium (mg/l) and (b) Bicarbonate (mg/l).



(a) Sodium

(b) Sulfate

Fig. 6 – Imputed values and continuous surfaces calculated by the inverse distance weighting (IDW) method. (a) Sodium (mg/l) and (b) Sulfate (mg/l).



(a) Total dissolved solids

(b) Fluoride

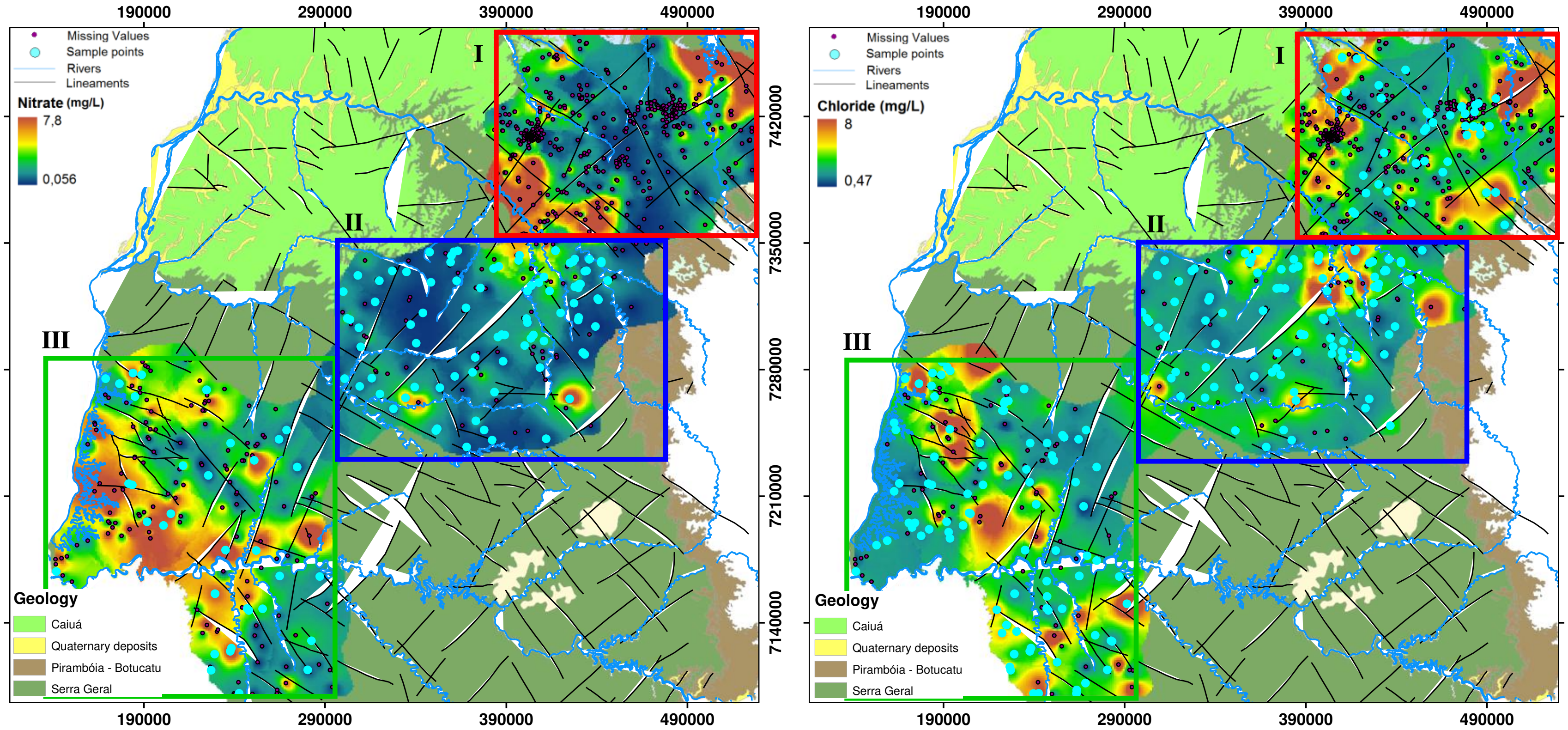
Fig. 7 – Imputed values and continuous surfaces calculated by the inverse distance weighting (IDW) method. (a) Total dissolved solids (mg/l) and (b) Fluoride (mg/l).

The land use and physical features can influence the chemical characteristics of recharging waters. Nitrate (Fig. 8a) usually has low concentrations in natural waters. High concentrations of nitrates in wells may result from direct infiltration of surface water or from polluted water into the free aquifer through the soil. Nitrate has a high spatial variability. In many groundwater systems nitrate is unlikely to have a relation with geological formations. Natural waters may contain large quantities of nitrate without causing serious health problems, but levels exceeding 5.00 mg/L represent an indicator of possible contamination by animal wastes or fertilizers (Rebouças and Fraga, 1988). The high nitrate content observed in the northeast and southwest sectors of subarea I is probably caused by surface contamination, given the proximity to two urban centers (Londrina and Maringá). Furthermore the area is intensely cultivated with cotton, coffee and soy (Licht, 2001).

An alternative to trace surface contamination is the chloride content (Fig. 8b). Chloride is taken as a highly mobile ion through most aquifer systems. Its source can be either antropic or natural. In the GAS, the chloride is likely originated from evaporitic rocks and from weathering of micas present in the Pirambóia and Botucatu Formations (Gastmans et al., 2010). In the SGAS, the chloride reflects surface intakes, GAS upward flows and weathering of basalt secondary minerals, like chlorite. The chloride non-reactive characteristics make hydraulic properties as support variables to analyze its spatial content variation.

The specific capacity (Fig.9a) refers to how much the water level decreases as a function of a given yield rate, i.e., it describes the aquifer's capacity for water supply and storage. In a fractured aquifer, the specific capacity is related to the density of structural discontinuities. The area denominated as the northern Serra Geral aquifer (Fraga, 1986) exhibits low specific capacity levels, whereas the southern Serra Geral aquifer shows higher values in areas close to central parts of the Parana basin, likely related to preferential flows caused by potentiometric gradients.

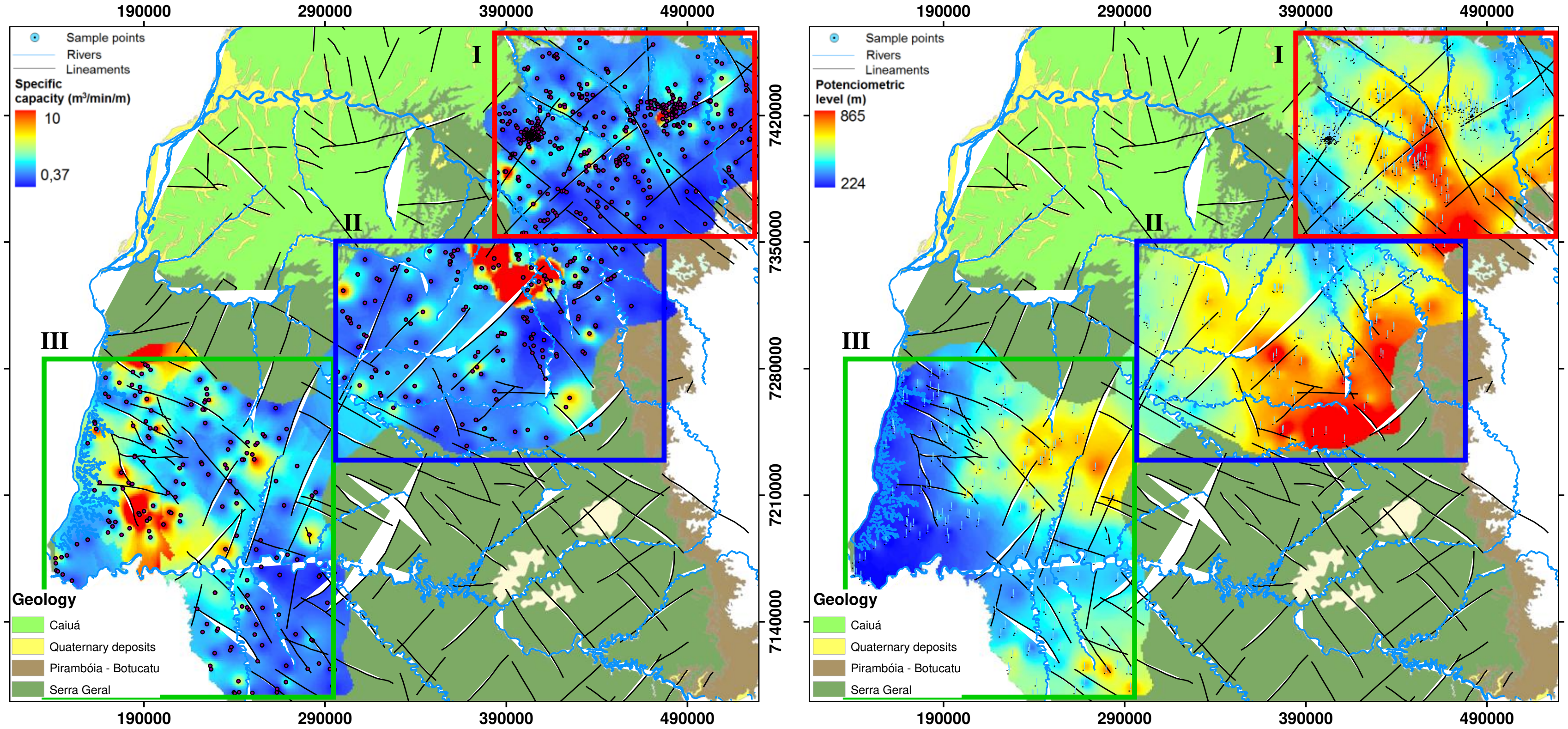
The potentiometric level (Fig. 9b) can be use to describe the hydrogeological flow direction preferences in isotropic aquifers, like the GAS, which is a porous aquifer hosted by the Botucatu and Pirambóia Formations. However, even considering the anisotropic feature of the SGAS, at a regional scale, low potentiometric level could indicate the path taken by the groundwater. The subarea III, having a low potentiometric level, indicates a potential recharge from the GAS when the other trait elements (Fig. 6a, 6b, 7a and 7b) concentrations are also considered.



(a) Nitrate

(b) Chloride

Fig. 8 – Imputed values and continuous surfaces calculated by the inverse distance weighting (IDW) method. (a) Nitrate (mg/l) and (b) Chloride (mg/l).



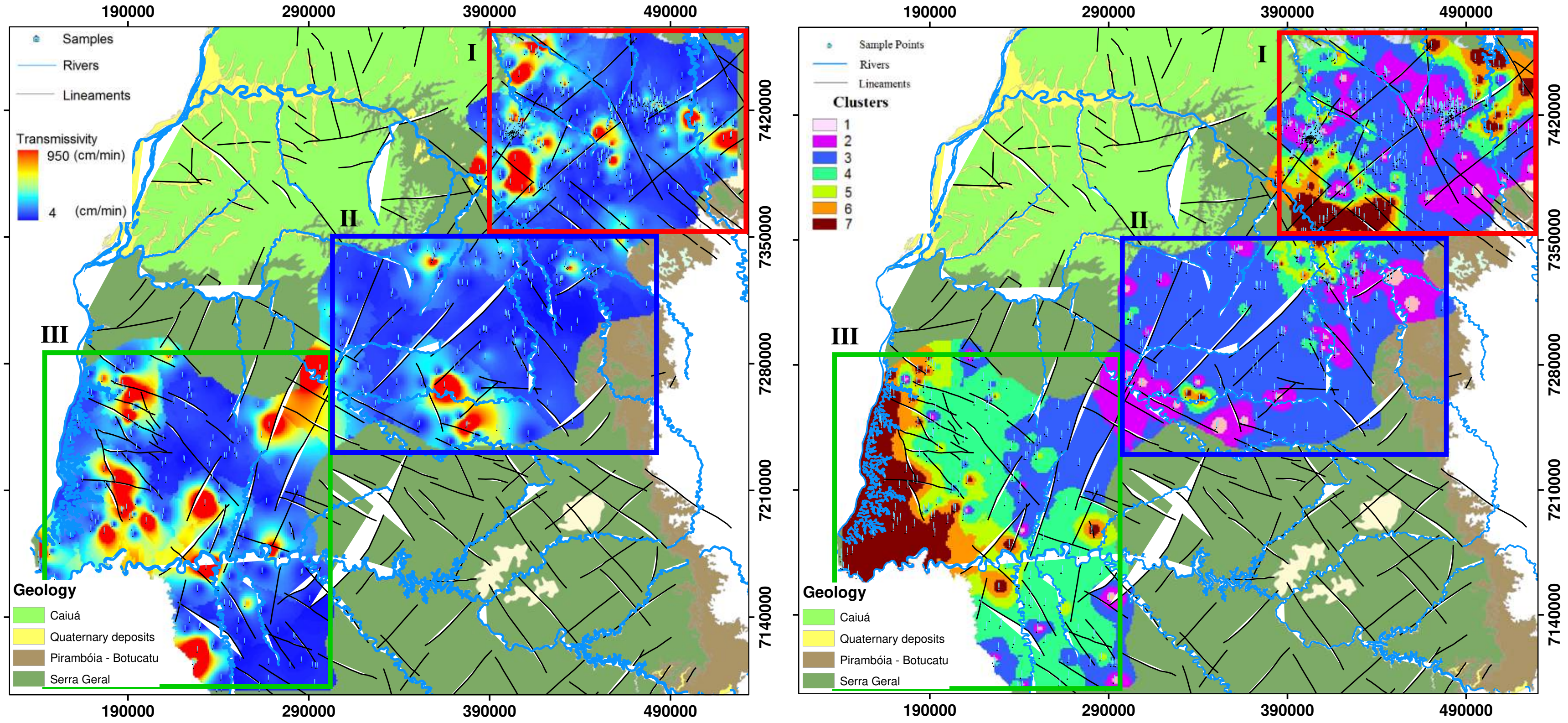
(a) Specific Capacity

(b) Potenciometric level

Fig. 9 – Imputed values and continuous surfaces calculated by the inverse distance weighting (IDW) method. (a) Specific capacity (m^3/h/m) and (b) Potenciometric level (m).

The hydraulic transmissivity (Fig. 10a) is a measure that describes the average speed at which water flows horizontally through the aquifer. Its variability depends on several factors, including potentiometric level and negative pressure. The horizontal movement of water is conditioned by the presence of discontinuities caused by the horizontal heterogeneity originated from a series of overlapping outflows (Rebouças and Fraga, 1988). The transmissivity map (Fig. 10a) represents only imputated values. The training dataset is located in the northeastern portion of subarea I, where data variability is high and correlations are negative, These are suitable conditions for SOM training to prevent overfitting (Rallo et al., 2002). The calculated hydraulic transmissivity supports hydrochemical analysis and is a parameter for numerical modeling of groundwater flow and solute transport.

The cluster map (Fig. 10b) summarizes the U-matrix values, which were classified using the k-means technique and the numbers of clusters were determined through Davies-Bouldin criteria. The U-matrix is a bi-dimensional representation for dissimilarities of n-dimensional code vectors (Fraser and Dickson, 2007). Subarea II is dominated by cluster three (Fig. 10b), which also occurs in subareas I and III. Cluster three, particularly if analyzed considering the spatial distribution of the elements, characterizes areas with recent recharge. It also displays a signature of a less weathered rock, expressed by its hydraulic and hydrochemical properties. Subarea I is greatly influenced by the density and magnitude of vertical structures, likely responsible for a diversity of clusters in this subarea. Cluster one describes outliers in general, whereas clusters five and six correspond to transitions between more distinct groups. Clusters four and seven comprise areas with potential hydraulic connectivity between the aquifers, whereas clusters two and three depict typical SGAS waters. The hydrochemical spatial model highlights areas with potential connectivity between the SGAS and the GAS (Fig. 11).



(a) Hydraulic Transmissivity

(b) Clusters

Fig. 10 – Imputed values and continuous surfaces calculated by the inverse distance weighting (IDW) method. (a) Hydraulic transmissivity (cm/min) and (b) Clusters.

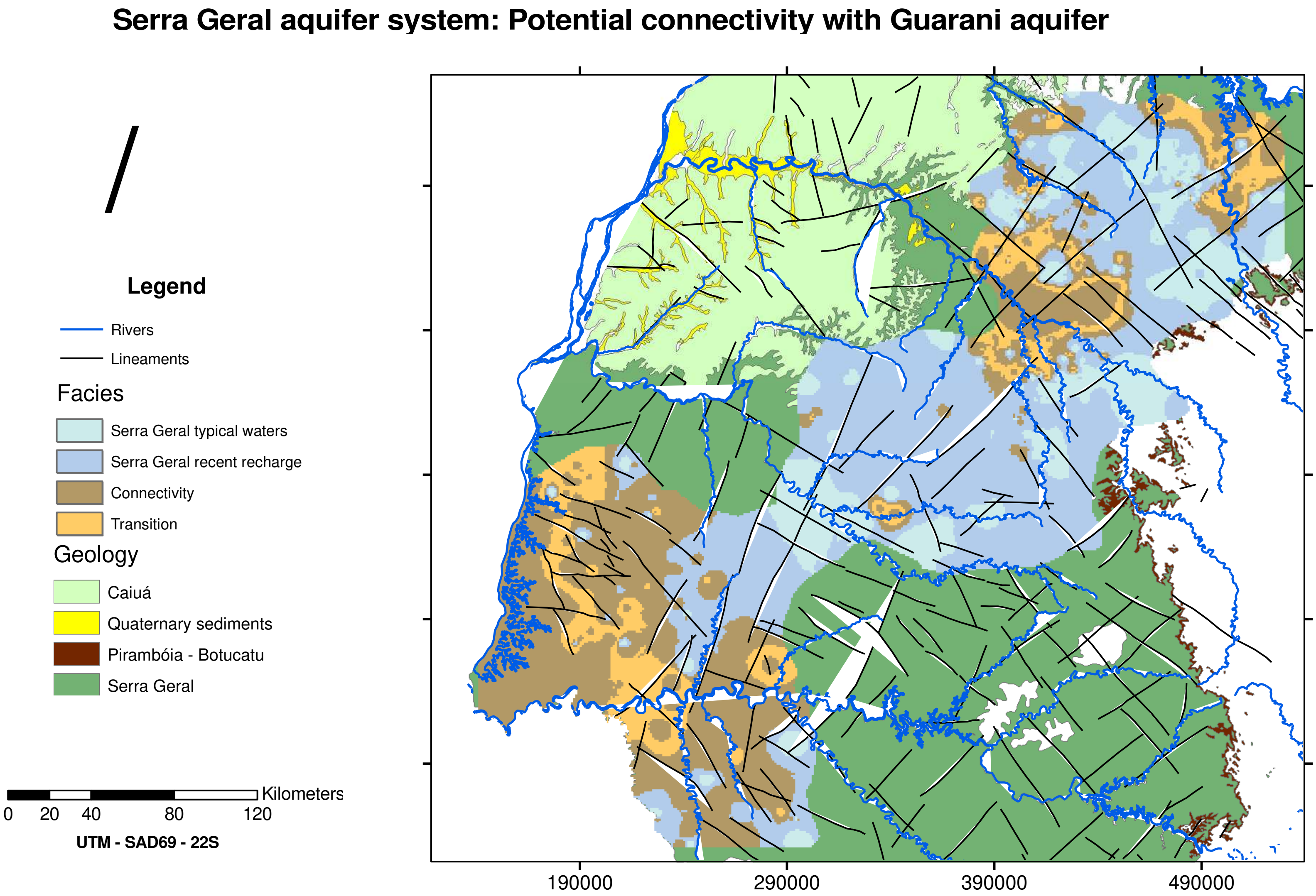


Fig. 11 – Map summarizing the hydrochemical spatial model.

6. Conclusions

A hydrochemical conceptual model from the Serra Geral fractured aquifer was built based on the spatial variation of chemical elements and compared with models previous established in the literature. Using self-organizing maps imputation, missing values for hydrochemistry and hydraulic transmissivity were estimated.

Considering the objectives of this work, important remarks are as follows:

(1) the SOMs were able to calculate the correlation matrix, determining nonlinear correlations between hydrochemical elements and explanatory variables. SOM imputation preserved the hydrochemical correlations described in the literature, as well as the statistical parameters (Table 3), showing no instability when predicting extreme values.

(2) the k-means clustering technique classified the variables based on their topological similarity, where the groups reflect (i) hydrochemical typical facies for recent recharge areas; (ii) potential connectivity between GAS e SGAS; (iii) regions featuring transition; (iv) water with confinement and log residence traits; and (v) typical waters from SGAS.

(3) the method proposed here to predict hydraulic transmissivity combining pumping tests, iterative equation and imputation of data was adequate to cope with incomplete information from wells database, proving to be an important advantage when conceiving numerical simulations.

(4) analysis of the spatial distributions of chemical elements and clusters maps has shown regions with potential flux connections between the Serra Geral and Guarani aquifer systems.

(5) The proposed method is suitable to survey hydrochemistry and groundwater physical properties, revealing and quantifying relationships in a large set of variables, which would not be possible to observe using parametric, multivariate statistical approaches.

We expect the proposed modeling method to be used as an alternative approach for further studies analyzing hydrochemistry and producing input parameters for numerical modeling of the Serra Geral and Guarani aquifers.

Acknowledgements

We are grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and Conselho Nacional de Desenvolvimento Científico e Tecnológico for their financial support; to Victor F. Labson, Director, Crustal Imaging and Geochemistry Science Center (CIGSC), U.S.

Geological Survey (USGS), Denver, Colorado, United States of America for providing the first author with the position of visiting scientist; to Instituto de Águas do Paraná and Companhia de Saneamento do Paraná for kindly providing the data.

References

Athayde, G. B., Müller, C. V., Rosa Filho, E. F., Hindi, E. C., 2007. Estudo sobre os tipos das águas do aquífero Serra Geral, no município de Marechal Cândido Rondon – PR. Águas Subterrâneas 21, p.111-122.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000a. Artificial neural networks in hydrology. I: preliminary concepts. Journal of Hydrologic Engineering 5, 115-123.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000b. Artificial neural networks in hydrology. II: hydrologic applications. Journal of Hydrologic Engineering 5, 124-137.

Astel, A., Tsakoviski, S., Barbieri, P., Simeinov, V., 2007. Comparison of self- organizing maps classification approach with cluster and principal components analysis for large environmental datasets. Water Research 41, 4566 – 4578.

Briggs, I. C., 1974. Machine contouring using minimum curvature. Geophysics 39, 39 – 48.

Benavides, A., Everett, M. E., Pierce Jr., C., 2009. Unexploded ordnance discrimination using time-domain electromagnetic induction and self-organizing maps. Stochastic Environmental Research and Risk Assessment 23, 169–179.

Bierlein, F.P., Fraser, S.J., Brown, W.M., Lees, T., 2008. Advanced methodologies for the analysis of databases of mineral deposits and major faults. Australian Journal of Earth Sciences 55, 79–99.

Bittencourt, A. V. L., Rosa Filho, E. F., Hindi, E. C., Buchman Filho, A. C., 2003. A influência dos basaltos e de misturas com águas de aqüíferos sotopostos nas águas subterrâneas do Sistema Aqüífero Serra-Geral na bacia do rio Piquiri, Paraná – BR. Revista Águas Subterrâneas 17. www.abas.org.br

Bradbury, K. R., Rothschild, E. R., 1985. A computerized technique for estimating the hydraulic conductivity of aquifers from specific capacity data. Ground Water 23, 240–246.

Davies, D.L., Bouldin, D.W., 1977. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227.

Dickson, B. L., Giblin, A., 2007. An evaluation of methods for imputation of missing trace element data in groundwaters. Geochemistry: Exploration, Environment, Analysis 7, 173–178.

Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. American Geophysical Union EOS 81, 583–585.

Ferreira, F. J. F., Portela Filho, C. V., Rosa Filho, E. F., Rostirolla, S. P., 2005. Conectividade e compartimentação dos sistemas aqüíferos Serra Geral e Guarani na região central do arco Ponta Grossa (Bacia do Paraná, Brasil). Revista Latino Americana de Hidrogeologia 5, 61-74.

Fessant, F., Midenet, S., 2002. Self-organising map for data imputation and correction in -surveys. Neural Computing & Applications 10, 300–310.

Fraga, C. G., 1986. Introdução ao Zoneamento do Sistema Aqüífero Serra Geral no Estado do Paraná. São Paulo, Master's thesis. Instituto de Geociências, Universidade de São Paulo (USP), São Paulo, Brazil. 125p.

Fraser, S.J., Dickson, B.L., 2007. A new method for data integration and integrated data interpretation: self-organizing maps. Exploration 07: Fifth Decennial International Conference on Mineral Exploration, pp. 907–910.

Friedel, M.J., Iwashita, F., *In press*, Nonlinear modeling of autocorrelated random variables for application to inverse problems, Environmental Modelling & Software.

Gastmans, D., Chang, H. K., Hutcheon, I., 2010. Groundwater geochemical evolution in the northern portion of the Guarani Aquifer System (Brazil) and its relationship to diagenetic features. Applied Gochemistry 25, 16-33.

Harris, C., Milner, S., 1997. Crustal origin for the Paraná rhyolites: Discussion of 'Description and petrogenesis of the Paraná Rhyolites, southern Brazil' by Garland et al (1995). Journal of Petrology 38, 299-302.

Hong, Y., Rosen, M. R., 2001. Intelligent characterization and diagnosis of the groundwater quality in a urban fractured-rock aquifer using an artifitial neural network. Urban Water 3, 193-204.

Iwashita, F., Friedel, M. J., Souza Filho, C. R., Fraser, S. J., *in press*. Hillslope chemical weathering across Paraná state, Brazil: A data mining–GIS hybrid approach, Geomorphology, 26p. http://dx.doi.org/10.1016/j.geomorph.2011.05.006

James, A. L., McDonnell, J. J., Meerveld, I. T., Peters, N. E., 2010. Gypsies in the palace: experimentalist's view on the use of 3-D physics-based simulation of hillslope hydrological response. Hydrological Processes 24, 3878-3893.

Jenson S. K. and J. O. Domingue., 1988. Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis. Photogrammetric Engineering and Remote Sensing 54, 1593–1600.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. & Kolehmainen, M., 2004. Methods for imputation of missing data in air quality data sets. Atmospheric Environment 38, p. 2895–2907.

Kalteh, A.M., Hjorth, P., 2009. Imputation of missing values in precipitation-runoff process database. Hydrology Research 40, 420-432.

Kohonen, T., 2001. Self-organizing Maps, third edition, Springer-Verlag, Berlin.

Licht, O. A. B. 2001. A geoquímica multielementar na gestão ambiental. PhD Thesis, Faculdade de Geologia, Universidade Federal do Paraná, Brazil.

López García, H., Machón González, I., 2004. Self-organizing map and clustering for wastewater treatment monitoring. Engineering Applications of Artificial Intelligence 17, 215–225.

Lu, R., Lo, S., 2002. Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. Water Research 36, 2265-2274.

Malek, M.A., Harun, S., Shamsuddin, S.M., Mohamad I., 2008. Imputation of time series data via Kohonen self organizing maps in the presence of missing data. Engineering and Technology 41, 501-506.

Manassés, F., 2009. Caracterização hidroquímica da água subterrânea da formação Serra Geral na região sudoeste do estado do Paraná. Master's thesis. Universidade Federal do Paraná. Curitiba, Brazil. 136p.

McLin, S. G., 2005. Estimating Aquifer Transmissivity from Specific Capacity Using MATLAB. Ground Water 43, 611 – 614.

Minty, B. R. S., 1991. Enhancement and presentation of airborne geophysical data. AGSO Jounal 17, 63 – 75.

Mocellin, R. C., 2009. Conectividade e compartimentação magnética-estrutural dos sistemas aquíferos serra geral e guarani na região sudoeste do Estado do Parana (Bacia do Paraná, Brasil). Master's thesis. Universidade Federal do Paraná. Curitiba, Brazil. 231p.

Nanni, A., Roisenberg, A., Marimom, M. P. C., Viero, A. P., 2009. The hydrochemical facies and anomalous fluoride content in Serra Geral aquifer system, southern Brazil: a GIS approach with tectonic and principal component analysis. Environmental Geology 58, 1247-1255.

Neter, J., Kutner, M. N., Nachtssheim, C. J., Wasserman, W. 1996. Applied linear statistical models, 4th Ed. WCB/McGraw-Hill, Boston.

Peate, D. W., Mantovani, M. S. M., Hawkesworth, C. J., 1988. Geochemical stratigraphy of Paraná continental flood basalts: borehole evidence. Revista Brasileira de Geociências 18, 212-221.

Penn, B. S. 2005. Using self-organizing maps to visualize high-dimensional data. Computer & Geosciences 31, 531-544.

Portela Filho, C. V., 2003. Condicionamento estrutural-magnético do sistema aqüífero Serra Geral na região central do arco de Ponta Grossa (Bacia do Paraná) e sua conectividade com o sistema aqüífero Guarani, Master's thesis. Universidade Federal do Paraná. Curitiba, Brazil. 163p.

Rabelo, J. L., Wedland, E., 2009. Assessment of groundwater recharge and water fluxes of the Guarani aquifer systems, Brazil, Hydrogeology Journal 17, 1733-1748.

Rallo, R., Ferre-Gine, J., Arenas, A., Giralt, F., 2002. Neural virtual sensor for the inferential prediction of product quality form process variables. Computers and Chemical Engineering 26, 1735-54.

Rebouças, A. C., Fraga, C. G., 1988. Hidrogeologia das rochas vulcânicas do Brasil. Revista Águas Subterrâneas 12, 29-55.

Rosa Filho, E. F., Salamuni, R., Bittencourt, A.V.L., 1987. Contribuição ao estudo das águas subterrâneas nos basaltos do estado do Paraná. Boletim Paranaense de Geociências 37, 22-52.

Rosa Filho, E. F., Bittencourt, A. V. L., Hindi, E. C., Bittencourt, A., 2006. Groundwater types and structural conditioning study of the Guarani aquifer system in the western of Paraná state (Brazil). Águas Subterrâneas 20, p.39-48.

Sánchez-Martos, F., Aguilera, P. A., Garrido-Frenich, A., Torres, J. A., Pulido-Bosch, A., 2002. Assessment of groundwater quality by means of self-organizing maps: Application in a semi-arid area. Environmental Management 30, 716-726.

Silva, A. B., 2007. Conectividade e compartimentação magnéticaestrutural dos sistemas aqüíferos Serra Geral e Guarani na região central do estado do Paraná. Master's thesis. Universidade Federal do Paraná. Curitiba, Brazil. 182p.

Souza Filho, O. A., Silva, A. M., Remacre, A. Z., Sancevero, S. S., McCafferty, A. E., Perrotta, M. M., 2010. Using helicopter electromagnetic data to predict groundwater quality in fractured crystalline bedrock in a semi-arid region, Northeast Brazil. Hydrogeology Journal 18, 905-916.

Sracek, O, Hirata, R., 2002. Geochemical and stable isotopic evolution of the Guarani aquifer system in the state of São Paulo, Brazil. Hydrogeology Journal 10, 643-655.

Suk, H., Lee., K. K., 1999. Characterization of a groundwater hydrochemical system through multivariate analysis: clustering into groundwater zones. Ground Water 37, 358-366.

Thiede, D. S., Vasconcelos, P. M,. 2010. Parana flood basalts: Rapid extrusion hypothesis confirmed by new 40Ar/39Ar results. Geology 38, 747-750.

Turner, S. P., Peate, D. W., Hawkesworth, C. J., Mantovani, M. S. M., 1999. Chemical stratigraphy of the Paraná bsalt sucession in western Uruguay: further evidence for the diachronous nature of the Paraná magma types. Journal of Geodynamics 28, 459-469.

Valeriano, M. M., Kuplich, T. M., Storino, M., Amaral, B., Mendes, J. N., Lima, D. J., 2006. Modeling small watersheds in Brazilian Amazonia with shuttle radar topographic mission-90 m data. Computer & Geosciences 32, 1169-1181.

Valeriano, M. M., Rosetti, D. F., Albuquerque, P. C. G., 2009. TOPODATA: desenvolvimento da primeira versão do banco de dados geomorfométricos locais em cobertura nacional. XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, pp. 5499-5506.

Wang, S., 2003. Application of Self-organising maps for data mining with incomplete data sets. Neural Computing and Applications 12, 42-48.

Wendland, E., Barreto, C., Gomes, L. H., 2007. Water balance in the Guarani aquifer outcrop zone based on hydrogeologic monitoring. Journal of Hydrology 342, 261-269.

Winter, C. L., Guadagnini, A., Nychka, D., Tartakovsky, D. M., 2006. Multivariate sensitivity analysis of saturated flow through simulated highly heterogeneous groundwater aquifers. Journal of Computational Physics 217, 166-175.

APPENDIX

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Author: Fabio Iwashita - fabio_iwashita@hotmail.com
%% Geosciences Institute -  Campinas University, Brazil
%% Crustal Geophysics and geochemistry Science Center - USGS, USA
%% March/15/2010
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function inTQs(arqdat)
% Function calculates transmissivity for a data vector

[Data Name Title]=readGEOEAS(arqdat);

[nline ncolumn]=size(Dados);
format short;
Tcol=1;
j=1;
for i=1:nline

  Q=Data(i,5);    conv=7.48;
  s=Data(i,7);    t=Data(i,6);
  L=Data(i,4);    S=Data(i,8);
  r=Data(i,3);    r=r/12;

  D=L;               err=0.000001;
  Tguess=1.0;

  a=2.948; b=-7.363; c=11.447; d=-4.675;
  G=(a+b*(L./D)+c*(L./D).^2+d*(L./D).^3);

  sp=((D-L)./L.*(log(D./r)-G));     %% sp= partial penetration factor,
  dimensionless
  test=1;

  Tcalc(i,j)=1440*Q*(log(2.25*Tguess*t/(1440*r^2*S))+2*sp(j))/(4*conv*pi*s);
  diff=abs(Tcalc(i,j)-Tguess); test=diff;
  while test>err
      calc(i,j)=1440*Q*(log(2.25*Tguess*t/(1440*r^2*S))+2*sp(j))/(4*conv*pi*s)
  ;
      diff=abs(Tcalc(i,j)-Tguess); Tguess=Tcalc(i,j); test=diff;
  end;
Ttot(i) = calc(i,j);
Tcol = Tcol+1;
end

nData=[Data Ttot]
csvwrite('transmis.dat', nData)
```

EVALUATING SRTM-90M INTERPOLATION USING SRTM-30 FROM U.S. TERRITORY

Abstract

The spatial resolution of SRTM (Shuttle Radar Topographic Mission) digital elevation models, currently available is set at 90m (or ~3”) and ~30m (or ~1”) for the USA territory. Refining the 90m grid through geostatistic methods has been an approach adopted by several users. However, models based on semivariograms generally exhibit distinct parameters for each sampled area. These particularities raise questions over the application of the same model throughout larger scales. The assessment of the interpolation effectiveness is another research topic of interest. This paper presents a methodology to measure the strength of SRTM data interpolation from 90m to 30m, and the feasibility to apply a single variogram model to larger areas. The study region lies near the Rocky Mountains in Montana State, USA. Initially, the SRTM was resampled from 30m to 90m, and then kriged to 30m. This interpolated data was compared with the 30m original grid through map algebra. The results from layers subtraction were evaluated with descriptive statistics and linear regression, and hypothesis test for β_1=1 e β_0=0. The regression residuals shows a submetric mean and the regression test accept the null hypothesis for both tests. These outcomes support the adoption of the kriging method for interpolation of SRTM-90m and the use of the same model adjusted for a sampled area to larger regions.

Keywords: digital elevation model, interpolation, geostatistics, kriging

1. INTRODUCTION

Digital elevation models (DEM) are fundamental sources to extract relief information over large areas or with difficult access. From elevation data it is possible to extract morphometric measures, slope and terrain aspect, that can be applied in a variety of studies, such as laminar erosion models (Araújo, 2006), meteorological models (Goovaerts, 2000), landslides risk assessment (Kääb, 2002), agricultural suitability mapping (Sommer et al., 1998), geological mapping (Demirkesen, 2008; Masoud and Koike, 2006) and forest structure mapping (Nelson et al., 2007). This work employs globally available data generated by the SRTM – Shuttle Radar Topographic Mission (http://edcsns17.cr.usgs.gov/NewEarthExplorer/), with a spatial resolution of 3 arcsec or ~90 meters (Farr and Kobrick, 2000), which can be improved by interpolation methods. For the United States territory, data are available with the prime, higher resolution, of 1 arcsec or ~ 30 m. The importance of a high spatial resolution lies in applications of the products derived from the DEM. The generation of grids with 30 m resolution by geostatistic-based interpolation methods has been adopted in some recent work (Rossetti and Valeriano, 2007; Araújo, 2006; Grohmann et al., 2007; Valeriano et al., 2006; Yun et al., 2005), where the evaluation of the data was based on descriptive statistics, terrain profiles and land morphometric characteristics, geostatistical analysis is essential to preserve terrain features (Valeriano et al. 2006).

The semivariogram depicts the degree of spatial dependence between samples over a specific support. For their construction, simply squared differences of pairs of values are obtained, assuming stationarity of the increments (Landim, 1998). The variogram measures the variability related to a distance. Such variability can be significantly different when considering different directions (Iwashita et al., 2005). A fundamental step in geostatistical analysis refers the determination of the variogram model, since it relies on parameters from the theoretical variogram model. Thus, for a reliable interpolation, there should be a good fit of the model to the experimental variogram, whereas different relief structures can lead to different models of variograms.

When working with large areas, there is an operational problem - the need to fit a large number of variogram models for each landscape segment. Besides, it is unclear whether the quality of interpolated SRTM data grids (30m) has sufficient details when compared to the original (30m from U.S. territory). It is also yet vaguely approached if a variogram model generated for a particular area is applicable for regions that are more extensive in the same SRTM data block. In this context, this work aims to demonstrate that is possible to validate a 90 m resolution SRTM dataset based on resampling of the original SRTM data at 30 m . Using a study area in the United States (where SRTM data is available at maximum, 30m resolution), the goals of this work are: (a) to refine a resampled SRTM 90 m grid to 30 m-using ordinary kriging, (b) to verify the applicability of variogram model with distinct parameters generated from different regions within the study area, and (c) to evaluate the results of the interpolation, through geoprocessing operations, descriptive statistics and linear regression.

2. METHODOLOGY

The study area is in the State of Montana, United States (Figure 1) - the fourth largest State in country, but one of the least populated, with 902,194 inhabitants. Montana displays a poor vegetation covering and a low population density (2.39/sq mi). This facilitates morphometric characterization, since cities and forest structures may change terrain features captured by the SRTM instruments. Considering the relief, Montana can be classified into two major regions: the Eastern region is dominated by the Great Plains and the West region is dominated by the Rocky Mountains.

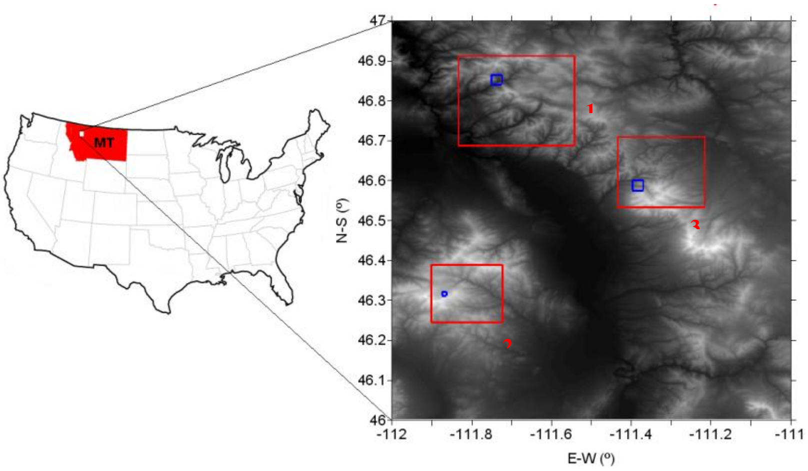


Figure 1 – Locality map of the study area. Sampling areas in blue and test areas in red.

The data processing firstly consisted of SRTM data resampling from 30 m to 90 m resolution (Figures 2). In order to include topographic diversity and variability, three sampling areas were employed for the construction of the semivariograms and three other areas were used to implement the ordinary kriging. Each sample area has approximately 900 points and areas of implementation from 250 to 400 thousand points. Considering these samples at 90 m spatial resolution, first order trend analysis was applied and the function residues extracted and were then used to build semivariograms.

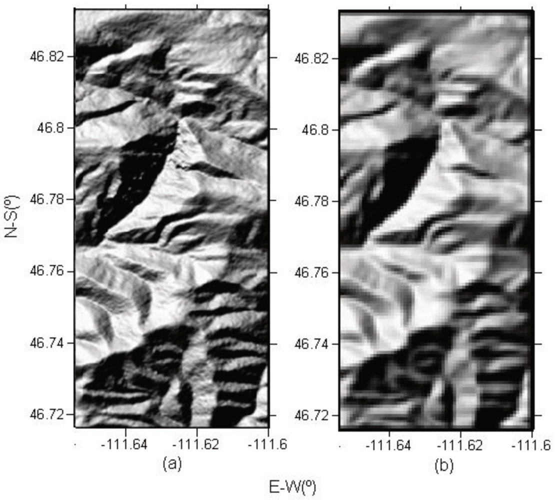
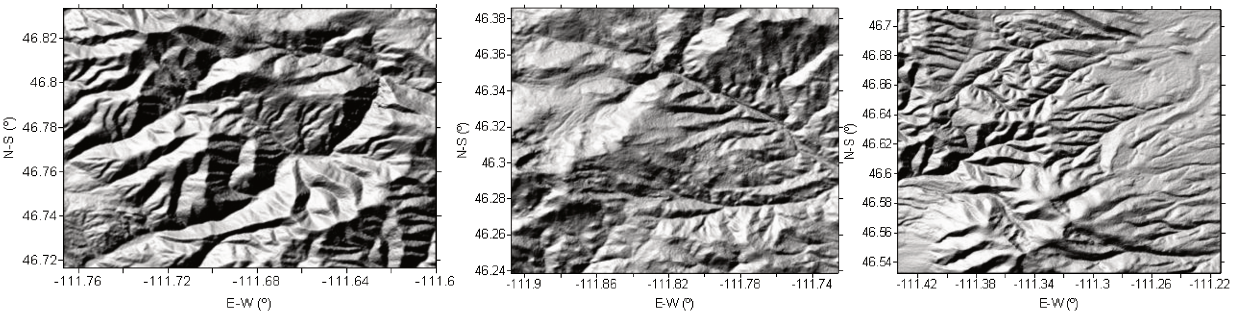


Figure 2 – Detail of Area 1with (a) 30 meters resolution (original grid) and; (b) 90 meters resolution (resampled grid).



(a) (b) (c)

Figure 3 – Areas 1 (a), 2 (b) and 3 (c) with 30 m resolution (original grid).

Valeriano (2004) conducted analysis of the SRTM data using ordinary kriging to generate digital elevation models (DEM) with 30 m spatial resolution for the Brazilian territory. Valeriano (2004) also designed a procedure for standardization and minimization of subjectivity in the process conceived to refine of the SRTM grids. The present work proposes a methodology to validate the interpolated data by subtraction between the original layers and that refined by ordinary kriging (Figure 3).

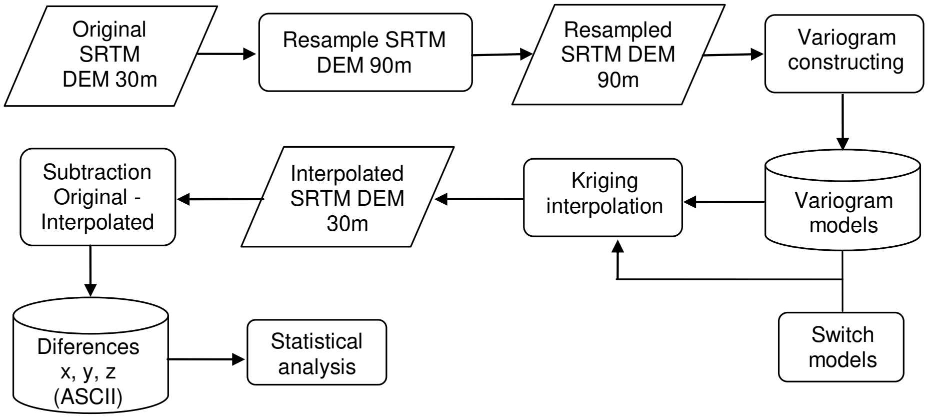


Figure 4 – Proposed method to evaluate surfaces generated through interpolation.

We highlight the existence of two sets of results: (i) raw, 30m SRTM resolution surfaces compared to 30m surfaces yielded from kriged, 90m SRTM resolution data; and (ii) 30m surfaces yielded from kriged, 90m SRTM resolution data, but using distinct interpolation models and parameters (Figure 4).

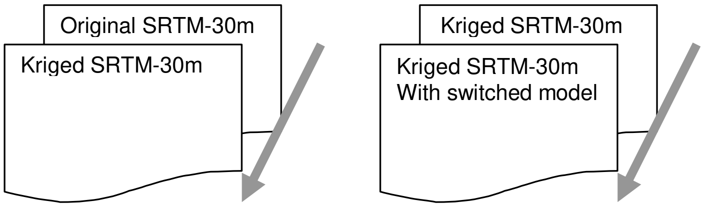


Figure 5 – Two different sets of results were evaluated: the surfaces compared to the original SRTM-30m dataset and the surfaces produced with switched models.

The evaluation of the interpolation procedures aims to verify that the model adopted, as well as their parameters, are suitable to represent the spatial variability of the variable (Isaaks and Srivastava, 1989). If the model adopted is appropriate, the residuals of the difference should have an average near zero and the relationship between the real and the calculated values should be linear. Based on this premise, the descriptive statistics of the residuals and the correlation between raw and interpolated data were here employed to evaluate the interpolation results. In summary, a good interpolation is one that generates results with minimal differences between the value taken as real and the value estimated by the interpolation. In this case, the ideal scenario is that where the differences between the elements of the surfaces are zero. Thus, the exploratory analysis provides a first idea of the quality of the procedure. Perfect matches would be an average of zero and minimum and maximum values the closest possible to zero.

Another approach used for evaluation of the results was the linear regression (Netter et al., 1996). Regressions were computed between the results of the interpolations and the original grid, as well as between the interpolated model and its permutations. When the results of other methods or models are statistically equal to an ideal model (model 1 applied to area 1, for example), or to the grid of 30 m, the slope of the regression has an inclination of 45^{o} and cross through the origin, i.e., hypothesis testing for $\beta_1 = 1$ and $\beta_0 = 0$ respectively.

3. RESULTS AND DISCUSSION

Models adjusted for the three selected areas are Gaussians (Figure 5) that contemplate, in essence, smooth variations at close neighborhood and high variations at medium and distant ranges. When executing the ordinary kriging, attention is required in choosing the search radius. It must be quite close to the value of the range of the variogram. This is because when the sill (which has a relationship with the statistical variance) is reached, the variogram displays only the random component. Models 1, 2 and 3 showed, respectively, ranges at 734m, 831m and 594m and sills at 26,000, 3200 and 12,500.

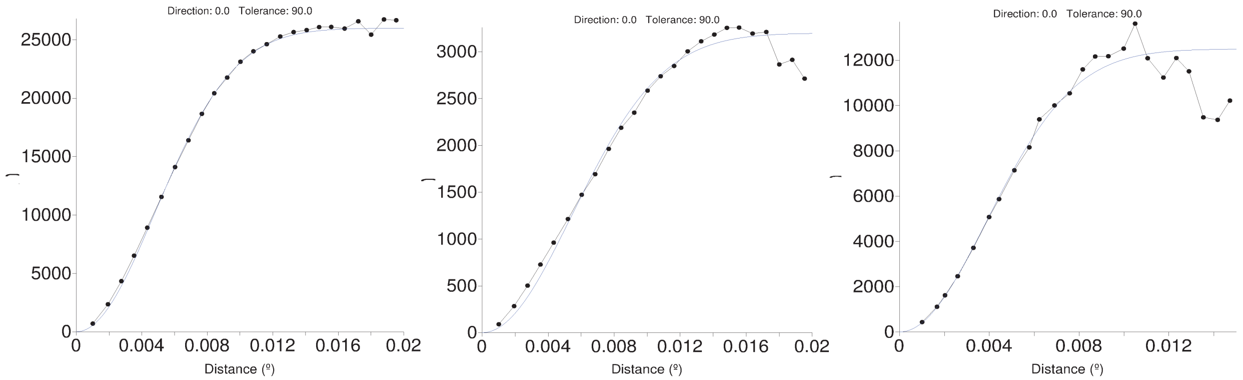


Figure 6 – Variograms for areas 1 (a), 2 (b) and 3 (c) (see Figures 1 and 2 for location)

The application of a single parameterized variogram model based on samples for a more regional scale has a twofold objective; to standardize and optimize the grid generation process and to preserve the morphometric characteristics and elevation values. In order to balance the quality of the interpolation with operational aspects, the semivariograms were adjusted considering an isotropic model. When comparing to the original 30m surfaces, the 30m surfaces produced by interpolation displayed smoothed features. This result is predictable, because ordinary kriging minimizes the variance via the Lagrange multiplier. Such smoothing affected the raw elevation data, because the interpolator works like a low-pass filter. Thus, it is important to quantify changes over the information. Another aspect is that the interpolation by window moving average causes loss of some points on the edge of the implementation area, depending on the size of the mask. These points should be disregarded during the statistical analysis of the difference between surfaces. Grids created using different models have similar aspects, making unfeasible a visual exploratory analysis (Figure 6). Then, map algebra operations are required to quantify these differences, like subtraction between maps.

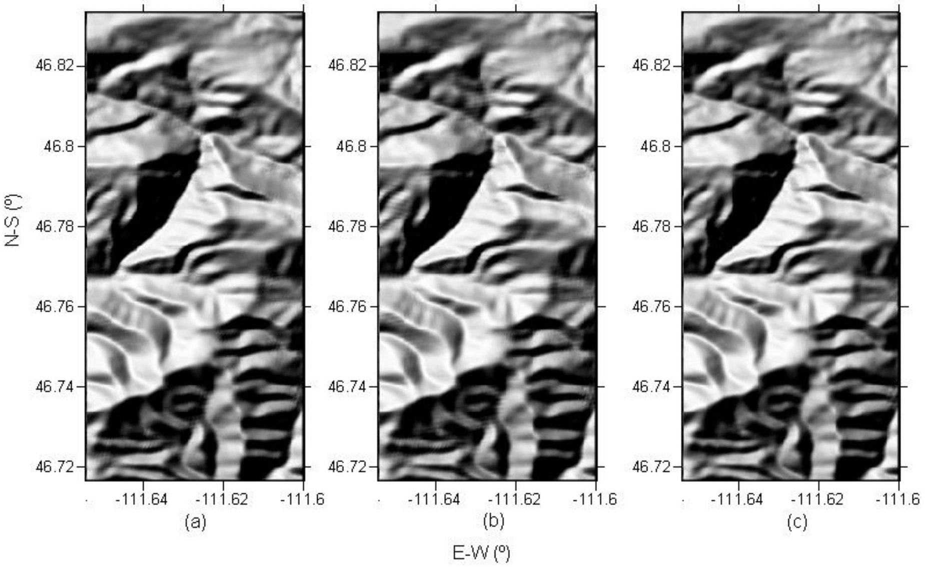


Figure 7 – Detail of area 1 with 30 m resolution, interpolated grid with model 1, (a), model 2 (b) and model 3 (c) (see Figure 5).

Table 1 presents statistics of the results for the subtraction between the raw surface areas with 30 m resolution without treatment with 30m resolution surfaces generated by kriging. All interpolated surfaces have sub-metric difference averages when compared to the original grid, which is towards an ideal evaluation scenario. The quartiles also provide a good perception of the refinement quality, because they are smaller than 5 m. The results produced by kriging with exchanged models for the same area showed minor differences (Table 2). The average differences between interpolations are close to zero and reach a maximum of 15 cm. The quartiles are sub-metric, reaching a maximum error of 2 m.

Table 1. Exploratory analysis of the difference between the raw SRTM dataset at 30m and interpolated (kriged) dataset at 30m.

	Area 1	Area 2	Area 3
Minimum	-52.08	-33.25	-50.34
Maximum	52.20	38.42	41.43
Average	0.51	0.38	0.56
Median	0.38	0.28	0.67
1º quartile	-4.29	-1.79	-2.23
3º quartile	4.96	2.45	3.38
Variance	48.61	13.02	25.52
Standard deviation	6.97	3.60	5.05

Table 2 – Exploratory analysis of the difference between interpolated data with switched models. Caption: A*x*M*y* – Area x interpolated with model y.

	A1M2	A1M3	A2M1	A2M3	A3M1	A3M2
Minimum	-8.47	-13.72	-22.23	-20.87	-41.04	-22.56
Maximum	8.54	11.97	20.59	20.23	38.42	39.52
Mean	-0.0002	-0.002	-0.02	-0.02	-0.44	0.0003
Median	0.0007	0.0009	-0.14	-0.15	-0.05	-0.001
First quartile	-0.29	-0.58	-2.15	-2.20	-1.76	-0.46
Third quartile	0.29	0.57	1.83	1.94	0.76	0.46
Variance	0.31	1.29	15.22	15.61	5.93	0.98
Standard deviation	0.56	1.13	3.90	3.95	2.43	0.99

Despite the fact that these assessments rely on descriptive statistics, the results show that it is possible to apply the same model over different areas, since the statistical parameters are not calculated based on individual samples, but over the entire sample population. This strengthens the representativeness of these results for larger regions within the same scene. Determination coefficients and confidence intervals (CI) for β_1 and β_0 at 95% (p-value = 0,05) are provided in Table 3. The R^2 can be interpreted as a measure of proportion for the modeled phenomena explained by the regression model, whereas the inclination test of β_1 represents the significance of the linear relationship between the dependent variable and explanatory variable. The three areas showed CI values for β_1 that contains the value; i.e. at 95%, the inclination β_1 can be considered equal to 1 for all three areas. Similarly, for each area, the CI for β_0 contains the value 0, pointing that at 95% all interceptors cross the origin point. These tests indicate that interpolated surfaces have a linear correspondence with the original values.

Table 3. Determination coefficient and confidence interval for β_1 e β_0

	Area 1	Area 2	Area 3
R^2	0.99	0.99	0.99
β_1 Cnf Lmt -95%	0.99	0.99	0.97
β_1 Cnf Lmt +95%	1.006	1.002	1.002
β_0 Cnf Lmt -95%	-11.40	-5.18	-6.30
β_0 Cnf Lmt +95%	5.42	5.40	3.63

4. CONCLUSIONS

The contribution of this work was to propose and test a method to quantify the interpolation errors associated to the enhancement of SRTM data spatial resolution from 90m to 30m, using a 'reference' grid. The quality of the SRTM-90m data interpolated to 30m was assessed in relation to 30m data available for a pilot area in the State of Montana (US).

Ordinary kriging generates results that are significantly close to the original grid. This is asserted both by the descriptive statistics yielded from residues of DEM differences, and by testing for β_1 and β_0. Eventual drawbacks of the procedure are intrinsic to kriging. These include changes on elevations values due to variance minimization and smoothing of high frequency relief features. Based on descriptive statistics, it is also possible to apply a well-fitted variogram model yielded for a small sampling area to larger adjacent regions. This is because different variogram models proved not to be significantly different for a dataset stemmed from the same source (i.e., SRTM) and covering contiguous regions.

Considering the method envisaged through this work, the globally available 90 m SRTM datasets can be plausibly transformed into higher resolution, 30 m resolution datasets for large neighboring regions and used in multiple applications outside North America.

REFERENCES

Araújo, E. P., 2006. Aplicação de dados SRTM à modelagem da erosão em microbacias por geoprocessamento. M.Sc. Thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil, 88pp.

Demirkesen, A. C., 2009. Quantifying geological structures of the Nigde province in central Anatolia, Turkey using SRTM DEM data. Environmental Geology 56, 865-875.

Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. American Geophysical Union EOS 81, 583–585.

Kääb, A., 2002. Monitoring high-mountaing terrain deformation from repeated air- and spaceborne optical data: examples using digital aerial imagery and ASTER data. ISPRS Journal of Photogrammetry & Remote Sensing 57, 39–52.

Goovaerts, P.A., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology 228, 113–129.

Grohmann, C. H., Riccomini, C., Alves, F. M., 2007. SRTM-based morphotectonic analysis of the Poços de Caldas Alkaline Massif, southeastern Brazil, Computer & Geosciences 33, 10-19.

Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics, Oxford University Press, New York, NY. 561 pp.

Iwashita, F.; Monteiro, R. C.; Landim, P. M. B., 2005. An alternative method for calculating variogram surfaces using polar coordinates. Computers & Geosciences 31, 801-803.

Landim, P. M. B. 1998. Análise estatística de dados geológicos, Fundação Editora da Unesp, São Paulo, SP, 225 pp.

Masoud, A.; Koike, K., 2006. Tectonic architecture through Landsat-7 ETM+/SRTM DEM-derived lineaments and relationship to the hydrogeologic setting in Siwa region, NW Egypt. Journal of African Earth Sciences 45, 467–477.

Nelson, A.; Oberthür, T.; Cook, S., 2007. Multi-scale correlations between topography and vegetation in a hillside catchment of Honduras. International Journal of Geographical Information Science 21, 145–174.

Neter, J.; Kutner, M.; Nachtshiem, C.; Wasserman, W., 1996. Applied Linear Statistical Models, Richard D. Irwin Inc, Chicago, IL, 659 pp.

Rossetti, D. F.; Valeriano, M. M., 2007. Evolution of the lowest amazon basin modeled from the integration of geological and SRTM topographic data. Catena 70, 253-265.

Sommer, S.; Hill, J.; Mégier, J., 1998. The potential of remote sensing for monitoring rural land use changes and their effects on soil conditions. Agriculture, Ecosystems and Environment 67,197–209.

Valeriano, M.M.; Kuplich, T.M.; Storino, M.; Amaral, B.D.; Mendes Jr.; J.N., Lima, D.J., 2006. Modeling small watersheds in Brazilian Amazonia with shuttle radar topographic mission-90m data. Computers & Geosciences 32, 1169-1181.

Valeriano, M. M., 2004. Modelo digital de elevação com dados SRTM disponíveis para a América do Sul. Report - INPE-10550-RPQ/756, Instituto Nacional de Pesquisas Espaciais. São José dos Campos, Brazil, 72 pp.

Yun, S.; Ji, J.; Zebker, H.; Segall, P., 2000. On merging high- and low-Resolution DEMs from TOPSAR and SRTM using a prediction-error filter. IEEE Transactions on Geoscience and Remote Sensing 43, 1682–1690.