Using self-organizing maps to analyze high-dimensional geochemistry data across Paraná, Brazil

Fábio Iwashita [1,2]
Michael James Friedel [2]
Carlos Roberto de Souza Filho [1]
Stephen James Fraser [3]

[1]Univesidade Estadual de Campinas – UNICAMP
{iwashita, beto}@ige.unicamp.br
[2]Crustal Geophysics and Geochemistry Science Center, U.S. Geological Survey, Denver, CO.
mfriedel@usgs.gov
[3]CSIRO Exploration and Mining, Brisbane, Australia
stephen.fraser@csiro.au

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 dataset consisted of 304 samples, 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, radiometrics and magnetic susceptibility measures) and 6 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. The self-organizing maps were used to identify and classify the relationships among solid-phase chemical element concentrations and GIS derived topographic models. The proposed method proved 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, where no statistical assumptions concerning the sampling dataset is required.

Keywords: self-organizing map, soil geochemistry, geomorphometry, Brazil

1. INTRODUCTION

Terrain morphometric features reflect physical and chemical processes responsible for their development. The proper reasoning of the weathering process helps understanding, at least partially, those phenomena that influence landscape formation. Chemical and physical weathering is a coupled process and a significant factor for hillslope shapening, since the mobility of chemical elements is strongly connected with the soil physical-chemical conditions, such as pH, moisture, temperature and porosity (Young, 1980). For example, concave areas on hillslopes are frequently associated with convergent hydrological fluxes, higher soil moisture and reducing environment, whereas convex areas, frequently associated with hilltops, are characterized by divergent hydrological fluxes, oxidizing environment and it is more susceptible to physical erosion (Heimsath et al., 1997).

Early numerical modeling and empirical approaches allowed the quantification of soil mass loss from a physical viewpoint. Such models considered hillslopes uniform along their extensions, i.e., rectilinear, not reflecting the heterogeneity of sediment transport and deposition rate. Most recent models for soil mass loss assume a nonlinear sediment transport rate; they consider that hillslopes, usually, have convex morphometry near hilltop, rectilinear at the middle section and concave at the base (Roering et al., 1999). Morphometric properties could be incorporated into the soil weathering modeling and applied to larger areas using digital elevation models and Geographical Information Systems (GIS) to calculate measures like, slope, aspect, horizontal curvature, vertical curvature and hydrological accumulated flux.

Empirical approaches to analyze soil geochemical data are usually based on statistical multivariate methods, such as multiple linear regression, cluster analysis and factor analysis.

These are robust and reliable methods, though with strong assumptions like normal distributed residuals, stationarity, and non co-linearity between the explanatory variables, where among these methods, multivariate linear regression, one of the most well established approach, penalize a high number of explanatory variables, seeking a balance between the number of variables and the information explained by the model. (Netter et al, 1996). An additional intricacy is the fact that according to Reimann and Filzmore (1999), geochemical data, at regional scale, do not show normal, neither lognormal distribution.

An alternative to analyze multivariate datasets are data mining techniques. Self-Organizing Maps (SOM) Kohonen (2001) are suitable to deal with noisy, non stationary and non Gaussian data. This method highlights nonlinear relationships by topological transformations of the original dataset. The absence of prior assumptions are one of the main advantages of data mining approaches, since traditional multivariate statistics, generally, assume linear relationship between the independent and dependent variables.

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 analyze nonlinear relations among published B-horizon soil geochemical, environmental, relief morphometry, and GIS data from 304 locations using the SOM (Kohonen, 2001) component planes visualization (Penn, 2005) and k-means clustering techniques.

1.1. Study area

Paraná is a state of Brazil, located in the South of the country. 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. The annual average precipitation ranges from 1.500 mm to 2.500 mm.

The geological record of Paraná is characterized by rocks with ages greater than 2.800 million years before present (Minerais do Paraná – MINEROPAR, 1986). The shield, composed of magmatic and metamorphic rocks older than 570 million years, is covered by Paleozoic and Mesozoic volcanic and sedimentary rocks comprising the Paraná basin (Figure 1). This coverage was eroded due to uplift of the continental crust, east of the basin, exposing the basement. More recent sediments (less than 1.8 million years) 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 Pre-Cambrian 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, intracratonic, evolved over the South American platform and its generation began during the Devonian period (approximately 400 million years ago) and ended in the Cretaceous period.

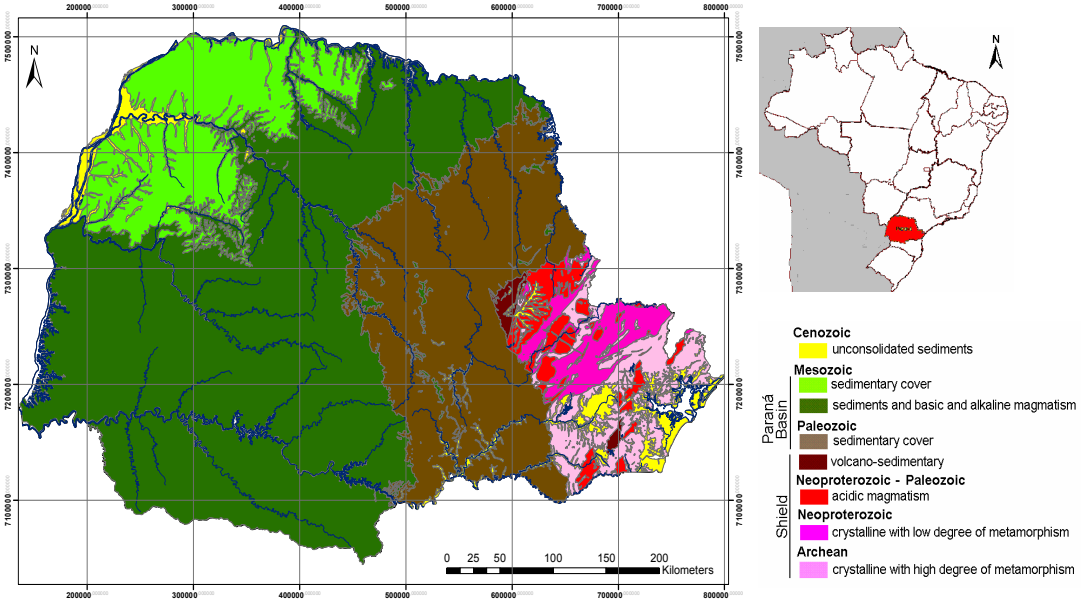

Figure 1. Simplified geological map of Paraná state (modified from Lich, 2001).

2. 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:

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

Where z is the standardized value; x is the raw score; 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 25 variables. Fourth, the k-means clustering technique (Vesanto et al., 2000) 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.

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. The Paraná Agronomic Institute performed analyses of the following variables: pH, $Al_{exchangeable}$ (mg/kg), $Ca_{absorbable}$ (cmolc/kg), $Mg_{absorbable}$ (cmolc/kg), $P_{absorbable}$ (cmolc/kg), $K_{absorbable}$ (cmolc/kg), $C_{organic}$ (g/kg), $H^{+}+Al^{3+}$, $Cu_{extractable}$ (mg/kg), $Zn_{extractable}$ (mg/kg), $Fe_{extractable}$ (mg/kg), $Mn_{extractable}$ (mg/kg), $S_{extractable}$ (mg/kg), $B_{extractable}$ (mg/kg), V% (base saturation), CEC (cations exchangeable capacity), Sum (sum of exchangeable bases), gama-spectrometry (channels total count, Uranium, Potassium and Thorium), and magnetic susceptibility (dimensionless). These data were assembled into a data base by the Paraná State Geological Survey and provided to project personnel (Licht, 2001).

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. Another derived measure, the vertical curvature depicts the hillslope profile: 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. 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.

2. 3. Self Organizing Maps

Self Organizing Maps (SOM) belongs to a category of Artificial Neural Networks (ANN) called competitive learning networks. These are computational models structured as a proxy for the neuron links that constitutes the human brain. The term 'self-organizing' comes from the unsupervised nature of the algorithm, having the ability to organize, or classify, the information without any specifications about the output pattern. The output maps are consisted in neurons organized on a regular two dimensional grid, usually represented as cells on hexagonal or rectangular lattice. Each neuron in the map is represented by a multi-dimensional weight vector $m=[m_1, m_2, \ldots, m_d]$, where d correspond to the dimension of the input vectors. Each neuron is connected to the adjacent neuron by a neighborhood relation, which defines the topology, or structure of the map (Vesanto et al., 2000).

Each sample is associated to a vector with properties that reflect its contributions relative to the other variables (Figure 2). From this 'cloud' of vectors the Best Matching Unit (BMU) is iteratively determined for each variable, by $\|x - m_c\| = \min_i \|x - m_i\|$ 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. Then the neurons within a specified neighborhood are updated by, $m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)]$, where t is the time step; $m_i(t)$ is the current weight vector, $\alpha(t)$ is the adaptation coefficient, and $h_{ci}(t)$ is the neighborhood kernel centered on the winning neuron c. The topology of the vectors is altered until convergence conditions are reached (Kohonen, 2001; Vessanto et al, 2000).

The resulting maps are organized in such way that similar data are mapped into the same node or into neighborhood nodes, i.e., it is a data classification based on their topology in the n-dimensional space (Figure 2). The U-matrix is compose by the BMUs, obtained from the weight vectors associated to the input vectors, thus, each variable produces a weight map, or component map, arranged in a grid that coincides to the U-matrix. These maps can be used to visualize the correlations, once the cells with similar colors and positions inside the map describe similar contributions for the construction of the U-matrix.

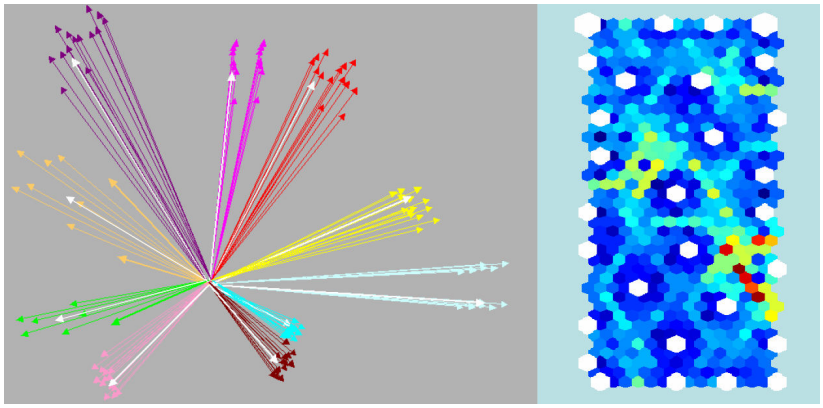

Figure 2. Each sample describes a vector in *n*-dimensional space (left picture). The white vector represents the best matching unit, these vector are initialized by a random seed vector which interactively change its position to produce the final representing node. Similar samples are gathered in the same node projected on a 2 dimensional map (right picture) where the colors represent dissimilarity values. Blue colors match up with low values, while reddish tones correspond to high values of dissimilarity (Fraser and Dickson, 2005).

Once component maps and the U-matrix are produced, the K-means clustering method was employed to classify the cells as a post-processing analysis. The method finds a vector of means from a specific group of cells so as to minimize the within-cluster sum of variances. The algorithm is deemed when convergence for a local minimum variance is reached and the assignments no longer change.

3. RESULTS AND DISCUSSIONS

Inspection of the component maps (Figure 4) revealed several soil geochemistry relations. For example, the elements B, Ca, Mg, K, Cu, P and Zn are strongly correlated, where Cu and Zn are linked to the presence of mafic rocks from the second and third plateau, while 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, base saturation and Al (blue boxes on Figure 4), and Fe and aspect (green boxes on Figure 4). 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 reduced the reduced form .

Cations exchangeable capacity (CEC) is a measure of allowance for solid particles to exchange positively charged ions with soil solution, constituting an approach to quantify soil fertility, i.e., the soil capacity to retain nutrients. From a hydrological point of view, this is an important parameter to characterize subsurface chemical weathering, since it is related to the capacity for adsorption of elements. Figure 5 shows the scatterplots for soil geochemistry variables (CEC and organic carbon) and measurements that can be obtained by remote sensors (gamma-spectrometry, elevation and slope). These ‘samples’ are actually the nodes stemmed from the SOM classification. Each node is associated to one or more samples according to a topological similarity. The colors represent different clusters, calculated using k-means technique.

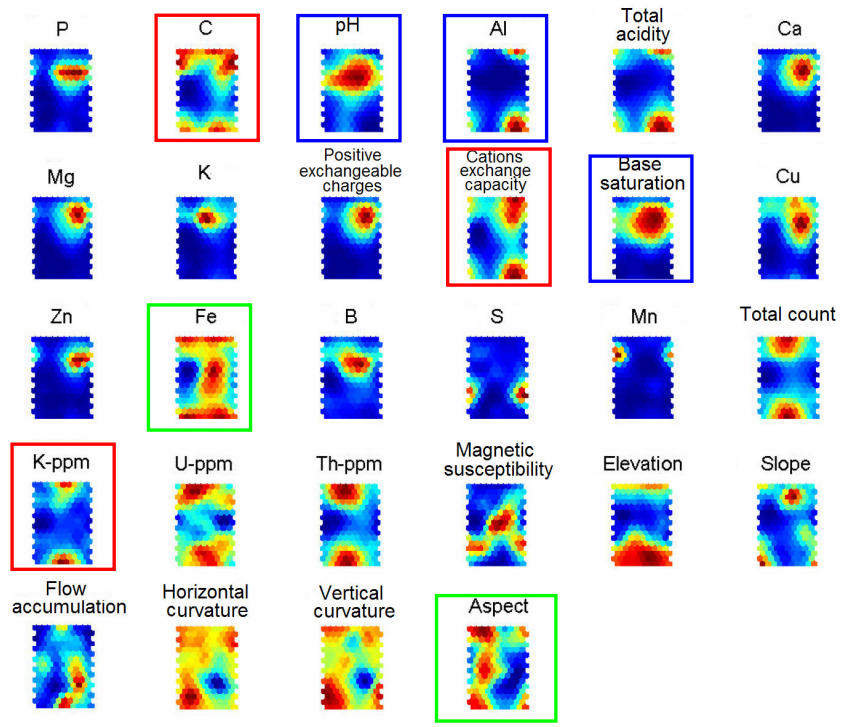

Figure 4. Component planes used to visualize nonlinear correlation. For example, the elements highlighted by boxes in similar colors, C, cations exchangeable capacity and K are correlated (similar colors), whereas pH and base saturation is inversely correlated with aluminum and iron is inversely correlated with aspect (opposite colors).

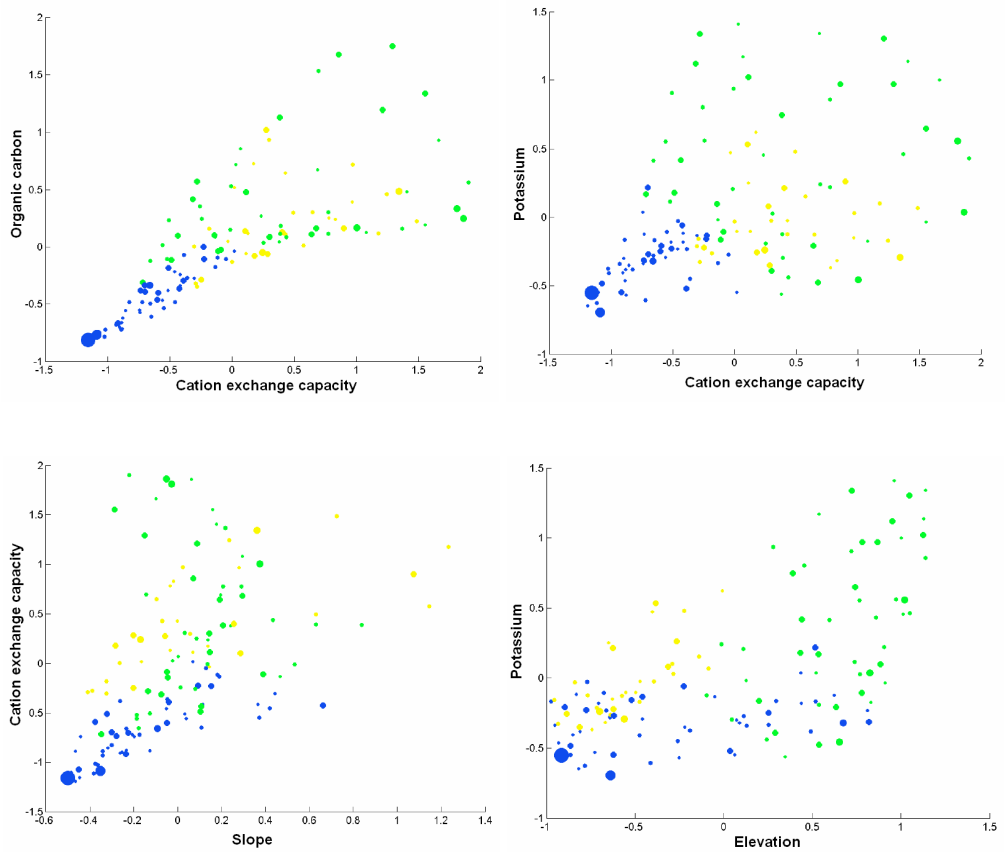

Figure 5. Scatterplots expressed as z-scores of the variables organic carbon, potassium (gamma-scpectrometry), cations exchange capacity, slope and elevation.

The graphs can be interpreted considering the general dispersion of each element, and the pattern of each cluster individually. The CEC is associated with electronegative charges from clay mineral, colloidal silica and humus content, here expressed by organic carbon. There is a degree of interactivity among the variables. A strong correlation between organic carbon and CEC is observed, whereas the CEC is correlated with the potassium content yielded from the gamma-spectrometry data. The correspondence between soil variables and relief features can be explained by the solubility of elements, which is associated with subsurface hydrology and weathering. More soluble elements present higher susceptibility to be transported and should show a sharper relation with the hydrological path on subsurface.

4. CONCLUSIONS

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 2-dimensional surface while preserving important nonlinear relations. 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 understand hillslope chemical weathering, erosion, and landscape evolution in other locations and environmental settings.

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, where no statistical assumptions concerning the sampling dataset is required

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

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

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

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

Friedel, M.J., Souza, O.F., Yoshinaga, S. P., Silva, A. M. 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. *in review.*

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

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

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

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

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

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

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

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.

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

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

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

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

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**, v. 32, p. 1169-1181, 2006.

Valeriano, M. M., Rosetti, D. F., Albuquerque, P. C. G. TOPODATA: desenvolvimento da primeira versão do banco de dados geomorfométricos locais em cobertura nacional. In: Simpósio Brasileiro de Sensoriamento Remoto (SBSR), 14, 2009, Natal. **Anais XIV Simpósio Brasileiro de Sensoriamento Remoto**. São José dos Campos: INPE, 2009. Artigos, p. 5499-5506.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. **SOM Toolbox for Matlab 5**. SOM toolbox team, Finland. Helsinki University of Technology, Laboratory of Computer and Information Science. 2000. Available at: http://www.cis.hut.fi/projects/somtoolbox/. Accessed on June, 21th, 2010.

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