Águas Subterrâneas (2013) 27(2): 47-65.

SPATIAL MODELING OF WATER SYSTEMS USING PARTIAL LEAST SQUARES: A CASE STUDY FOR PARACATU BASIN (SF7) IN MINAS GERAIS STATE, BRAZIL

MODELAGEM ESPACIAL DE SISTEMAS HÍDRICOS POR MÍNIMOS QUADRADOS PARCIAIS: ESTUDO DE CASO PARA A BACIA DO RIO PARACATU (SF7), MINAS GERAIS, BRASIL

Vitor Vieira Vasconcelos[1], Paulo Pereira Martins Junior[2], Renato Moreira Hadad[3]

Artigo recebido em: 25/11/2012 e aceito para publicação em: 04/06/2013.

Abstract: Through a cartographic approach, the aim of this paper is to bring useful information to understand the recharge of basins water systems. The Partial Least Squares modeling is applied to investigate the spatial relations between the environmental attributes of nested sub-basins of Paracatu Basin (a tributary of São Francisco River) and the total flow, quickflow, interflow and baseflow for these respective sub-basins. The results of the regression indicate the role of each environmental attribute in the hydrological and hydrogeological processes. The cartographic products are specific flow maps useful for basin management, obtained by the weighted overlay of statistical model results, and also maps that evaluate the prediction uncertainty and the hypothesis of regional groundwater flows. The model also does the cartographic regionalization of the specific flow for the mouth of Paracatu River where there is not any gauging station. The results of this methodology are instrumental for the sustainable management of land use and water resources, and can be part of master plans and environmental zonings.

Keywords: Hydrology. Partial Least Squares. Hydrogeology. Watersheds. Specific Flow.

Resumo: Este artigo tem como objetivo trazer subsídios para a compreensão da recarga dos sistemas hídricos, em uma perspectiva cartográfica. Utiliza-se o método de modelagem por Mínimos Quadrados Parciais para investigar relações espaciais entre as características ambientais de subbacias hidrográficas aninhadas na Bacia do Rio Paracatu (afluente do Rio São Francisco) em relação à vazão total, ao fluxo rápido, ao interfluxo e ao fluxo de base dessas respectivas subbacias. Os resultados da regressão fornecem indicações do papel de cada atributo ambiental nos processos hidrológicos e hidrogeológicos. Apresentam-se, como produtos cartográficos úteis para a gestão de bacias, mapas de vazão específica obtidos pela conjugação ponderada dos resultados das modelagens estatísticas, bem como mapas que avaliam a incerteza de predição e a hipótese de fluxos hidrogeológicos regionais. O modelo também regionaliza a vazão específica cartograficamente para a região da foz do Rio Paracatu, não abarcada por estações fluviométricas. Os resultados apresentam informações para a gestão territorial do uso do solo e da água, podendo ser agregados a planos diretores e zoneamentos ambientais.

Palavras-chave: Hidrologia. Mínimos Quadrados Parciais. Hidrogeologia. Bacias Hidrográficas. Vazão Específica.

INTRODUCTION

The reclamation and maintenance of the hydrological cycle, both in quantitative and in qualitative terms, are dependent of the correct planning of the soil and biota environmental impacts. Keeping the vegetal covering and handling soils in significant areas for aquifer recharging are crucial measures to ensure the water conservation. This facilitates the water percolation through soil, and ensures a steadier water flow into the surface water courses, mainly in the dry season (MARTINS and VASCONCELOS, 2005). Therefore, the awareness of the hydrological process is a wise way to accomplish the integration between the man-

[1]Doutorando em Geologia na Universidade Federal de Ouro Preto. Consultor Legislativo de Meio Ambiente e Desenvolvimento Sustentável na Assembleia Legislativa de Minas Gerais (vitor.v.v@gmail.com).
[2] Pesquisador Pleno da Fundação Centro Tecnológico de Minas Gerais - CETEC-MG. Professor do Departamento de Geologia da Universidade Federal de Ouro Preto (paulo.martins@cetec.br).
[3] Professor da Pós-Graduação em Tratamento da Informação Espacial / Geografia da Pontifícia Universidade Católica de Minas Gerais (PUC-Minas). Pró-Reitor da PUC-Minas, campus Barreiro (rhadad@pucminas.br).

agement of the soil occupation and the management of the hydraulic resources. In particular, the relationship between the environmental attributes and the underground, surface and subsurface water flows differ spatially, and may provide a basis for the best practice of agricultural projects, engineering works and other uses of the land. Understanding the aquifer recharge and discharge is of paramount importance for an integrated management of the surface water and ground water (ARRAES, 2008).

Objectives

This study intends to present multivariate partial least squares regression to show the influence of environmental characteristics (independent attributes) on the components of total flow, base flow, interflow and quickflow of the water systems regarding the watershed of each gauging station Based upon the weighted attribute is therefore possible to make a crossing weighted overlay of cartographic databases to obtain a more detailed map of the areas of greater significance for Paracatu's river water systems recharge, taken as the study case.

It is important to notice that the bibliography on multivariate statistics applied to Hydrology is chiefly headed to water flow estimates. In the case of multivariate models incorporating flow and physiographic attributes, classical models are oriented mainly to estimate river flows at points of watercourses where there is no measurement stations, or to fill gaps and extension of data series. Such considerations can be easily verified in reference works for Hydrology, for instance in Tucci (2002; 2009) and Naghettini and Pinto (2007).

In the present study, the evaluation of the dependent variable is no more than an interim by-product. The ultimate goal refers to the preparation of a map of a recharge water system with the highest possible amount of geo-information that might bring a deeper understanding of the hydrological and hydrogeological processes and, as a consequence of this, to support a sustainable territorial occupation for the watershed. This is a theoretical yet little studied field for theories of multivariate statistics, and hydrological and hydrogeological modeling do not have consolidated tools specialized to meet such practical goals.

This article does not aim at establishing a complete and accurate geo-mathematical model to estimate dependent variables of flow, because we are aware that the studied phenomenon is too complex and because several other important variables were not incorporated in the modeling. Although recognizing that the maturation of this scientific area still lies in its exploratory phase, tools are proposed for assisting the geoscientist and the environmental manager to extract useful information from the mapping and hydrological data available.

Theoretical Background

Modeling of water circulation in environmental systems - the challenge of multicollinearity

Holtschlag (1997), Brandão and Gomes (2003), Flynn and Tasker (2004), Latuf (2007) and Gomes (2008) employed multivariate analysis integrating cartographic and hydrogeological data (primary or inferred from the surface water flow) so as to constitute hydrological models able to explain processes of the water climatic and hydrogeological balance. It must be emphasized that all five authors acknowledged that those models have severe limitations. The main limitations are related to the cartographic variables, theoretically, of the utmost importance. They were dropped from the model for not presenting a statistically significant relationship with the dependent variable. New variables and new statistical tools, as pointed out by the above mentioned authors, would be required to advance into this novel territory of scientific knowledge. This is an open path to move forward on this study.

As for the variables used as independent variables in environmental studies, we start from the assumption that an intrinsic multicollinearity exits among analyzed data, due to the fact that attributes of different layers of environmental information present similar spatial distribution. For example, a sandstone tends to form more sandy and less fertile soils, that under a stable geomorphotectonic environment tends to drive into deep quartzarenic neosols on a plan to wavy relief which, in the case of a unimodal climate (year divided into wet and dry season) supports a shallower vegetation such as savannah and meadow (KHEORUENROMNE et al., 1998; RETALLACK, 2008) This multicollinearity between the independent variables hinders the separation of their effects on the dependent variable, increases the standard deviation of the regression, disturbs the normal operation of the significance tests and obtain unstable estimators (HAIR JUNIOR et al., 2009).

In conventional regionalized hydrological models, chiefly oriented to stream flow forecasting, usually the attempt to use few variables ultimately leads to the selection of varia-

bles strictly related to the spatial attributes. Such spatial variables embody within an index, the most complete explanation about the hydrogeological variables regarding the system of the basin, hence, a high degree of colinearity with those variables (SILVA JUNIOR et al., 2002).

Recurring examples of parametrically spatial variables are the altimetry and the watershed area. Such variables are linked to the scale of the watershed analysis, and clearly demonstrate the variation of hydrogeological processes from upstream to downstream in the watershed. Sub-basins of smaller area near the headwaters and ridges generally have a steeper slope and are more affected by the orographic precipitation (TUCCI, 2009). These basins have a fast response to precipitation and the concentration of the runoff and their flows are deeply affected by short-term convective precipitations (TUCCI, 2009). As for the hydrogeological processes in these sub-basins, soil characteristics conditioned by altitude lead to a spatial predominance of areas where the groundwater recharge phenomenon is more expressive, although the typology of the discharges into resurgences are also relevant. Tucci (2002), on his turn, emphasizes the downward trend of the specific flow rates in relation to the increase of the areas of the watersheds, based on data from several Brazilian water basins.

However, the use of these strictly spatial variables diminishes the power to explain the difference in the roles that each environmental attribute has, and masks the heterogeneity of the hydrogeological processes within and among the reference sub-basins. When evaluating the incorporation of other environmental variables, in a typical process of a stepwise multiple regression, the explanatory power would have already been entirely taken by the spatial variable(s) of higher correlation, and this leaves spurious the remaining coefficients of partial determination. In a nut shell, the choice is being usually made looking for the simplicity of the prediction, leaving aside the complexity of the explanation of hydrological processes. Therefore, it becomes necessary to seek statistical techniques that may be more effective on the incorporation of the relationship of multicollinearity among the independent variables that should be used in hydrological and hydrogeological models.

Another challenge for the modeling of hydrological processes is the consideration of regional groundwater flow into the streams, because it may pass under the gauging station and outflow farther downstream. Furthermore, such flows may not even respect the drainage dividers, in which case the boundaries of the watershed do not coincide with the boundaries of the hydrogeological basin.

Usually, it is assumed that the underground regional flow is primarily related to the base flow, as it corresponds to the deep aquifer outflow. However, some quickflow and interflow may also come from regional groundwater flows, due to karstic pipelines, river/fracture systems, flow sharing among river and alluvial aquifers, and also due to the hydrogeological piston flow effect.

The hydrogeological piston flow effect was studied by Kirchner (2003) and Gonzales et al. (2009), based upon geochemical tracers and isotope dating. It shows that the groundwater flow generally responds to the regional rainfall as quickly as the surface flow does. This happens as a result of the rainfall on the recharge basin areas, which causes pressure waves on the aquifers as it was a system of communicating vessels. On the other hand, clay lands in the low level terrains retard and dilute the subsurface flow over time.

Partial Least Squares – PLS – Regression

The partial least squares – PLS is a regression model suitable for treating multicollinearity. The PLS mathematical-statistical basis is a non-linear algorithm that, at each iteration, seeks to maximize the variance of the dependent variables explained by the independent variables. A more detailed explanation of the algorithm can be found in Haelein and Kaplan (2004). PLS rotates the reference system of the feature space and aligns the axles with the highest explanation vectors on the dependent variable (HAIR JUNIOR et al., 2011). Thus, it reduces the independent variables to a minimum of optimized vectors to accomplish the regression. Subsequently, it redistributes the weights among the original variables. Differently from the original environmental data, the new vectors have minimal multicollinearity among themselves.

PLS was developed by Wold (1981; 1985) originally aimed at econometric applications. Recently, PLS started to be used for remote sensing in environmental studies (KOOISTRA et al., 2001; SCHMIDTLEIN and SASSIN, 2004; NOBRE, 2006) and hydrology (GEBREHIWOT et al., 2011).

As a soft modeling technique, PLS is theoretically less dependent on assumptions of

distribution, such as multivariate normality and absence of multicollinearity (HAIR JUNIOR et al., 2009; GARSON, 2010). The theoretical base of the conventional linear regression method assumes the normal distribution of the dependent variable (GARSON, 2010). However, the flow variables generally follows a Gama distribution (NAGHETTINI and PINTO, 2007), mainly a log-Pearson type III (INTERAGENCY ADVISORY COMMITTEE ON WATER DATA, 1982). Such distribution is theoretically expected for non-negative variables with positive skewness and elongated right tail. In this context, this paper proposes that non-parametric statistical techniques for soft modeling could bring more reliable results for the hydrologic modeling.

PLS application aiming at exploratory purposes has been accepted for populations as small as 20 samples (TOBIAS, 1997; VILARES et al., 2010; HENSELER et al., 2009; GARSON, 2010), although the prediction reliability increases as the sample universe expands (HUI and WOLD, 1982; CHIN and NEWSTED, 1999; MARCOULIDES and SAUNDERS, 2006).

PLS is just like a structural equation model (SEM), as it also enables complex modeling with latent formative and reflective constructs (composite variables) within the same model (ANDREEV et al., 2009). In reflective constructs, variables with assumed similar behavior (i.e., presenting multicollinearity) will be merged into one or more major components (as a factor analysis or a major components analysis). Conversely, in the formative constructs, variables are supposed to have distinct behavior (bearing no multicollinearity), and to these variables different coefficients will be allocated, accordingly to its effect on the dependent variable (similarly to a conventional multiple regression). Each component in PLS operates as a reflective latent construct of the independent variables, whereas the sum of the components has a formative effect regarding the dependent variable.

In reference to the dependent variables of flow (quickflow, interflow or base flow), it should be emphasized that it is restricted to the limited number of gauging stations in each study area. In this context, due to the wide range of environmental attributes (independent variables) and to the complex methods of multivariate modeling, an overfitting situation is always possible - that is, a complex modeling with a large determination coefficient which nevertheless fits itself largely to the prediction drift (due to noise and/or random errors) instead of adjusting itself to the environmental processes.

Hypothetically, an overfitting model would have difficulty in demonstrate its validity if tested against additional gauging stations on the watershed and, furthermore, if an attempt was made to extrapolate it to other watersheds. *Cross validation* and *jack-knifing* are recommended methods for evaluation of the overfitting through indexes acquired by resampling. In these methods, the regression model is repeated *n* times while some cases are withdrawn for observation to evaluate the stability of the sample as a whole.

As it is essentially a non-parametric method without assumptions of distribution, PLS precludes testing the variance ratios (F tests) and other conventional goodness of fit indices (HENSELER et al., 2009; CHIN, 2010). For model evaluation, accuracy indices made up from resampling techniques are used (UMETRICS, 2008; GARSON, 2010), which has the additional advantage of making possible the model overfitting evaluation.

Characterization of the Paracatu River Basin

The Paracatu River Basin is almost entirely in the State of Minas Gerais, with some small highland areas entering in the State of Goiás and in the Federal District of Brasília (Figure 1). This basin has 45,154 km^2, and it is the largest river basin among the direct tributaries of the São Francisco River. The climate is rainy megathermal of the Aw type (IGAM, 2006). It is a typical tropical rainy climate, with high temperatures and unimodal rainfall fluctuation concentrated between October and April, when the average rainfall reaches 93% of the annual total (MULHOLLAND, 2009).

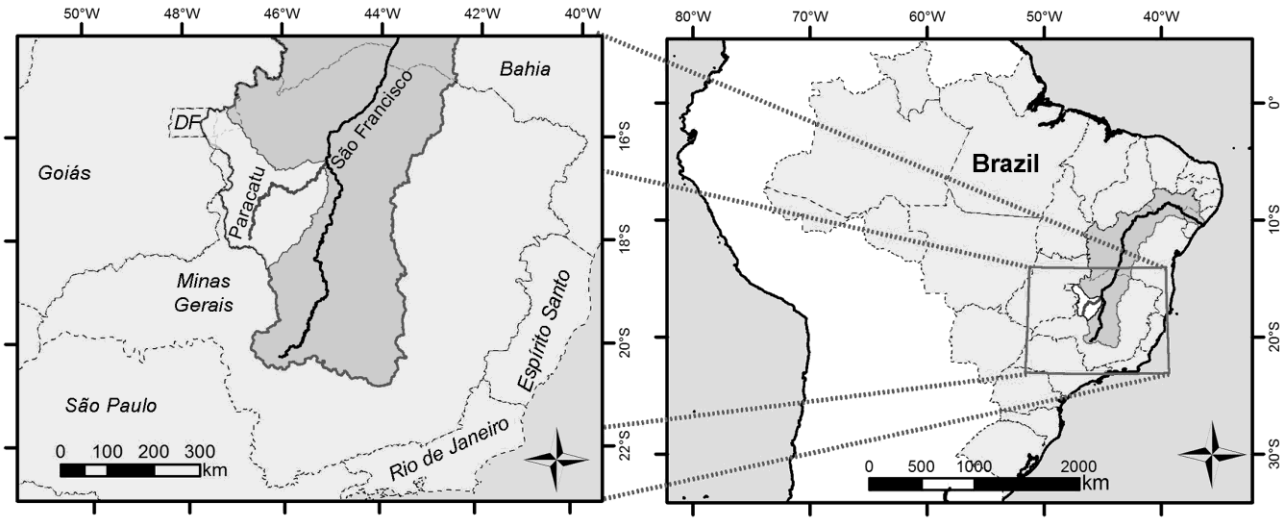

Figure 1 – Geographical Location of the Paracatu River Basin

The stratigraphy of the Paracatu River Basin affects different rock systems of bearing aquifers (Figure 2).

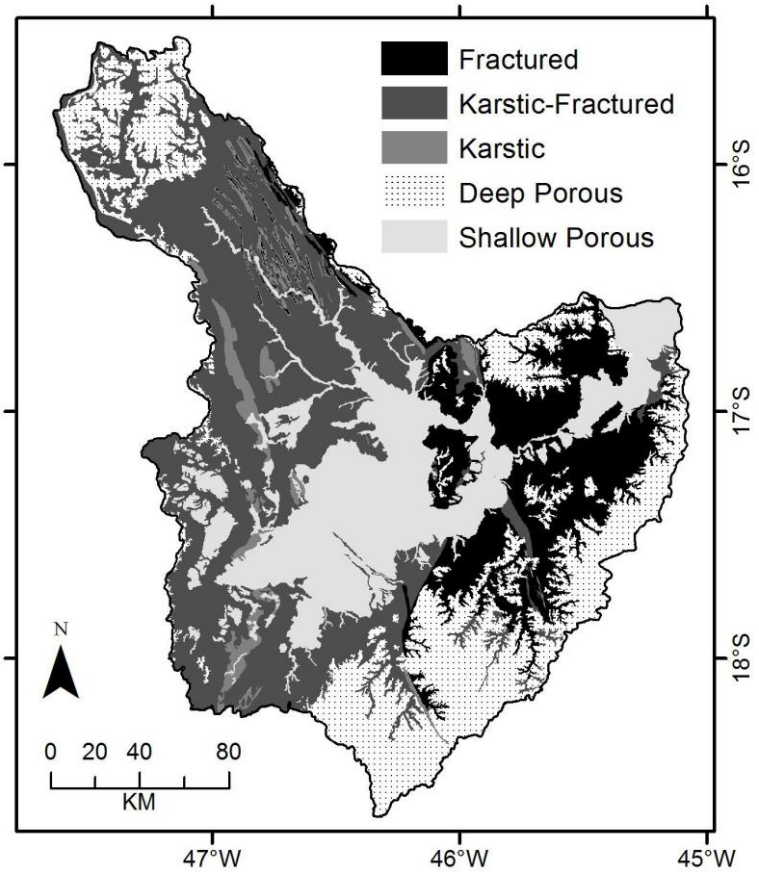

Figure 2 – Systems bearing rocks aquifers of the Paracatu River Basin, inferred from Martins Junior's lithostratigraphic bases (2006).

The deep sedimentary beddings (Cretaceous stratigraphy and Tertiary-Quaternary detrital cover in the highlands bedside) present themselves as the major potential areas for recharge and storage of groundwater, according to CETEC-MG (1981). The shallow detrital Tertiary-Quaternary toppings of the lowlands and the alluvial quaternary toppings possibly have a secondary role more akin to flow regulation (RURALMINAS, 1996). The aquifers connected to karstic and metamorphic beddings depend heavily on the presence of ductile and brittle structures, whose spatial heterogeneity comes from the structural geological history of the basin.

METHODOLOGY

Statistical Modeling by PLS

The SIMCA-P+13 program was used to run a Partial Least Squares (PLS) regression projection model to latent structures. This paper uses the term 'model' addressing the numerical model generated through PLS, and not to any prior algorithms or procedures used to calculate independent or dependent variables.

Following Barclay et al (1995) guidance to handle the PLS, a maximum of one predictor component was used for every 10 cases of the sample population. Thus, the model was limited to two components extracted from the independent variables for the regression of each dependent variable. As recommended by Marcoulides and Saunders (2006) and Rouse and Corbitt (2008), the overfitting of the predictive model was assessed by resampling techniques.

Statistical procedures were based on the following recommendations Chin (2010) for the PLS method: correlation analysis between the variables and analysis of the weights and accuracy indicators. Regression was analyzed based on the coefficient of determination (R^2), compared to the standard deviation of the residuals and to the Q^2 (cumulative variance that can be predicted by its components), as recommended by Umetrics (2008). The standard deviation and the Q^2 in PLS are calculated by resampling (through jack-knife and cross-validation, respectively) of the dependent and independent variables which are rescaled for standardization (Z), thus making possible the comparison between the different models to be tested. The larger R^2 and Q^2, and the lower the residual standard deviation, the more suitable model is. Q^2 is calculated by:

$$Q^2(cum) = (1.0 - \Pi(PRESS/SS)a) \quad \text{(Equation 1)}$$

Where, a = 1 ... A;

Π(PRESS/SS)a = PRESS/SS index times each individual component;

PRESS = predicted residual sum of squares;

SS = sum of squares of the observations;

A = total number of components.

Preliminarily, we carried out a hierarchical analysis of the clusters using the Euclidean quadratic distance between the standardized variables (Z) for each section of the nested basins, in order to view their correlation through a dendrogram.

Therefore, while working on the regression modeling, the incorporation of each variable aimed at the evaluation of the prediction gain (coefficient of determination), and of the overfitting (Q^2 and standard deviation), but also aimed at the possibility of theoretical hydrogeological explanation, as well as at increasing and detailing cartographic information in the final products. In the regression model, the variables were analyzed for their influence value on the projection (VIP – Variable Importance in the Projection – Equation 2), by their standardized coefficients (making it possible to compare the variables) and through the respective standard deviations of the VIP and coefficients. These indices were obtained by a resampling technique on the dependent and independent variables rescaled for standardization [Z], according to the guidelines of Umetrics (2008). The VIP makes possible to visualize the influence of each independent variable in the model, as if there were no other correlated variables. Conversely, for the standardized coefficient, the weights are redistributed among the correlated variables. VIP may be calculated as the square root of the sum of (SSY) of the PLS weights (w_{ak}) of the regression for a certain independent variable K. Thus, its formula can be expressed as:

$$VIP_{Ak} = \sqrt{\sum_{a=1}^{A} (w_{ak}^2 * (SSY_{a-1} - SSY_a)) * \frac{K}{(SSY_a - SSY_A)}}$$

(Equation 2)

We have also performed analyzes of scatter plots showing loads of each variable according to the axis of the regression components. This kind of chart allows us to see the correlation between variables and propose hypotheses of the application of different processes on the same independent variable for each PLS component axis (UMETRICS, 2008).

Dependent and independent variables

As dependent variables the models encompass total flow and its components of base flow, interflow and rapid flow estimated from the daily flow data from gauging stations with a reference period between 1976 and 2000. Vasconcelos et al. (2013) separated these components using BFLOW recursive digital filters (LYNE and HOLLICK, 1979) calibrated (a) by the influence of the surface runoff (LYNSLEY et al., 1975) and (b) by the inflection in the re-

cession curve at the dry season period (BARNES, 1939), according to Figure 3. The location of the gauging stations is shown in Figure 4.

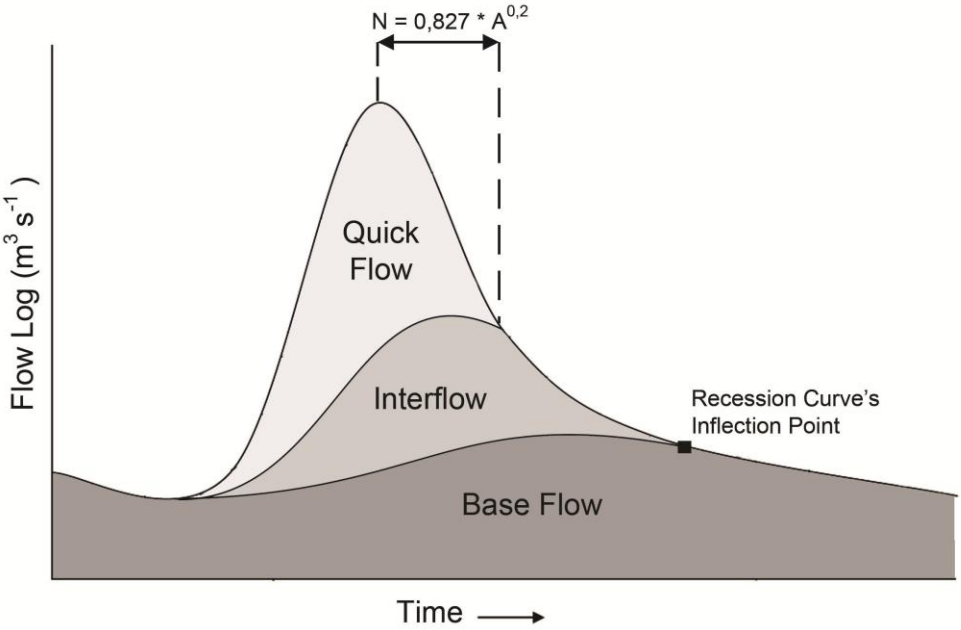

Figure 3 – Conceptual runoff hydrograph. N is the number of days after a peak in the hydrograph to cease the participation of a rainfall event in the runoff, and A is the catchment area in km^2 – Empirical formula of Lynsley et al. (1975).

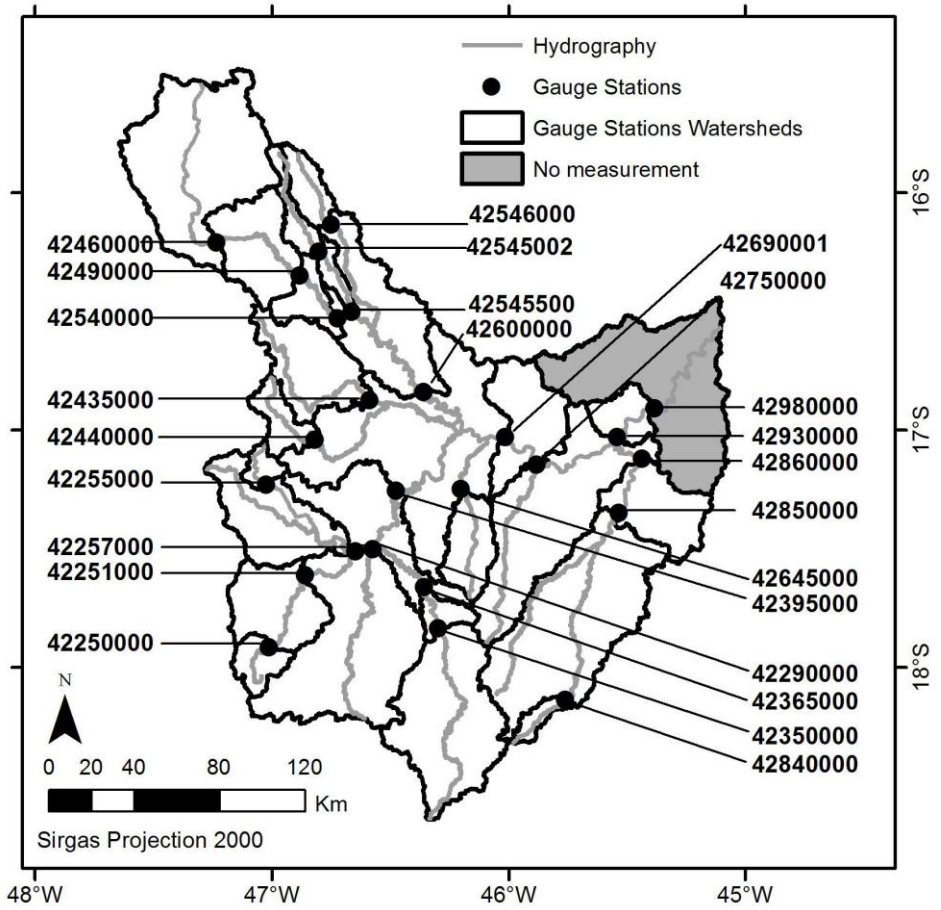

Figure 4 – Gauging Stations in Paracatu River Basin and their respective watersheds.

The independent variables are based on cartographic bases that are available or can be built in most environmental studies. The databases are shown in Table 1. More detailed expla-

nation about the calculation of the independent variables besides their cartographic visualization can be found in Vasconcelos et al. (2012). Variables were standardized before entry into the model, as recommended by Garson (2010). PLS estimation bias, which occurs for reflective constructs with fewer than 10 indicators (Chin, 1995), were avoided because the model uses 39 indicators to form each reflective construct.

It should be noticed that we were unable to find in the literature a previous incorporation of the absolute curvature, distance to brittle structures (mesofractures), and water springs level in a model that investigates component flows. The experience presented in this paper is also pioneering regarding the incorporation of data from the Brazilian Underground Water Information System (SIAGAS) aiming at using the data from water wells for a better understanding of the flows at gauging stations.

Morphometric and hydro-morphometric variables were calculated using SAGA 2.0.8, ENVI 4.8 and ArcGIS 10 (Spatial Analyst extension) programs, over a Digital Elevation Model Hydrologically Consistent – DEMHC. The MEDHC was prepared using the extension Hydrotools 2.0 for ArcGIS 10 program and the free program Saga 2.0.8.

Table 1 – Information sources.

	Attribute	**Source**	**Scale**
Independent variables	Morphometric variables: elevation, normalized elevation, standardized elevation, mass balance, slope height, slope, accumulated slope of the watershed, curvature, absolute curvature, convergence index, ruggedness index, vectorial ruggedness index, flow dispersion, topographic wetness index, topographic index of subsurface flow, sky view factor, land view factor, sky visibility, total annual insolation, diurnal anisotropic heating, prevailing windward index (East-Northeast - ENE), prevailing leeward index (ENE), prevailing wind effect index (ENE), effective strength of the prevailing air flow (ENE)	Hydrologically consistent Digital Elevation Model (DEM) based on data from the Shuttle Radar Topography Mission (SRTM) compared to IBGE hydrography data	1:100,000
	Morphometric drainage variables: channel network base level, water springs level, vertical distance to channel network base level, horizontal overland distance to watercourse, vertical overland distance to watercourse, distance to basin outfall (mouth)	IBGE hydrography and altimetry from the SRTM satellite	1:100,000
	Distance to Brittle Structures	Performed through aerial photographs, Martins Junior (2006)	1:50,000
	Average annual rainfall	Regionalized rainfall stations, Nunes and Nascimento (2004)	5,221 km^2/station in the interpolation mesh (stations inside and outside the basin)
	Drilled Wells Attributes (flow stabilization, specific flow, dynamic level, water table lowering)	Underground Water Information System (SIAGAS) accessed in 3/28/2012	148 km^2/wells inside the watershed
	Space Variables (latitude, longitude, distance to the edge of the basin)		
Dependent Variable	Total Flow, Base Flow, Interflow and Quickflow	Gauging stations in National Water Agency (ANA) network, accessed in 3/20/2011	1,802 km^2/ station

The MEDHC's primary source was the altimetry data image from the Shuttle Radar Topography Mission (SRTM) and the official hydrography from the *Instituto Brasileiro de Geografia e Estatística* (Brazilian Institute of Geography and Statistics – IBGE) (1:100,000). First, the lentic water bodies corresponding areas delimited by IBGE base were altimetrically leveled. Then we used the reconditioning digital elevation model called AgreeDem (HELLWEGER and MAIDMONT, 1997) with a 2 cells buffer (120 meters for either river bank) gentle deepening down to 10 meters, and a channel deepening of 5 meters. Reconditioning was completed through the Saga program ensuring that the drainage depth was deepened at least one meter in relation to the minimum elevation of the adjacent cells to the watercourses. With the reconditioning, we try to compensate the height of the riparian canopy of the hydrography along the savannah and even improve the hydrological consistency of the elevation model. The removal of the depressions occurred in two steps, as proposed by Ferrero (2004). In the first step, we used the technique of removing depressions barriers through deepening the outflow channels down to a maximum of 4 meters deep. About 50% of depressions were eliminated this way. Then, we have filled up the remaining depressions (sinks), keeping their tilt towards the point of lowest elevation, using Wang and Liu (2006) algorithm.

Regarding the vector elements such as mesofractures and drainage, the literature on statistical modeling of hydrological regionalization traditionally uses density variables, i.e., value / area (NAGHETTINI and PINTO, 2007*)*. However, aiming at to retranslate cartographically the modeling results more accurately we should use the distance variables to the vector element, because this will bring a precise physical quantification of the final cartographic products for each raster grid cell in the ArcGIS program.

Regarding the mesh of gauging stations, the World Meteorological Organization - WMO - recommends a minimum density of 3,000 km^2 per station (RURALMINAS, 1996), and this rate was achieved for Paracatu basin. It is worth noting, however, that the amount of gauging stations is a boundary condition affecting the number of regression cases and influencing the reliability of the statistical results. Such data availability, although not inhibiting the statistical analysis of smaller data sets, turn them more useful to accomplish exploratory functions than confirmatory functions.

Modeling the hypothesis of regional flows

Two statistical models have been tested. As this study deals with nested basins, the first model groups the variables on each section of the watershed and limits the drainage to the gauging stations - thus, each portion of the basin is used only once for regression.

In the second model we assumed regional flows crossing the watersheds sections and outflowing into the watercourse after the gauging station. For this model we have grouped the variables by the total drainage area of each gauging station, under the assumption that the entire area upstream of the station will influence its flow components.

Cartographic Representation of Results

Finally, once the most reliable regression model had been evaluated and chosen, the weigh the multiple regression assigns to the reflective components was transferred back to the cartographic base of the original environmental variables. The weight of the independent variable times its value in each raster grid cell, has generated raster layers with its relative influence over each dependent variable. The layers of all independent variables were summed by overlay, outputting the cartographic product named map of Specific Flow Units.

The map of Specific Flow Units has a companion map explaining the statistical uncertainty of the regression associated with its spatial distribution. This last map was calculated based on the distance between the predicted and the specific discharge flow for each section of the basin. Such proposal gives transparency to the spatial heterogeneity of the uncertainties related to the model prediction. The observed and predicted specific flow was redistributed by section concatenated from Equation 3.

$$Q_{section} = [(Q_{container} * A_{container\ flow's\ area}) - (Q_{internal} * A_{internal})] / (_{container\ flow's\ area} - A_{internal}) \quad \text{(Equation 3)}$$

Where,

Q = Specific flow;

A = Container basin or internal drainage area.

The same map carries a scheme proposing underground regional flows, which could balance the model and cancel the differences between the predicted and observed flow in each section, while maintaining the topological flow consistency along the drainage. These flows hypothesize that the regional groundwater flow occurs through the inner basins upstream or from the border with the adjacent basins to Paracatu Basin

RESULTS

Dendogram

The dendrogram of the used variables is presented in Figure 5. From the dendrogram, it was possible to identify certain groups of variables related to the topographic relief (macro, meso, micro), as well as the water depth of shallow aquifers, the ruggedness of the relief, the depth of the aquifer, the river flow and the relief of the valleys. At a higher level of abstraction regarding the dendrogram, groups of variables related to the topographic relief are all nested in the same cluster, which also approaches the variables related to the depth of the aquifer. In branches with greater independence, there are clusters related to the flow of the rivers, and to the relief of valleys and other isolated variables, such as longitude, latitude and the flow from wells.

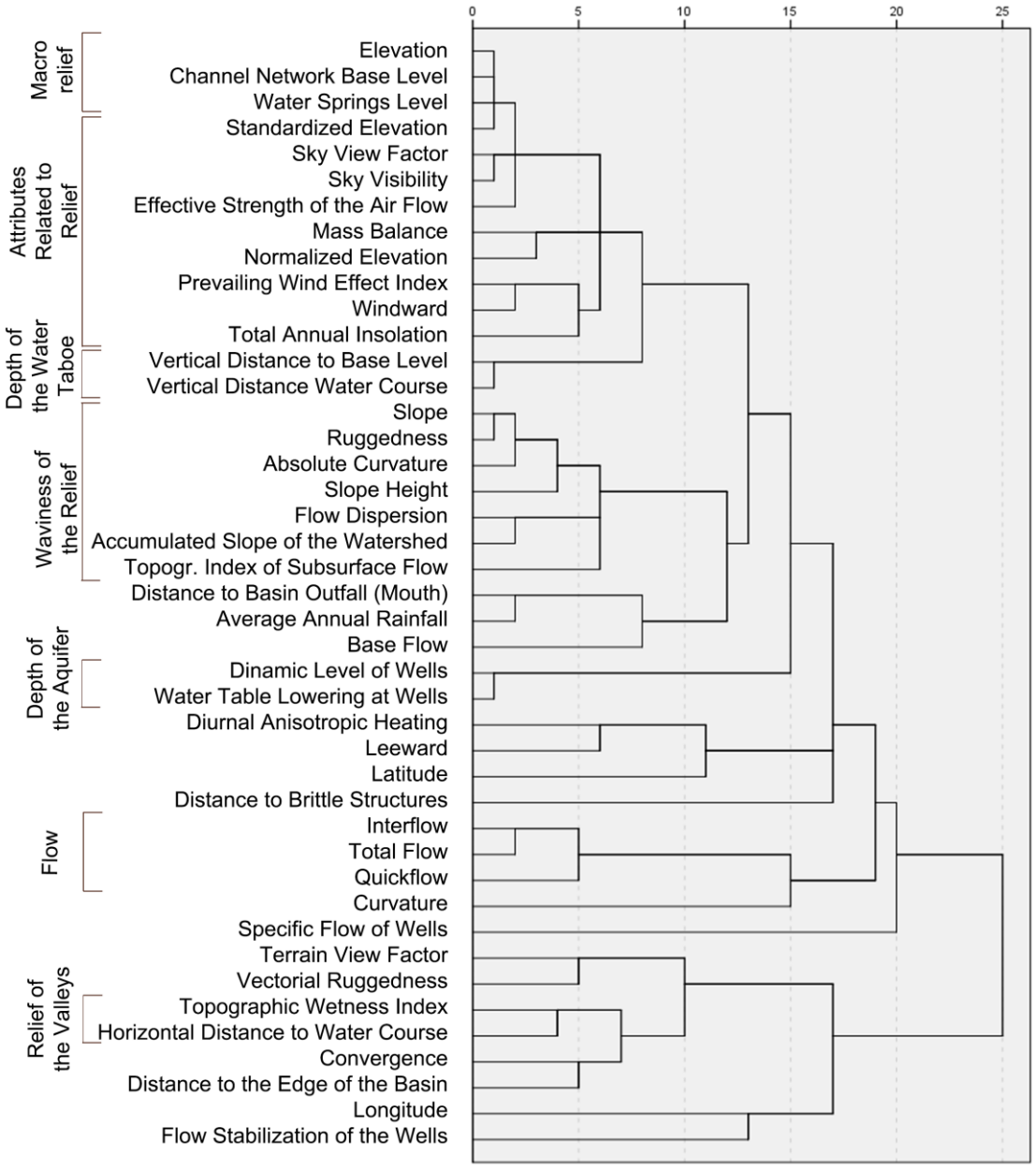

Figure 5 – Dendogram with the hierarchic clustering of used variables

Regression Model

The results of the two regression models used are shown in Table 1. It should be noticed that the model assuming regional flows had improved in all regressions in R^2 and in standard deviation, whereas there was a small decrease in Q^2 for the interflow regression. A separated analysis shows that the improvements were consistent with the hypothesis that the greater influence of the regional underground flow would be on the base flow (through the deep aquifer flow), followed by the rapid flow (through the karstic pipelines and the piston flow effect - Kirschner, 2003) and, to a lesser extent, by interflow (alluvial aquifers and local river / fracture systems), given that the latter would be delayed due to the more clayey soil in the river valleys. The analysis of the signals and weights of the coefficients was also more consistent in the model with the assumption of regional flows, and this corroborated to its choice.

Table 1 – Regression Results.

Regression	Model without regional flows			Model with regional flows		
	R^2	Q^2	Standard deviation	R^2	Q^2	Standard deviation
Total flow	0.35	-0.10	0.84	0.84	0.73	0.42
Quickflow	0.40	-0.12	0.81	0.76	0.43	0.51
Interflow	0.40	-0.04	0.81	0.43	-0.12	0.79
Base flow	0.35	-0.21	0.85	0.84	0.67	0.42

Coefficients of the independent variables

Figure 6 shows the Variable Importance in the Projection (VIP) values for the independent variables multiplied by the respective coefficients signals for each model. Detailed data model, although not discussed in this paper, can be accessed in http://pt.scribd.com/doc/127873830/tabela-graficos-pls where the real and standardized coefficients, the VIP, the respective standard deviations, the comparative charts obtained from the VIP and standardized coefficient can be found, as well as the scatter plots of the loads per components.

It should be emphasized that small VIP and standardized coefficients values do not mean that the independent variable has an insignificant role in the operation of the water systems. After all, there may be non-linear relationships not captured by the statistical model, which could be later explained with more sophisticated models. For example, if one of the independent environmental variables has an exponential or logarithmic relationship with a flow component, such variable would not be optimally measured by this model.

Analyzing the coefficients, it is noticeable the strong positive influence of the elevation slope and the elevation toward the rivers (micro relief) on the total flow and base flow regressions. This relationship can be based on the microclimatic predisposition for orographic rainfalls, and to a greater predisposition of water infiltration on hilltops.

The diurnal anisotropic heating had a positive coefficient, playing a significant role on total flow and acting even more importantly on the base flow. Hence the model corroborates the microclimatic hypothesis that slopes with increased exposition to solar radiation would have higher evapotranspiration, contributing less to infiltration, and therefore with less flow into the water system.

The regression on the base flow kept a consistent positive influence of variables related to the macro relief (e.g., absolute altitude, springs and base level, as well as rainfall and distance to the outfall), to the meso relief (e.g. altitude standardized), and to the micro relief (e.g. slopes height, mass balance, normalized altitude, altitude over the base level, elevation above the river, horizontal distance in relation to rivers), indicating that the higher areas would be most important for groundwater recharge. In the regression on the quickflow, this group of variables appeared to be reversed, which shows that such variables influence the separation between rainwater runoff and seepage. It is quite expressive that the closest areas to the rivers (both, horizontal and vertical) as well as the shortest vertical distance to the baseline level contribute with the largest quickflow, which is consistent with the idea that the soil in these fluvial valleys saturates more quickly during precipitation, consequently, displacing the flow directly to the runoff. However, despite the inverted coefficients between the quickflow and the base

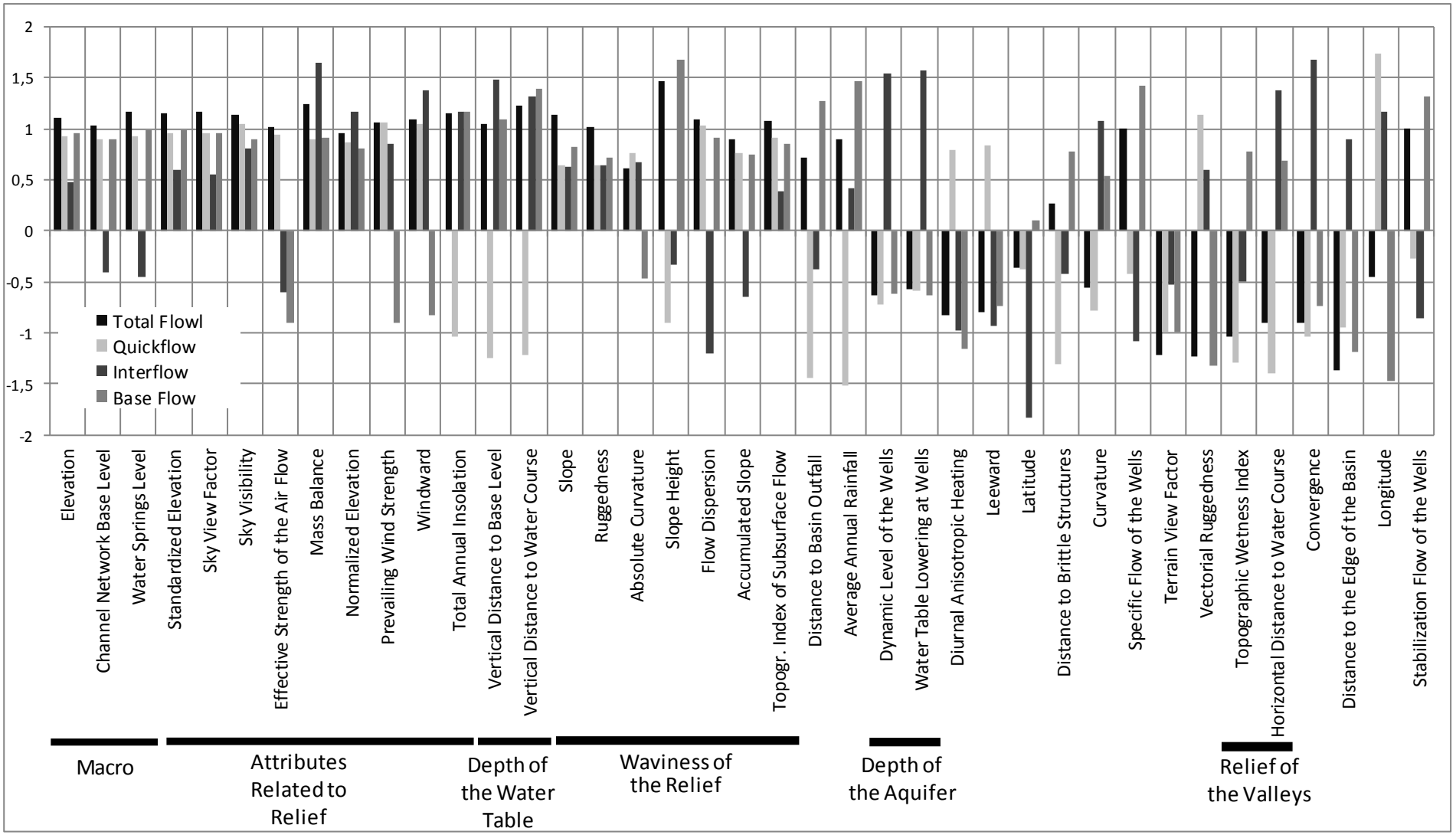

Figure 6 – Variable Importance in the Projection (VIP) values of the independent variables, multiplied by the signal of the respective c oefficients.

flow, the variables weight was not symmetrical. Furthermore, the total flow regression to evaluate the weight shows that the influence of the variables does not happens only in flow separation, but also in the overall contribution of water to the water system as a whole.

Also, it is remarkable how the base flow from the stabilization flow and the specific flow rate of the wells had positive influence, which shows that aquifers with higher flow contribute to both wells and water courses. These coefficients were negative for the quickflow, showing the influence of the flows in the division, although they have been clearly positive in the computation of total flow.

With respect to the density of fractures, the less dense areas (i.e., greater distance to the fractures) were shown to be more favorable to the base flow, while the denser surface areas were favorable to quickflow, and this variable had no influence on the total flow. This contrast can refer to the areas of porous aquifers (and, consequently, to the formation of sandier and more drainable soils, but with fewer fractures), where the deeper infiltration can be found, while areas of fractured aquifers have a greater tendency to runoff towards the rivers. Furthermore, in the case of the fractured-karstic areas, the systems of fractures may contain preferred ducts directing quickly the waters of the rivers.

Regarding the interflow, despite the lower explanatory power of the model, the convergence (more concave valleys) in areas with greater mass balance (highland areas of the micro relief) correspond to areas of greatest contribution to this flow. Furthermore, there is also an interesting positive influence related to how deep is the lowering depth of the wells, which is typical in local confined aquifers: this influence may lead to the hypothesis that the confining aquifer can also prevent the local surface percolation location, redirecting it to the rivers as interflow.

The significant influence of strictly spatial variables such as latitude (for interflow), longitude (for base flow and quickflow, in an inverted and nearly symmetrical mode) and the distance to basin edges (for total flow and base flow) may indicate that there are important environmental processes that have not yet been adequately captured in the regression model, but whose correlation has spatial orientation. For example, it is noteworthy what CETEC-MG (1981) remarks about areas close to the basin boundaries: Beyond the expected effect of the orographic rainfalls, the aquifers might contain wider and deeper fractures due to their more active tectonic history. Because of these fracture attributes, aquifers would have higher water flow capacity, including the possibility of receiving groundwater from other catchments in recharge areas of surrounding highlands (the central highlands of Urucuia and San Marcos neighbor watersheds) in cases where the adjacent hydrogeological basin does not exactly match the watershed, as suggested by Martins Junior (2009) for Paracatu Basin.

Streamflow regionalization

The regressions also made it possible to regionalize the flow data for Paracatu river basin, using the same cartographic databases. Table 2 presents the regression prediction data and the adjustment of the prediction deviation as to the last station upstream of the outfall. The sum of the components (Quickflow, Interflow and Base Flow) proved to be consistent with the predicted volume of total flow, corroborating to the consistency of the statistical model.

Table 2 – Regionalization of streamflows and respective components for the Paracatu River Basin.

	Predicted		**Corrected with prediction deviation for the last station upstream**	
	Specific Flow ($m^3.s/km^2$)	Annual Avg, Flow ($m^3.s$)	Specific Flow ($m^3.s/km^2$)	Annual Avg, Flow ($m^3.s$)
Total flow	4.31	194719.87	4.41	199070.7
Quickflow	1.48	66703.57	1.50	67917.31
Interflow	1.19	53411.52	1.10	49518
Base flow	1.67	75184.58	1.81	81797.97
Sum of Components	4.33	195299.67	4.41	199233.3

Cartographic analysis

The maps with the redistribution of coefficients on dependent variables are shown in Figure 7.

As the maps were calibrated with a regression based on mean values for each river basin, it is likely that high standard deviation are noticed for the most extreme values by the time the coefficients are computed into the maps. Because it is an exploratory model, the cartographic results for each grid cell stand as benchmarks for comparison of the relative differences within the basin, and not as an absolute reference of flow. In these terms the statistical evaluation is done for the average accuracy of the model, but not to the accuracy of each given place. Therefore, it is more advisable to interpret the maps as an indicative of the most favorable areas for aquifer recharge, rather than rely on the absolute value of the specific flow of each raster cell grid.

It is observed that in areas that present geomorphological compartments of micro wavy relief, the specific flow values have the highest spatial heterogeneity. This is mainly due to the diversity of geotopes in wavy areas, with multiple combinations of convergences, slopes, roughness, and slope positions (exposure to solar radiation and winds), and curvature at each point of the hills and between one hill to the next. The quickflow map was most affected by the wavy landscape, followed by interflow, total flow and, lastly, by the base flow.

The raster cell grids trespassed by the water courses have shown the lowest contribution to the base flow and interflow, especially in the case of the enclosed valleys, but showed the higher values for quickflow and total flow. This spatial arrangement is consistent with the assumption that the rainfall on rivers or on the immediately adjacent area is almost all transformed into in quickflow, which is more accentuated the steeper is the valley (such as v-valleys).

Hypotheses of regional flows

Figure 8 shows the maps with the difference between observation and prediction, as well as the hypotheses of regional flows for the balance of the model. The largest prediction deviations refer to sections upstream from stations 4298000 (Porto Alegre, in the Lower Paracatu River), 4254000 (Santo Antônio do Boqueirão, in Middle Preto River) and 4229000 (BR-040 Bridge in Lower Prata River) Therefore, the specific flow predicted for such areas should be examined on the map of Figure 7 with cautious.

The proposed regional flows between each section present considerable uncertainty, since it is not possible to define to what extent the discrepancy between observed and predicted flow refers to the inaccuracy of the model or to the real existence of these flows. For example, the incorporation of new cartographic variables could explain part of the difference of flows, helping to partially clear out some of them. Nevertheless, the proposal of regional flows serves as an exploratory tool for addressing future hydrogeological researches.

It is interesting to noticed how the estimated regional flows increase from the tributaries of the middle Paracatu (central basin area) up to the low Paracatu river, later emerging to the watercourse after the station 42930000 (Porto do Cavalo). Flows are also significant between stations 42490000 and 42540000 (Middle Preto River) and between stations 42350000 and 42365000 (Lower Prata River). It should be noticed that the larger regional flows occur in areas of porous shallow lateritic detritus aquifers receiving waters from fractured-karstic aquifers, as inferred from the interpretation of lithostratigraphic aquifers in Vasconcelos et al. (2012). As to the possibilities of capturing recharge areas outside the basin, the flow from Alto Rio Preto (1326 m^3.s, adjacent to the basins of the São Bartolomeu, Paraná e Urucuia rivers) and Upper Prata river (1507 m^3.s adjacent to Alto Parnaiba Basin) are noteworthy.

CONCLUSIONS

The hierarchical cluster analysis using a dendrogram was useful to indicate the spatial multicorrelation among the environmental and hydrological attributes of the water basin. The PLS regression, on its turn, proved to be effective in dealing with this multicollinearity. The hypothesis of regional flows increased the explanatory power of the model and, in general terms, decreased the risk of the overfitting. Based on the influence of the independent variables, it was possible to draw hypotheses about the hydrogeological, microclimate and geomorphic processes that operate on the water basin systems.

The combination of the coefficients for the generation of the maps of water system recharge has shown the spatial diversity of this phenomenon on the Paracatu River Basin, also regionalizing the specific flow to the area not

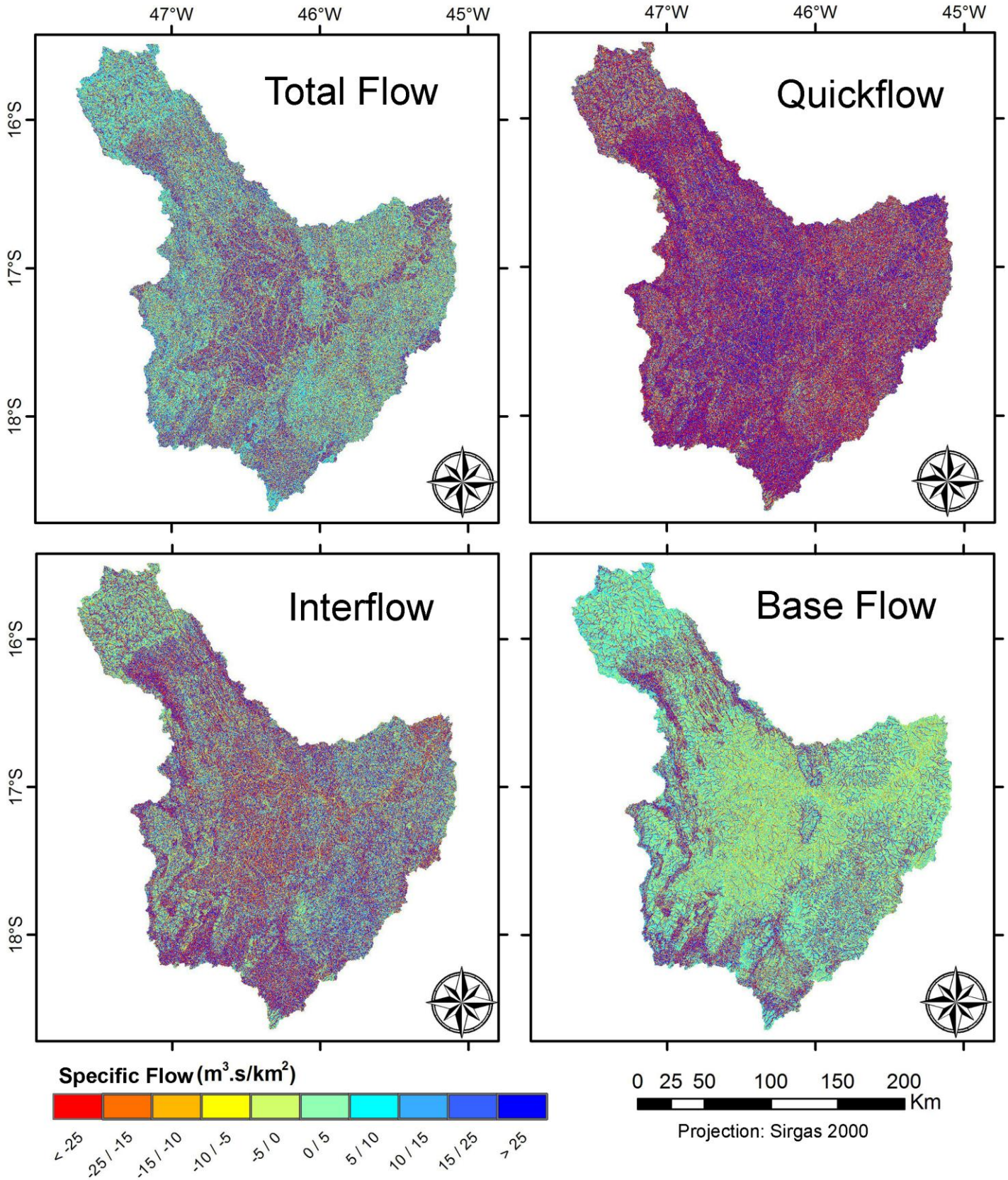

Figure 7 – Specific streamflow maps of the Paracatu River Basin.

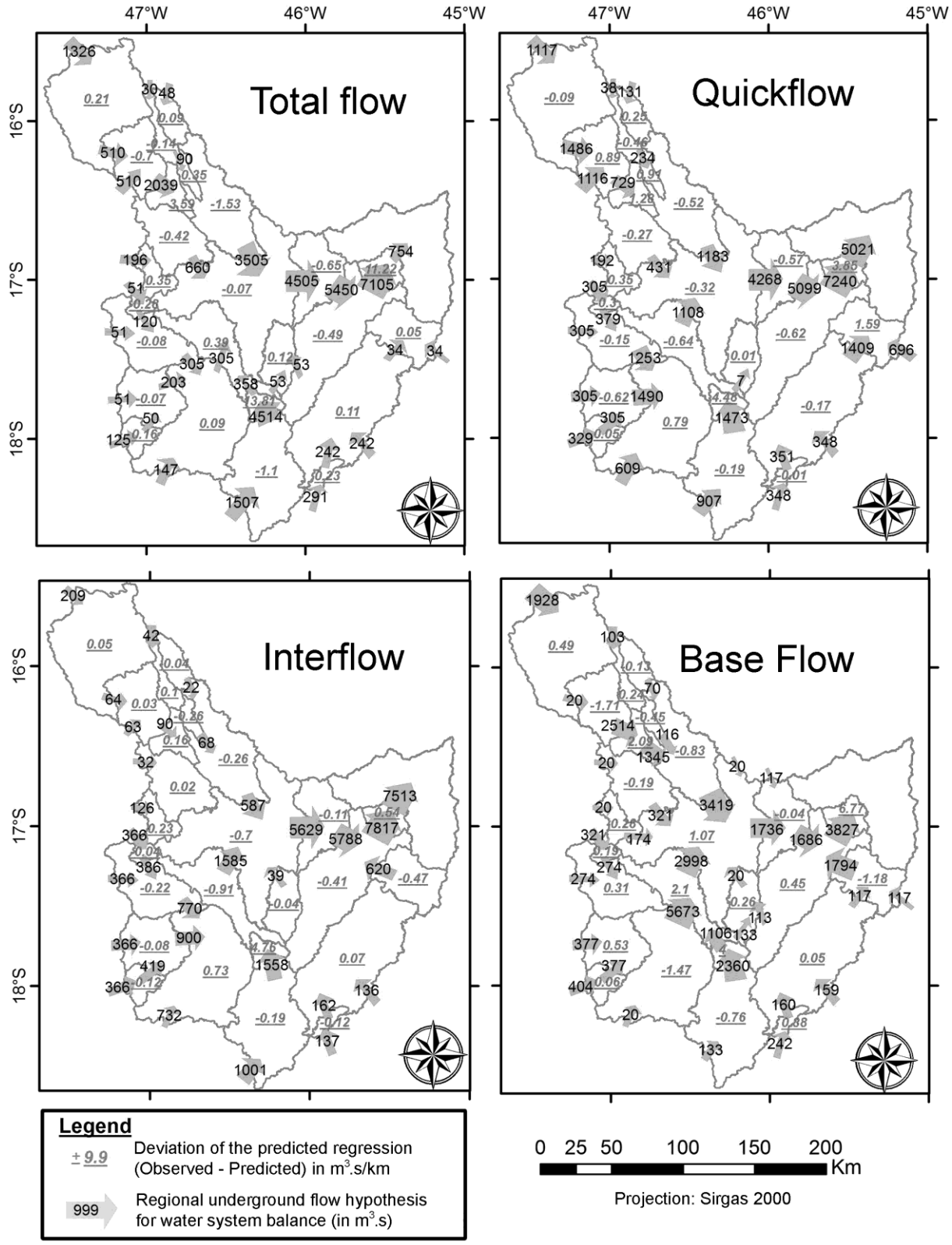

Figure 8 – Maps showing the predicted deviation and the hypotheses of regional groundwater flow for the total stream flow and for each flow component.

covered by the gauging stations. The map of the prediction deviation assisted in the analysis of the predictions showing the distance between the original cartographic model and the original measurement in each section of the basin. The mapping of regional flows for the model's balance brought up hints about possible significant groundwater flows.

The products of the proposed methodology offer useful information to the environmental policy instruments involved in territorial plans, such as master plans and zoning. The paper presents a flexible methodology, to can be adjusted on a case-by-case basis to the availability of cartographic data for the studied region. This would facilitate its replication for water basins in developing countries, where there is a great lack of systematic environmental data.

As to the use of cartographic products, it is important to notice the need to keep water and soil conservation in the most relevant areas for recharging the base flow, in order to reduce the water conflicts over the use of water during the drought periods. On the other hand, the quickflow map can point out the most interesting areas to reserve and regularize the water flow, keeping it when it rains and aiming at its multiple use, or even release during the dry season. In the case of basins with flooding problems, the quickflow maps would gain even greater importance for the sustainable management of the land use.

As a suggestion for future studies, we recommend the application of the regression methodology for basins with more gauging stations in order to increase the sample population. Only by doing this the expansion of the sampling universe would increase its reliability of predictions through the PLS method (HUI and WOLD, 1982; MARCOULIDES and SAUNDERS, 2006), decrease the risk of overfitting, and also would become feasible the selection of new PLS components, even with smaller coefficients (CHIN and NEWSTED, 1999). In the case of a sampling universe greater than 100 sample cases, it would be possible to move from PLS techniques to the more robust Structural Equation Modeling (SEM) and, thus, to go beyond the exploratory drilling by making use of confirmatory statistical indicators of fit (goodness of fit) regarding the measurement model (ROUSE and CORBITT, 2008; GARSON, 2010).

Moreover, it is emphasized that one of the weaknesses of the regression methodology associated with the hypothesis of regional groundwater flows is the overlapping areas between gauging stations, because it brings forth a partial violation of the independence assumption among samples. Thus, if two gauging stations are close to each other, sharing very similar areas, this shared area will have a greater effect on the outcome of the dependent variables coefficients, besides the influence it will also have on the R^2 coefficient. In such cases, the Q^2 and the residual standard deviation indexes, being based on resampling techniques, will be crucial to attest the quality of the model. In future studies, the development of a system of weights for each observed basin could be carried out in order to balance such area overlapping.

As a final caveat, it must be recognized that the components of separation of the hydrograph are extracted only based on patterns of pulses waveform representing of stream flow pulse and, therefore, they do not address directly the traceability of the flow components. Thus, the results of automated methods are related to the hydrogeological processes only to the extent that respectively, the base flow, the interflow, and the quickflow actually correspond to the underground, subsurface, and surface waterflows.

Moreover, the assessment of the reliability of the results cannot give up of a more comprehensive understanding of the hydrological, hydrogeological and climatic phenomena of the watersheds to be analyzed. Every information acquired by tracers, fluid balances, geophysical prospection and groundwater reserves estimates would bring additional data to evaluate the hypotheses previously raised by the PLS modeling.

ACKNOWLEDGMENTS

We address our heartful thanks to FAPEMIG (Research Support Foundation of Minas Gerais), CAPES (Coordination for the Improvement of Higher Education Personnel), CNPq (National Counsel of Technological and Scientific Development), and FINEP (Funding Agency for Studies and Projects), for funding the required researches that made feasible the performance of this work.

BIBLIOGRAPHY

ANDREEV, P., HEART, T., MAOZ, H., e PLISKIN, N. Validating Formative Partial Least Squares (PLS) Models: Methodological Review and Empirical Illustration, **Thirtieth International Conference on Information Systems**, Phoenix, Arizona, p. 1-17. 2009.

ARLOT, S. e CELISSE, A. A survey of cross-validation procedures for model selection. Statistics Surveys. Vol. 4, 40–79. ISSN: 1935-7516. 2010.

ARRAES, T. M. **Proposição de Critérios e Métodos para Delimitação de Bacias Hidrogeológicas.** 2008. 125p. Dissertação de Mestrado. Instituto de Geociências – UNB. Brasília, 2008.

BABYAK, M.A. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. **Psychosomatic Medicine. Statistical Corner**. American Psychosomatic Society. 66:411–421, 411. 2004.

BARCLAY, D.W.; HIGGINS,C., e THOMPSON, R. The partial least squares approach to causal modeling: Personal computer adoption and use as illustration. Technology Studies, 2(2), p. 285–309. 1995.

BARNES, B.S. The structure of discharge recession curves. **Transactions of the American Geophysical Union**, 20: 721-725. 1939.

BRANDÃO, R.L. e GOMES, F.E.M. Técnicas de geoprocessamento e sensoriamento remoto aplicadas na avaliação do potencial hidrogeológico da Folha Irauçuba. **Revista de Geologia,** Fortaleza, v. 16, n. 1. 2003.

CETEC-MG. FUNDAÇÃO CENTRO TECNOLÓGICO DE MINAS GERAIS. **II Plano de Desenvolvimento Integrado do Noroeste Mineiro: Recursos Naturais.** Belo Horizonte. 1981.

CHIN, W.W. PLS is to LISREL as principal components analysis is to common factor analysis. **Technology Studies**, 2, 315-319. 1995

CHIN, W.W. How to write up and report PLS analyses. In: V. ESPOSITO VINZI et al. (eds.), **Handbook of Partial Least Squares**: Concepts, methods, and applications. Springer Handbooks of Computational Statistics, p. 655-690. 2010.

CHIN, W. W., e NEWSTED, P. R. Structural Equation Modeling Analysis with Small Samples Using Partial Least Squares. In: **Statistical Strategies for Small Sample Research**, R. H. Hoyle (ed.), Sage Publications, Thousand Oaks, CA, p. 307-341. 1999.

FERRERO, V.O. Hidrología Computacional y Modelos Digitales del Terreno: Teoría, práctica y filosofía de una nueva forma de análisis hidrológico. 2004. 391 p. Available at: http://www.gabrielortiz.com/descargas/Hidrologia_Computacional_MDT_SIG.pdf, access in 4/16/2012.

FLYNN, R.H. e TASKER, G.D. **Generalized Estimates from Streamflow Data of Annual and Seasonal Ground-Water-Recharge Rates for Drainage Basins in New Hampshire.** USGS Scientific Investigations Report 2004-5019. New Hamphshire, 2004. 72 p.

GARSON, G.D. **Statnotes**: Topics in Multivariate Analysis. North Carolina State University. College of Humanity and Social Sciences, 2010. Available at: http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm, access in 12/13/2010.

GEBREHIWOT, S. G., ILSTEDT, U., GÄRDENAS, A. I., and BISHOP, K.: Hydrological characterization of watersheds in the Blue Nile Basin, Ethiopia, Hydrol. Earth Syst. Sci., 15, 11-20, Hess, 15-11-2011.

GONZALES, A.L.; NONNER, J.; HEIJKERS, J.; UHLENBROOK, S. Comparison of different base flow separation methods in a lowland catchment. **Hydrology and Earth Sciences Discussions**, 6, 2483-3515, 2009.

GOMES, F.E.M. Geoprocessamento em Ambiente SIG aplicado à Hidrogeologia. In: FEITOSA, A.C. (org.) **Hidrogeologia: conceitos e aplicações.** 3 ed. rev. e ampl. Rio de Janeiro: CPRM: LABHID. 2008.

HAELEIN, M.; KAPLAN, A.M. A Beginner's Guide to Partial Least Squares Analysis. **Understanding Statistics**, 3(4), 283–297, 2004.

HAIR JUNIOR, J.F.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E.; TATHAM, R.L. **Análise Multivariada de Dados.** 6ª edição. Porto Alegre: Bookman, 2009. 688p.

HAIR JUNIOR, J.F.; CHRISTIAN, M.R.; SARSTEDT, M. PLS-SEM: Indeed a Silver Bullet. **Journal of Marketing Theory and Practice.** Vol. 19, nº. 2, spring, 2011. p. 139-151. DOI: 10.2753/MTP10069-66791900202.

HAWKINS, D.M. The Problem of Overfitting. **Journal of Chemical Information and Computer Sciences**, 44, 1-12. 2004.

HELLWEGER, F.; MAIDMONT, D. 1997, AGREE-DEM **Surface reconditioning system**. University of Texas. 1997. Available at: http://www.crwr.utexas.edu/gis/gishyd98/quality/agree/agree.htm, access in 10/27/2003.

HENSELER, J.; RINGLE, C. M.; SINKOVICS; R. R. The use of partial least squares path modeling in international marketing. New Challenges to International Marketing, Advances in International Marketing, Volume 20, 277–319, 2009.

HOLTSCHLAG, D.J. A generalized estimate of ground-water-recharge rates in the Lower Peninsula of Michigan. U.S. Geological Survey water-supply paper 2437. 1997. 44p.

HUI, B. S., e WOLD, H. Consistency and Consistency at Large in Partial Least Squares Estimates, in **Systems Under Indirect Observation, Part II,** K. G. Jöreskog and H. Wold (eds.), North Holland, Amsterdam, 1982, p. 119-130.

INSTITUTO MINEIRO DE GESTÃO DAS ÁGUAS – IGAM. **Plano Diretor de Recursos Hídricos do Rio Paracatu: Resumo Executivo. Governo de Minas Gerais.** Comitê da Sub-bacia Hidrográfica do Rio Paracatu. Belo Horizonte: Instituto Mineiro de Gestão das Águas. 2006. 384p.

INTERAGENCY ADVISORY COMMITTEE ON WATER DATA. **Guidelines for determining flood-flow frequency**, Bulletin 17B of the Hydrology Subcommittee, Office of Water Data Coordination: U.S. Geological Survey, Reston, Va., 1982, 183 p.

KHEORUENROMNE, I.; SUDDHIPRAKARN, A.; KANGHAE, P. Properties, Environment and Fertility Capability of Sandy Soils in Northeast Plateau, Thailand. Kasetsart J. (Nat. Sci.) 32 : 355 - 373 (1998).

KIRCHNER, J.W. 2003 A double paradox in catchment hydrology and geochemistry. **Hydrological Processes**, 17, 871–874.

KOOISTRA, L.; WEHRENS, R.; BUYDENS, L.M.C.; LEUVEN, R.S.E.W.; NIENHUIS, P.H. Possibilities of soil spectroscopy for the classification of contaminated areas in river floodplains. **Journal of Applied Earth Observation and Geoinformation** (JAG) 3/4: 337-344, 2001. (verschenen in 2002).

LATUF, M.O. **Mudanças de Uso do Solo e Comportamento Hidrológico nas Bacias do Rio Preto e de Entre-Ribeiros.** Dissertação (Mestrado). Universidade Federal de Viçosa, Programa de Pós-Graduação em Engenharia Agrícola. 2007. 103 f.

LYNE, V. e HOLLICK, M. Stochastic time-variable rainfall-runoff modeling. **Institute of Engineers Australia National Conference.** Pub. 79/10, 89-93, 1979.

LYNSLEY R.K.; KOHLER M.A.; PAULHUS, J.L.H.; WALLACE J.S. **Hydrology for Engineers**, 2th edition. McGraw Hill, New York. 1975.

MARCOULIDES, G.A.; SAUNDERS, C. PLS: A Silver Bullet? MIS Quarterly Vol. 30 No. 2, pp. iii-ix/June, 2006.

MARTINS JUNIOR, P. P. (coord.). **Projeto CRHA - Conservação de Recursos Hídricos no âmbito de Gestão Agrícola de Bacias Hidrográficas.** MCT/Finep/CT-Hydro 2002-2006. Relatório Final em 2006.

MARTINS JUNIOR, P. P. (coord.). **Projeto GZRP - Gestão de Zonas de Recarga de Aquíferos Partilhadas entre as Bacias de Paracatu, São Marcos e Alto Paranaíba.** CETEC/FAPEMIG - 2007-2009. Relatório Final em 2009.

MARTINS JUNIOR, P. P.; VASCONCELOS, V.V. Comentários à Legislação sobre Águas em Correlações ao Uso, Outorga, Conservação e Preservação. Nota Técnica nº 15, 2005, 242p. In: MARTINS JUNIOR, P. P. (coord.). Projeto CRHA - Conservação de Recursos Hídricos no âmbito de Gestão Agrícola de Bacias Hidrográficas. MCT/Finep/CT-Hydro 2002-2006. Relatório Final em 2006.

MULHOLLAND, D.S. **Geoquímica Aplicada à Avaliação de Qualidade de Sistemas Aquáticos da Bacia do Rio Paracatu (MG)**. Dissertação de Mestrado. Brasília-DF: IG/UNB. 2009. 95p.

NAGHETTINI, M. e PINTO, E.J.A. **Hidrologia Estatística**. Belo Horizonte: CPRM. 2007. 552p.

NOBRE, S.A.S. O algoritmo PLS Model. Dissertação de Mestrado. Instituto Superior de Estatística e Gestão da Informação. Universidade Nova de Lisboa. Portugal. 2006. 95p.

NUNES, H.T. e NASCIMENTO, O.B. **Base de Dados Meteorológicos.** Belo Horizonte: Nota Técnica NT-CRHA 17/2004. Em: MARTINS JUNIOR (coord.). Projeto CRHA. Memória Técnica da Fundação CETEC. 40p.

REFAEILZADEH, P.; TANG, L; LIU, H. Cross-Validation. **Encyclopedia of Database Systems**. 2009. p. 532-538. Available at: http://www.informatik.uni-trier.de/~ley/db/reference/db/c.html#RefaeilzadehTL0, access in 11/25/2012.

RETALLACK, G. J. (2008). **Soils of the past**: An introduction to paleopedology. Wiley-Blackwell. 404p.

TUCCI, C. E. M. **Regionalização de vazões**. Porto Alegre: Ed. Universidade: UFRGS, 256p. 2002.

TUCCI, C. E. M. (org.). **Hidrologia**: Ciência e Aplicação. Segunda Edição. ABRH. Universidade Federal do Rio Grande do Sul. Porto Alegre: Editora da Universidade. 2009. 944p.

UMETRICS**. User Guide for SIMCA P+**. Version 12.0.1. Kinnelon: UMETRICS. 2008.

VASCONCELOS, V. V.; MARTINS JUNIOR, P. P.; HADAD, R. M. Caracterização Ambiental da Bacia do Rio Paracatu. In: Martins Junior, P. P. (coord.) Projeto SACD – Sistemas de Arquitetura de Conhecimentos e de Auxílio à Decisão na Gestão Geo-Ambiental e Econômica de Bacias Hidrográficas e Propriedades Rurais. 2012. 84p. Available at: http://pt.scribd.com/doc/98405182/caracterizacao-ambiental-da-bacia-do-rio-paracatu, access in 4/19/2013.

VASCONCELOS, V. V.; MARTINS JUNIOR, P. P.; HADAD, R. M. Estimation of flow components by recursive filters: case study of Paracatu River Basin (SF-7), Brazil. **Geologia USP** – Série Científica, vol.13, n.1, pp. 3-24, 2013.