Evaluating the potential of Genetic Programming as an exploratory data analysis in soil science

Research · September 2015

CITATIONS
0

READS
51

2 authors:

Lorenzo Menichetti
Swedish University of Agricultural Sciences
28 PUBLICATIONS **286** CITATIONS
SEE PROFILE

Alberto Paolo Tonda
French National Institute for Agricultural Resea…
87 PUBLICATIONS **259** CITATIONS
SEE PROFILE

Evaluating the potential of Genetic Programming as an exploratory data analysis in soil science

L. Menichetti[a*] and A. Tonda[b]

[a]Swedish University of Agricultural Sciences, Department of Soil and Environment, P.O.Box 7014, 75007 Uppsala, Sweden
[b]UMR 782 GMPA, INRA, 1 Av. Lucien Brétignères, 78850, Thiverval-Grignon, France
[*]corresponding author, tel. +46768549268, e-mail: Lorenzo.Menichetti@slu.se

Abstract

Genetic Programming is a powerful optimization technique, able to deliver high-quality results in several real-world problems. One of its most successful applications is symbolic regression, where the objective is to find a suitable expression to model the underlying relationship between data points, with no aprioristic assumptions. In this paper, we propose the application of a Genetic Programming technique to a dataset on soil respiration and soil properties, in order to investigate possible influences of soil properties on soil respiration through symbolic regression. The best candidate models obtained by the technique are then studied to determine possible differences in the relationships related to environmental factors. Recurring patterns in the best solutions proposed by the search algorithm are identified, and the suitability of symbolic regression in soil science is evaluated and discussed. Genetic Programming proves to be an extremely promising data mining technique for soil scientists, as it is able to uncover relationships that could otherwise remain hidden, while remaining completely neutral and bias-free. We suggest its application for routine data analysis, as the technique presents particular interest for environmental modeling and development of pedotransfer functions.

1. Introduction

As new field methods are developed, making field measurements cheaper and denser, and new studies are published, the amount of data available to the scientific community grows more that linearly over time. An unprecedented amount of data is now at disposal of ecosystem scientists, and there is a need for methods able to treat it in a comprehensive and objective way. This objective involves the use of new algorithms and data mining procedures, as the field slowly adopts more and more automatic processes.

Semi-empirical relationships are widely exploited in soil science. For example, it is common to predict soil properties, which would be too costly or difficult to estimate otherwise, through the use of pedotransfer functions (Bouma, 1989) that exploit easily measurable variables. Because of the high global concern for climate change and the emission of greenhouse gases in the atmosphere, another variable that received a lot of attention in the last decades is soil respiration. Its relation with different soil edaphic properties and soil processes is nevertheless still not completely clear. While it is widely accepted that soil respiration is linked with temperature and moisture conditions (see Lloyd & Taylor, 1994 and Moyano *et al.*, 2011 for some examples), there is a lack of understanding on how site-specific properties can modify these relationships in the field. Part of the observed error in these relationships is probably accountable to yet unknown links between soil respiration and environment.

One possible approach to the issue is to explore the space of possible dependencies between elements, while being as unbiased as possible towards the shape of the solution. The rise in complexity of available data has led the machine learning community to develop refined methods able to uncover relationships between variables in huge datasets. For natural laws, evolutionary-based computation has been successfully used to detect hidden dependencies, especially in field of physics (Schmidt & Lipson, 2009). While the expertise of human scientists is irreplaceable, machine learning can be exploited to obtain a large number of *candidate solutions*, that is, equations proposing a connection between variables.

Evolutionary algorithms have been sometimes applied to soil science problems for the development of pedotransfer functions (Crowe et al, 2006, Padarian et al, 2012) and more often to hydrology problems (Johari et al., 2011, Pedroso et al, 2011), but the potential of the technique in soil science is still largely unexplored. In this paper, we propose to apply a state-of-the-art evolutionary algorithm to a real-world dataset obtained by crossing two freely available datasets on soil respiration and soil properties. The most promising solutions obtained through the automatic approach are then examined, and recurring patterns are found, hinting at possible

strong, uncovered relationship between variables. While further experiments are required to draw more definite conclusions, preliminary results show great promises for the coupling of automatic approaches and human expertise.

2. Material and Methods

2.1 Evolutionary Algorithms and Symbolic Regression

The term *Evolutionary Algorithms* (EAs) groups a great variety of bio-inspired stochastic meta-heuristics for optimization, loosely inspired by the paradigm of Neo-Darwinian natural evolution. In EAs, an *individual* is defined as a candidate solution for a given problem. A *population* of solutions is randomly created, and then evaluated with a *fitness function*, that examines their efficacy with regards to a target problem. The fittest individuals are then selected for *reproduction*, usually performed by slightly altering some elements of the solution (*mutation*) or by mixing the information contained in two individuals (*crossover*). The result of the reproduction step is a new generation of candidate solutions, which are subsequently evaluated with the fitness function. The worst individuals are removed from the population, and the loop resumes from reproduction, until a user-defined *stop condition* is reached.

After the seminal work on *Genetic Algorithms* (Holland, 1975) carried on by Holland during the 60s, where solutions are modeled as bit strings, other independent research lines led by Fogel and Schwefel gave birth to *Evolutionary Programming* (Fogel, 1962) and *Evolution Strategies* (Schwefel, 1965), powerful algorithms focused on real-value optimization. At the beginning of the 90s, John Koza presented *Genetic Programming* (GP) (Koza, 1992), an EA whose individuals are modeled as trees: the expressive power of this idea made it possible to approach extremely complex problems, where the shape of a solution could range from a network layout to a complete Assembly-language program.

Thanks to the development of GP, the EA community tried to answer to the pressing practical need for improved forms of scientific data mining (Clery & Voss, 2005 and Valdés-Pérez, 1999) with the *symbolic regression* technique. In symbolic regression, the objective is to find a mathematical expression linking variables' values in a dataset, without making assumptions on the structure of the expression itself. Candidate equations to solve the problem are modeled as trees, while the fitness function usually aims at minimizing the absolute or squared difference from experimental data. From the first promising results (Koza, 1992), a research line led by Schmidt and Lipson produced an extremely efficient GP-based algorithm (Schmidt & Lipson, 2009), able to deliver high-quality solutions in small amounts of time. The derived software, *Eureqa Formulize* (http://formulize.nutonian.com/, accessed on 25 September 2013), is now considered the state of the art in the field.

2.2 The dataset

In this study, we use data from the updated soil respiration database (SRDB) (Bond-Lamberty & Thomson, 2012): in particular, we consider variables describing soil respiration, mean annual temperature and mean annual precipitations. Latitude and longitude specified in the corresponding study are used for combining this dataset with the harmonized world soil database (HWSD) (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). Topsoil and subsoil gravel, sand, silt and clay content, topsoil and subsoil pH and cation exchange capacity (CEC), topsoil and subsoil soil organic carbon (SOC) content are taken from the HWSD, while soil respiration data and the environment ecologic identification are obtained from the SRDB.

The target variable for the study is soil respiration, normalized by the SOC content in the topsoil, as the latter is already known to explain most of the observed variation.

In order to improve the effectiveness of the search algorithm, outliers outside two times the interquartile range are removed. Data are then normalized, so that each variable has mean 0

and variance 1, and then multiplied by 100 in order to obtain a medium magnitude. The dataset is then randomly divided between a training set (80% of the samples) and a validation set (20% of the samples).

2.3 Data treatment

The target expression is:

$$R_{norm} = f(lat, long, T, P, Gravel_{tops}, Sand_{tops}, Silt_{tops}, Clay_{tops}, BD_{tops,} pH_{tops}, CEC_{tops},$$

$$Gravel_{subs}, Sand_{subs}, Silt_{subs}, Clay_{subs}, BD_{subs}, pH_{subs}, SOC_{subs}, CEC_{subs})$$

where *tops* denotes topsoil and *subs* denotes subsoil. The term R_{norm} denotes soil respiration, normalized by topsoil SOC content (in g C m^{-2} per unit % of SOC content), the term *T* the mean annual air temperature (in ° C). The two terms *lat* and *long* denote latitude and longitude, respectively. The terms *Gravel, Sand, Silt, Clay* and *BD* denote gravel, sand, silt and clay percentage and bulk density, respectively. The term *pH* denotes the soil pH measured in H_20, the term *SOC* denotes the soil organic carbon content in percentage and the term *CEC* denotes the cation exchange capacity in cmol kg^{-1}. The following basic functions are used as building blocks during the GP search: *constant, integer constant, input variable, addition, subtraction, multiplication, division, negation, sine, cosine, tangent, exponential, natural logarithm, factorial, power, square root, minimum, maximum, modulo, floor* and *ceiling*. After the search, the ten best solutions proposed by the software are tested against the validation dataset, and residuals for each point are computed. Residuals are then plotted, divided by ecosystem group.

As a measurement of the fit of the possible models, we consider the following indicators: mean error (ME), mean absolute error (MAE), root mean squared error (RMSE), normalized root mean squared error (NRMSE), percent bias (PBIAS), Nash-Sutcliffe Efficiency (NSE), index of agreement (d), Pearson's correlation coefficient (r) and coefficient of determination (R^2).

Candidate solutions are also visually compared through a principal component analysis (PCA) (Venables & Ripley, 2002) on the residuals.

The machine used to run the search is a 64-bit workstation with 64 GB of RAM, mounting 2 Intel Xeon 2-Ghz E5-2650 processors, using a total of 16 cores and 32 threads. The software used for the experiments shows several statistics to detect convergence: in this case, we observe *maturity*, a metric that describes diversity inside the population. When an EA is close to convergence, most of the candidate solutions inside the population closely resemble each other, with minimal differences between them: in such a condition, the EA is focusing on exploitation of a small part of a search space, and it is unlikely to produce dramatically different solutions. We stop the experiment when the maturity score of the population reaches 90%, after about 25 hours of computation. It is important to notice that the same results could have been achieved on a standard desktop computer in a reasonable amount of time (around one week).

3. Results

3.1 The selected candidate solutions

Our search evaluated 2.4×10^{12} solutions over approximately 8 million of generations. We selected the 10 best solutions presented by the search algorithm according to the best compromise between complexity (the size of the function) and squared error minimization. The selected solutions are the following:

$$R_{norm} = 1.79 \cdot pH_{tops} + mod\left(Clay_{tops}, 0.63\right) - 0.31 \cdot Silt_{subs} - 1.7 \cdot pH_{subs} - 0.21 \cdot T^2 \quad (1)$$

$$R_{norm} = 0.24 + 2.17 \cdot pH_{tops} + 1.60 \cdot min\left(min\left({pH_{tops}}^2, 1.56 - pH_{tops}\right), T\right) - T - 2.09 \cdot pH_{subs} \quad (2)$$

$$R_{norm} = 0.27 + 2.14 \cdot pH_{tops} + 1.62 \cdot min\left(min\left({pH_{tops}}^2, 1.12 + Clay_{subs} - Silt_{tops}\right), T\right) - T - 2.04 \cdot Ph_{subs} \quad (3)$$

$$R_{norm} = 1.65 \cdot pH_{tops} + 0.25 \cdot Clay_{tops} + 0.16 \cdot BD_{tops} - 1.68 \cdot pH_{subs} - 0.22 \cdot T^2 \quad (4)$$

$$R_{norm} = 1.70 \cdot pH_{tops} + 0.19 \cdot BD_{tops} + 2.05 \cdot Clay_{tops} \cdot max(0.12, Gravel_{subs}) - 1.73 \cdot pH_{subs} - 0.22 \cdot T^2 \quad (5)$$

$$R_{norm} = 0.41 + 1.73 \cdot pH_{tops} + 0.26 \cdot Clay_{tops} + 0.19 \cdot BD_{tops} - 1.76 \cdot pH_{subs} - 0.24 \cdot T^2 \quad (6)$$

$$R_{norm} = pH_{tops} + 0.32 \cdot Clay_{subs} + 0.21 \cdot BD_{tops} - pH_{subs} - 0.22 \cdot T^2 \quad (7)$$

$$R_{norm} = mod\left(BD_{tops} + 1.61 \cdot Clay_{tops}, 0.83\right) - 0.15 \cdot T^2 \quad (8)$$

$$R_{norm} = 0.27 + 2.14 \cdot pH_{tops} + 1.61 \cdot min\left(min\left({pH_{tops}}^2, 1.48 + 1.29 \cdot Gravel_{subs} - pH_{subs} - pH_{tops} \cdot Gravel_{subs}\right), T\right) - T - 2.03 \cdot pH_{subs} \quad (9)$$

$$R_{norm} = 1.64 \cdot pH_{tops} + mod(Clay_{subs}, 0.53) - 1.61 \cdot pH_{subs} - 0.22 \cdot T^2 \quad (10)$$

Considering the following parameters: R_{norm} T, BD_{tops}, $Silt_{tops}$, $Clay_{subs}$ and $Silt_{subs}$.

3.2 The fit of solutions by ecosystem groups

The residuals of the selected solutions look similar when the selected solutions are tested against the validation dataset (Fig. 1). In savanna ecosystems, the variation of residuals is quite high, and the obtained functions do not seem to perform well.

The only functions that seem to reproduce data in the validation dataset are Eq. 2 and Eq. 3.

The PCA analysis of the residuals (Figure 2) does not find relevant differences by ecosystem group, but helps to highlight the differences between Eq. 2, Eq. 3 and all the others.

The variables selected by the search do not include latitude, cation exchange capacity or mean annual precipitation, and all the variability is explained according to mean annual temperature, pH and soil texture. The two most performing functions, Eq. 2 and Eq. 3, do not include exponential terms for the mean annual temperature, and both are almost linear, differently from the others.

4. Discussion

4.1 The candidate solutions

All the selected equations present an R^2 value on the training dataset between 0.46 and 0.50, but do not perform accordingly on the validation dataset. Eq. 2 and Eq. 3 are the only two solutions that can be considered to explain some of the variability in the validation dataset. Both functions suggest a linear relationship between soil respiration, topsoil pH and temperature, while introducing also a small nonlinear factor for topsoil pH. The better fit of Eq. 3 seems to be related to the inclusion of soil texture in the function.

The bad fit for most of the functions on the validation dataset, together with the relatively good fit on the training dataset, can be explained considering the specificity of the constant terms proposed by the algorithm. Furthermore, in the machine learning community, there is evidence that GP models with a high degree of complexity might overfit the training set, introducing terms that increase the fitting by a minimal amount, exploiting specific characteristics of the dataset that do not generalize well (Rosca, 1996)**.** The information on the possible relationships between the data that all the selected functions carry is nevertheless potentially valuable, as many of the relationships that have been found might contain relevant information on the shape of potential dependencies between variables.

In general, the algorithm discards most of the chemical information contained in the CEC values, and retains pH as the only chemical variable. Temperature is present in all the selected functions, sometimes in a linear form and more often in an exponential form. Soil texture appears quite often, but never using coarse fractions of the topsoil as a predictor, and just seldom considering the gravel content of subsoil (that could be a proxy of other variables as water infiltration or aeration). Sand is never used, while finer fractions seem to play a role in predicting soil respiration, probably because of their interaction with soil organic matter.

4.3 Suitability of the method in the context of soil science

The symbolic regression algorithm finds several potential correlations in the dataset. The first benefit of this technique is to find hidden relationships between data in a way that is totally neutral toward the solution and carries absolutely no human bias.

Although only two of the selected solutions could be used for predictions, the main asset of the technique in our case concerns the exploration of possible relationships rather than predictions, and in this respect the technique presents a good potential. The identification of potential relationships between variables in a mathematical form and in a way that it is not biased by the beliefs of the experimenter is an invaluable asset for any model study, and might be significantly superior to traditional correlation analyses. The suggestion for possible numerical transformations contained in the best equations found by the EA can represent an important aid for modelers, although at the moment the technique should be followed by a second phase of "traditional" modeling with a human expert. We must anyway consider that the accuracy of the technique is extremely dependent on the number of generations, and therefore any increase in computing power (foreseeable in a near future on common desktop machines, or already achievable with relatively cheap infrastructures such as rented cloud grids or clusters) could increase such accuracy.

5. Conclusions

The EA-based search identifies a set of solutions performing relatively well in predicting soil respiration over the training dataset, although performances with the validation dataset are comparable only in a few cases. The selected solutions contain, nevertheless, relevant information on possible relationships between the predicted variables and all potential predictors.

The main benefit of this technique is the totally unbiased estimation of possible links between the variables. The technique explores the most promising part of all possible combinations of numerical transformations to apply on the data, inside a subset of transformation functions defined by the user. This allows for a much deeper assessment of correlations between the variables than traditional techniques of correlation analysis. Still, as an asset over other machine learning techniques, the EA-based search retains complete transparency to the user. Solutions found by symbolic regression, although not directly usable for mechanistic modeling, are a useful tool for data interpretation and could be used for the development of a more mechanistic model. We therefore advocate for the adoption of symbolic regression techniques in the early part of the routine analysis workflow of soil related datasets, as an explorative data mining technique, and particularly as an explorative method for modeling purposes and for the development of pedotransfer functions.

References

Bond-Lamberty, B.P. & Thomson, A.M., 2012. A Global Database of Soil Respiration Data, Version 2.0. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. http://dx.doi.org/10.3334/ORNLDAAC/1070

Crowe, A., Mcclean, C., & Cresser, M., 2006. An application of genetic algorithms to the robust estimation of soil organic and mineral fraction densities. Environmental Modelling & Software, 21, 1503–1507.

Koza, J. R., 1992. Genetic programming: on the programming of computers by means of natural selection. MIT Press, Cambridge, MA, USA

Clery, D. & Voss, D., 2005. All for one and one for all. Science 308, 809

FAO/IIASA/ISRIC/ISSCAS/JRC, 2012. Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Fogel, L. J., 1962. Autonomous automata. Industrial Research 4, 14-19.

Holland, J. H., 1992. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. MIT Press Cambridge, MA, USA

Johari, A., Javadi, A. A., & Habibagahi, G., 2011. Modelling the mechanical behaviour of unsaturated soils using a genetic algorithm-based neural network. Computers and Geotechnics, 38, 2–13.

Padarian, J., Minasny, B., & McBratney, A., 2012. Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system. Soil Research, 50, 443–446.

Pedroso, D. M., & Williams, D. J., 2011. Automatic calibration of soil–water characteristic curves using genetic algorithms. Computers and Geotechnics, 38, 330–340.

Lloyd, J., & Taylor, J., 1994. On the Temperature Dependence of Soil Respiration. Functional Ecology, 8, 315–323.

Moyano, F., Vasilyeva, N., Bouckaert, L., Cook, F., Craine, J., Curiel Yuste, J., Don, A., Epron, D., Formanek, P., Franzluebbers, A., Ilsted, U., Kätterer, T., Orchard, V., Reichstein, M., Rey, A., Ruamps, L., Subke, J.-A., Thomsen, I. K. & Chenu, C., 2011. The moisture response of soil heterotrophic respiration: interaction with soil properties. Biogeosciences Discuss, 8, 11577–11599.

Rosca, J. P., 1996. Generality versus size in genetic programming. In Proceedings of the First Annual Conference on Genetic Programming, pp. 381-387. MIT Press.
Schwefel, H.-P., 1956. Cybernetic Evolution as Strategy for Experimental Research in Fluid Mechanics (Diploma Thesis in German). Hermann Föttinger-Institute for Fluid Mechanics, Technical University of Berlin
Valdés-Pérez, R. E., 1999. Discovery tools for science apps. Communications of the ACM 42.11 (1999): 37-41.
Venables, W., & Ripley, B., 2002. Modern Applied Statistics with S. Springer-Verlag.

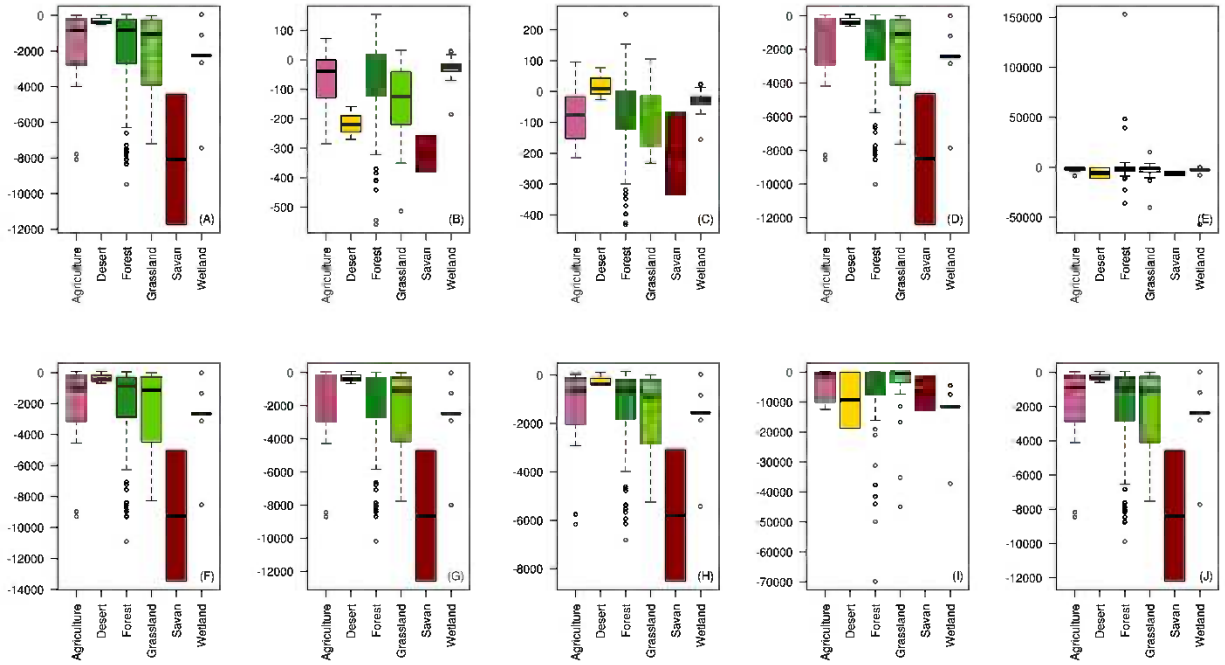

Figure 1: Residuals of the selected functions. A) Equation 1, B) Equation 2, C) Equation 3, D) Equation 4, E) Equation 5, F) Equation 6, G) Equation 7, H) Equation 8

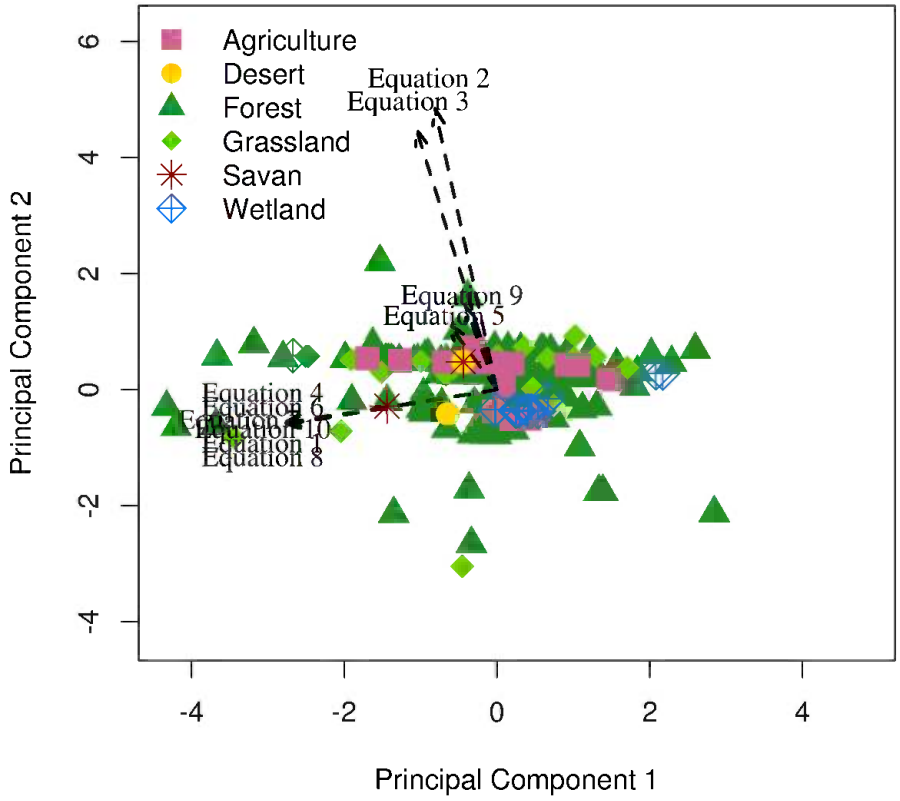

Figure 2: PCA analysis of the residuals

Tables

	Function 1	Function 2	Function 3	Function 4	Function 5	Function 6	Function 7	Function 8
ME	-2101.6	-73.6	-63.2	-2138.0	-1244.5	-2324.8	-2175.7	-1470.1
MAE	2103.4	101.7	90.4	2141.1	4906.5	2327.7	2178.3	1474.2
RMSE	3433.3	148.4	126.9	3505.4	13412.4	3811.7	3566.0	2417.6
NRMSE %	3684.7	159.3	136.2	3762.1	14394.5	4090.8	3827.1	2594.7
PBIAS %	28932.6	1012.7	869.6	49411.5	30274.3	53729.6	50283.1	33974.7
NSE	-1362.3	-1.6	-0.9	-1409.0	-20166.4	-1666.1	-1458.1	-669.7
d	0.0	0.4	0.6	0.0	0.0	0.0	0.0	0.0
r	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
R2	-2101.6	-73.6	-63.2	-2138.0	-1244.5	-2324.8	-2175.7	-1470.1

Table 1: the goodness of fit indicators considered for each function. ME = mean error, MAE = mean absolute error, RMSE = root mean squared error, NRMSE = normalized root mean squared error (-100% <= nrms <= 100%), PBIAS = percent bias, NSE = Nash-Sutcliffe Efficiency, d = index of agreement (0 <= d <= 1), r = Pearson's correlation coefficient, R^2 =

coefficient of determination (0 <= R2 <= 1). These indexes have been calculated on the whole dataset, without removing the outliers.