ΠΕΔΟ METRON

The Newsletter of the Pedometrics Commission of the IUSS

Issue 35, September 2014 **Chair: A-Xing Zhu** **Vice Chair: Dick. J. Brus**
Coordinator: Murray Lark **Layout: Jing Liu**

Inside this Issue

From the Chair

Dear Colleagues,

With this issue, I conclude my term as the chair of the Pedometrics Commission. This issue should have been published a month ago but due to my travel schedule I was not able to tally the numbers and write this summary message until very recently.

First of all, I want to thank you all for the great contributions you have made to this community during my term. In particular, I would like to thank Dick for being working closely and productively with me. Jing Liu assisted me in managing the Pedometrics website which lately became quite a task due to the technical glitches from the service provider. Jing also helped with the layout of Pedometron. I thank her for your help to me and her contribution to the Commission. I also want to thank Murray Lark for his generous help with the material collections for the first few issues of Pedometron during my term. I received enthusiastic support and great advice from the board on many issues. The board consists of (in alphabetical order of first name): Alex McBratney, Bertin Takoutsing, Bas Kempen, Bob Macmillan, Brendan Malone , Budiman Minasny, Bui Le Vinh, David Joseph Brown, Dominique Arrouays, Gerard Heuvelink, Janis Boettinger, Jing Liu, Joulia Meshalkina, Philippe LAGACHERIE, Leigh Winowiecki, Lou Mendon'a Santos, Marc Van Meirvenne, Lark Murray, Raphael Viscarra-Rossel, Pierre Roudier, Sabinee Grunwald, Thorsten Behrens, Tom Hengl, Lin Yang.

Financially, the commission now has a higher balance than when I took over (See Table 1 for details). The balance as July31 is **$7,777.24**.

Table 1: Financial details of the Commission (in US currency):

Income	Amount	Balance
Transfer from IUSS	$6,375.00	$6,375.00
Interests	$5.47	$6,380.47
Pedometrics 2013 registration fees*	$7,644.62	$14,025.09
Expenses		
Wire Transfer Fee+	$368.50	$13,656.59
Payment to Pedometrics 2013	$5,500.00	$8,156.59
Pedometrics web hosting fee	$379.35	$7,777.24

*The portion which paid through the Commission account.
+Bank fee for receiving wire transfers

Academically, the commission has been very actively. In addition to its regular Pedometrics Conferences (Pedometrics'2011 at Trest and Pedometrics'2013 at Nairobi), the commission organized a session at the Soil Carbon Conference in Madison, U.S.A. (June 3-6, 2013), a session at the Division 1 Conference in Ulm, Germany (Sept. 30-Oct. 3, 2013), two symposia at the 20th WCSS in Jeju, Korea. These activities have made this Commission as one of the most active commissions in IUSS.

Administratively, the commission continues to recognize the contributions from its members through its awards. It has completed the annual Best Paper Awards for 2010, 2011, and 2012 (See Pedometron 34

From the Chair

for the list of these papers) and awarded the 2014 Webster Medal to Dr. Gerard Heuvelink for his marvelous achievements and great contribution to the Commission (See the Citation for the award in this issue). The Commission is currently calling for nomination for the Best Paper Award for 2013 (see the call for nomination below).

Structurally, we have initiated discussions at the business meetings in Nairobi and Jeju, respectively. The following suggestions were made at these meetings: (1) Creation of a treasure position (supervised by the Chair and Vice Chair); (2) Creation of a webmaster position; (3) Financial contribution to the Commission from the Pedometrics conferences; (4) Synchronization of Pedometric conference and conferences of working groups: DSM, Proximal Soil Sensing and Soil Monitoring; (5) Inclusion of young and active Pedometricians in the decision bodies of this Commission (such as the Board and the award committee).

In summary, with your strong support I was able to complete this term successfully and hand over the commission in a good condition. Thank you for your trust, for the opportunity and for working with me!

Best wishes,

A-Xing Zhu

News and Updates

❖ **Pedometrics 2015**

Dr. Budiman Minasny

New elected Chair of Pedometrics Commission

Dear Colleagues

We are happy to announce that Pedometrics 2015 will be held in Córdoba, Spain.

The organizing committee, Tom Vanwalleghem, Ana Tarquis, Juan Vicente Giráldez, and Karl Vanderlinden invite you to the famous world-heritage town.

This Pedometrics Conference will also incorporate meetings for the IUSS WG on Soil Landscape Modelling and Soil Monitoring.

The dates are: 15-18 September 2015. Please keep these dates free!

Pre-conference workshop is on 14 Sept 2015, with activities from the IUSS WG on Soil Landscape Modelling and Soil Monitoring.

Topics include:

1. Soil-landscape modelling: mechanistic & empirical
2. Soil Morphometrics (image analysis, remote sensing, 3D soil imaging)
3. Sampling and monitoring
4. Field experimental design
5. Digital soil mapping and proximal soil sensing
6. Bayesian statistics and Hierarchical Modelling in soils
7. Fuzzy cognitive mapping
8. Soil Spatial and Temporal Scaling
9. Soil Ecosystem Services

The organizing committee is working hard to realize this conference, and we will update you with further information very soon.

Kind regards

Budi.

News and Updates

❖ Award of the Richard Webster Medal 2014

David G Rossiter
Chair Pedometrics Committee on Prizes and Awards for 2014--2017

I am pleased to report that the IUSS Richard Webster Medal corresponding period 2011-2014 was awarded at the 20th World Congress of Soil Science (June 2014) to Dr. Gerard B. M. Heuvelink, Associate Professor in the Soil Geography and Landscape group at the Wageningen University (NL), as well as a Senior researcher at ISRIC - World Soil Information. He was nominated by two colleague pedometricians (Tomislav Hengl and Dick Brus), and evaluated by the five members of the Pedometrics Committee on Prizes and Awards. Gerard amply satisfied the criteria for the award, viz:

a) "**a distinction in the application of mathematics or statistics in soil science through their published works**":

Gerard has co-authored 40 ISI papers between consecutive World Soil Congresses. Among his most influential publications, based on Google Scholar, are:

- "A Propagation of Errors in Spatial Modelling with GIS" (Citation rate: 32/year)
- "A generic framework for spatial prediction of soil variables based on regression-kriging" (CR: 38/year)
- "Optimization of sample patterns for universal kriging of environmental variables" (CR: 20/year)
- "Modelling soil variation: past, present, and future" (CR: 17/year)

The latter is a comprehensive review article from 2001, written with Richard Webster, that has served as guide for soil scientists trying to apply pedometric techniques to soil geography

a) "**innovative research in the field of pedometrics**":

He is so recognized by his peers in Pedometrics, e.g., by the 2006 Best paper award in Pedometrics:

- Heuvelink G.B.M., Schoorl J.M., Veldkamp A., Pennock D.J. 2006. Space-time Kalman filtering of soil redistribution. *Geoderma* 133:124-137.

This followed the more theoretical and motivating paper:

- Webster, R., and G.B.M. Heuvelink. 2006. The Kalman filter for the pedologist's tool kit. *European Journal of Soil Science* 57(6): 758–773.

c) "**leadership qualities in pedometrics research**":

Since 2012 Gerard has lead the largest research project at ISRIC: AfSIS (Africa Soil Information Services); chaired the Dutch Soil Science Socitey from 2004-2008; project leader or (co)-supervisor of 4 current and 13 completed research projects registered by the (Dutch) National Academic Research and Collaborations Information System (NARCIS); see list at http://www.narcis.nl/person/RecordID/PRS1284143/

d) "**contributions to various aspects of education in pedometrics**":

Gerard teaches geostatistics, spatial uncertainty analysis and pedometrics to students of Wageningen University. The teaching is embedded in the Landscape Properties and Variability course to undergraduate students of Soil, Water, Atmosphere. Gerard also contributes to the Spatial Modelling and Statistics course to MSc students of Geo-information Science of Wageningen University and the Inventory Techniques for Land Science and Frontiers in Land Science courses of the MSc Soil Science. Recent post-graduate courses taught by Gerard are the Space-Time Geostatistics course to employees of Wageningen IMARES, the Statistical Methods for Spatial Data Analysis and Modelling course to PhD-students of the Production Ecology and Resource Conservation graduate school, and the Spatial Uncertainty Propagation workshop organised prior to the Pedometrics 07 conference. He has been a member of 18 PhD committees since 1996.

e) "**service to pedometrics**":

Gerard chaired the Pedometrics Commission in the period 2002-2006; during his mandate, the Digital Soil Mapping working group has been established; pedometrics got much more visibility within the IUSS (it was promoted to a commission in 2004; Pedometrics is now one of the most active research groups within IUSS). He chaired the Dutch Soil Dcience Socitey from 2004-2008, after being vice-president from 2001-2004). He is the co-editor of the Geoderma journal, and has edited several issue of the Pedometron newsletter and is still one of the main

contributors to that newsletter (notably with the Pedomathemagia column).

Gerard is originally from the tiny village of Kranenburg, a dot in the road between the small towns of Ruurlo and Vorden in the Dutch "Achterhoek" (rough translation: "way back"). He attended the University of Twente, majoring in applied mathematics, and found his way to Peter Burrough's group at the University of Utrecht, where he wrote an influential thesis (1993) "*Error propagation in quantitative spatial modelling: applications in Geographical Information Systems*", later (1998) published in book form by Taylor & Francis, with a glowing introduction by Michael Goodchild. He then spent about ten years at the University of Amsterdam, before joining Wageningen University, first with Alterra and then ISRIC as well as in the academic Soil Geography and Landscape group. On his Wageningen professional directory page he lists his expertise as "Statistics" with the keywords "uncertainty analysis, geostatistics, pedometrics". With the award of this medal we are pleased to confirm the last!

Gerard and Dick Webster

❖ Call for nominations for the Best Paper in Pedometrics 2013

David G Rossiter
(e-mail: dgr2@cornell.edu)
Chair Pedometrics Committee on Prizes and Awards for 2014—2017

Nominations are now open for the **"Best Paper in Pedometrics 2013"** award, to be presented at **Pedometrics 2015** (September 2015, in Córdoba, Andalucía, Spain). Early next year (2015) we will repeat the exercise for the best paper of 2014, but we want to spread the work out and also have papers nominated while they are still fresh in your minds. This is a prestigious award, which recognizes work that is judged to be of importance and of excellent quality by your peers. It stimulates us all to do top-quality, influential and innovative work.

The procedure is as follows:

1. You are all now invited to **nominate one or more papers**. They must be **relevant to pedometrics** and have been **published in recognized journals** with a **final publication date in calendar year 2013**. These can be from the usual journals for pedometricians, such as Geoderma and European Journal of Soil Science, but can also be from journals where we do not publish so much -- this would encourage us all to scan these journals. You may nominate a paper on which you are (co-)author. If you are confused about what exactly is "pedometrics" for the purposes of this award, please see the definition as approved by the IUSS, see http://pedometrics.org/ "What is Pedometrics?"
2. The nominations and justifications will be assembled by me, and sent for review to the committee:
   - David Rossiter (Cornell University, USA)
   - Sabine Grunwald (University of Florida, USA)
   - Alex McBratney (University of Sydney, Australia)
   - Margaret Oliver (University of Reading, UK)
   - Gerard Heuvelink (Wageningen University and ISRIC - World Soil Information, NL)

   The committee will independently grade the papers (0 to 10). I will average the scores. The five papers with the highest average will be then definitively nominated. In the case that a committee member is the (co)-author of a nominated paper, s/he is not allowed to grade her/his own paper, so the average is from the remaining members.
3. The nominated papers will be placed on the

pedometrics website, announced in the pedometrics Google group, in the IUSS LinkedIn group, and in Pedometron. All self-declared pedometricians are encouraged to read the nominated papers and rank them in the single transferable vote (Hare) system (first choice, second choice... up till the last paper the voter is willing to vote for). Votes should then be sent to me from a traceable e-mail address (to prevent over-voting), over a period of at least a month, to be announced. A pedometrician is not allowed to vote for a paper where s/he is a (co-)author.

4. I will tally the votes according to the Hare system and determine the winner. Votes will be kept secret and you will just have to trust me to tally them honestly.

Time line:

- DEADLINE for nominations is 15-September-2014.
- The list of selected papers will be announced by 01-November-2014.
- Pedometricians will then have two months to read and vote.

❖ Obituary

Dick Brus, vice chair pedometrics
dick.brus@wur.nl

Ben Marsman passed away at his home in Ede, Netherlands on Sunday, June 1, 2014 at the age of 83. Ben was nine years old when the Second World War began. He was not able to finish his education at Highschool. After the war he got a job as a soil scientist at the former Soil Survey Institute. Ben was an intelligent and studious man, and became a selfmade pedometrician avant la lettre. He took the initiative to collect statistical data for quantifying the quality of soil maps, which was in the beginning not very much appreciated by his colleagues. He worked closely together with Jaap de Gruijter on designing a probability sample for validation of the nationwide Soil Map of the Netherlands at scale 1:50 000. His most influential publication is B.A. Marsman and J.J. de Gruijter, 1986, "Quality of Soil Maps. A comparison of survey methods in a sandy area". Soil Survey Papers 15. Nowadays this publication would have easily passed a PhD examination committee. Besides he is co-author of a couple of papers published in international scientific journals. Ben stopped working at the age of 60, so that he could fully support his loved ones. We remember him as an excellent, honest and warm soil scientist.

Reports

❖ EGU meetings

Murray Lark
Environmental Statistician IM(3)
British Geology Survey

Statistics and Informatics at the European Geosciences Union

In 2013 the Soil System Sciences Divison of the European Geosciences Union (EGU) voted to set up a subdivision to look at statistics and informatics. The officers of the subdivision are Murray Lark (British Geological Survey), Ana Maria Tarquis (Universidad Politécnica de Madrid) and Beate Zimmermann (Forschungsinstitut für Bergbaufolgelandschaften). The primary role of the subdivision is to organize seminars for the annual EGU congress in Vienna, so 2014 was our first outing.

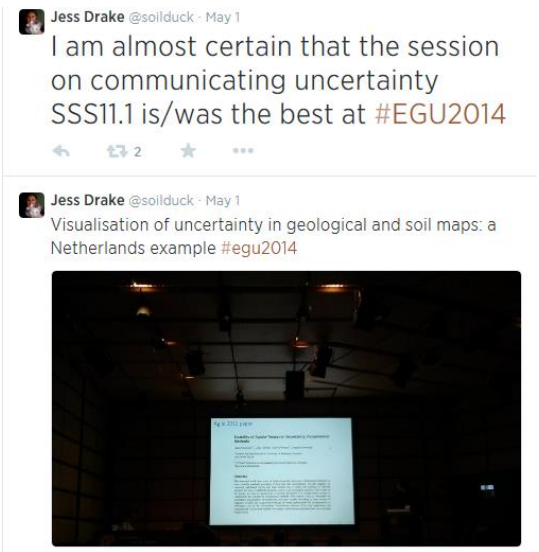
Along with a seminar organized by the GEMAS (Geochemical Mapping of Agricultural and Grazing Land Soil) project, and one on Digital methods for field mapping, three sessions were specifically organized on the subdivisions initiative. One covered a range of topics in the measurement and modelling of soil variation with particular focus on engineering and management problems. A second covered questions in soil sampling and digital soil mapping. Werner Müller gave an overview of issues in spatial sampling design and Budiman Minasny talked about some work on the use of digital soil maps to stratify for field sampling, a collaboration with Jaap de Gruijter. There was a case study on sampling for multivariate DSM from Hungary and presentations on various sensing technologies for DSM.

One session tackled a rather different issue. It was entitled *Communication of uncertainty about information in earth sciences* and exemplified the strongly interdisciplinary flavour of EGU meetings because, along with a healthy dose of soil-related material, there were presentations involving geology, greenhouse gas inventory and psychology.

Poster Session at EGU 2014. How do we communicate the uncertainty in soil maps?

As pedometricians we are used to the idea of quantifying uncertainty (of confidence limits, posterior prediction intervals, kriging variance) but have you ever encountered a blank look when presenting these ideas to non-specialists, including policy makers or managers? I certainly have, and this motivated the session. One of the speakers was Dr Adam Harris, a psychologist at University College London, who has examined the "verbal scale" used by the Intergovernmental Panel for Climate Change (IPCC) to convey the degree of certainty attached to different findings or predictions. For example, when the IPCC recently stated that "It is very likely that there is a substantial anthropogenic contribution to the global mean sea level rise since the 1970s" this means that the probability is over 90% (which is roughly the same as the probability of getting more than one "Head" in ten tosses of a fair coin). Clearly it matters that voters and politicians clearly understand the strength of the evidence. Unfortunately, psychologists have found that in general the verbal scale is interpreted regressively, large or small probabilities are generally interpreted too close to 0.5. We need to find ways of doing this better. Research has been done on this, and has been put to use. See, for example, this paper in Geoderma.

Various case studies were presented covering work in the UK on the agricultural greenhouse gas emissions budget, and how the uncertainty in this is presented to different audiences, and some novel ideas on how to present uncertainty in soil maps and in contours in spatial data. It was a stimulating and interesting meeting, and I don't think I was the only one who thought so, note that on the verbal scale "almost certain" means >90% so don't interpret this regressively!



We return to EGU next year, and I would urge Pedometron readers to get involved. There will be sessions on sampling, scaling and soil variation, modelling and visualization and a second round on communication of uncertain information. There will also be short courses on statistical topics. So mark the date of EGU 2015 in your diary (12th–17th April) and keep an eye on the EGU Soil System Sciences webpages as the programme evolves and YOU have the chance to influence it. For more information contact me at mlark@bgs.ac.uk.

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

L. Menichetti[1*] and A. Tonda[2]
[1]Swedish University of Agricultural Sciences, Department of Soil and Environment, P.O.Box 7014, 75007 Uppsala, Sweden.
[2]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 than 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:

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

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

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

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

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

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

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

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

$$R_{norm} = 0.27 + 2.14 \cdot pH_{tops} + 1.61 \cdot min(min({pH_{tops}}^2, 1.48 + 1.29 \cdot Gravel_{subs} - pH_{subs} - pH_{tops}.$$

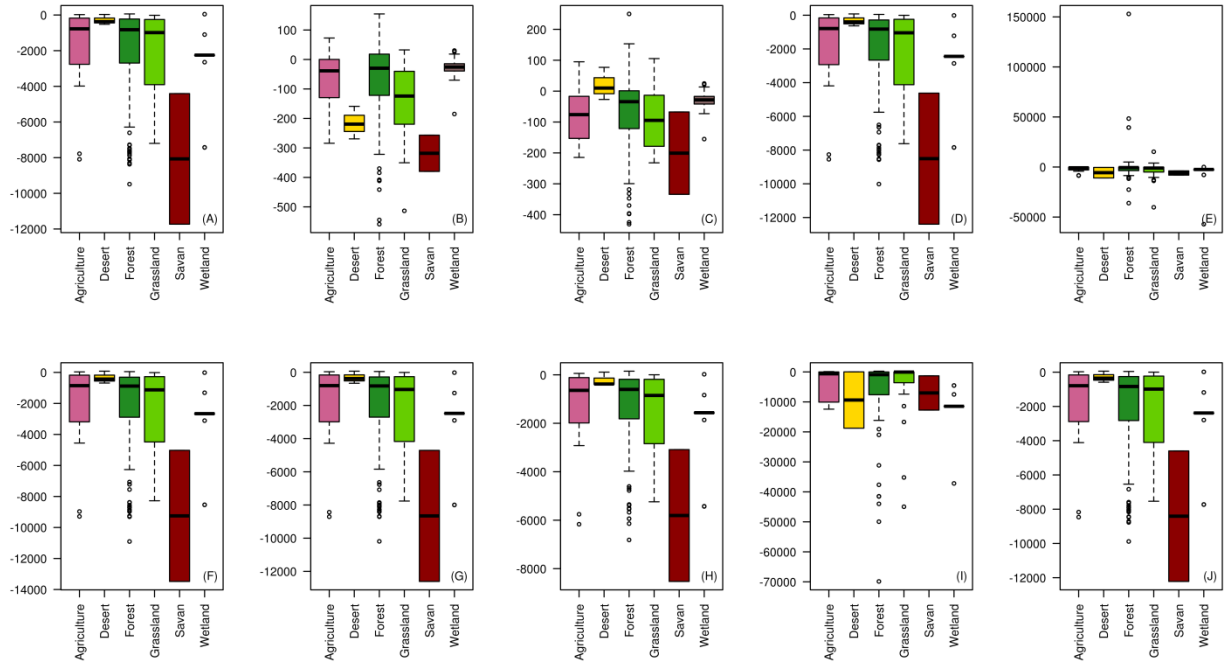


Figure 1: Residuals of the selected functions. A) Equation 1, B) Equation 2, C) Equation 3, D) Equation 4, E) Equation 5, F) Equation 6, G) Equation 7, H) Equation 8, I) Equantion 9, J) Equation 10.

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 <= R^2 <= 1). These indexes have been calculated on the whole dataset, without removing the outliers

	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
R^2	-2101.6	-73.6	-63.2	-2138.0	-1244.5	-2324.8	-2175.7	-1470.1

The PCA analysis of the residuals (Fig. 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.

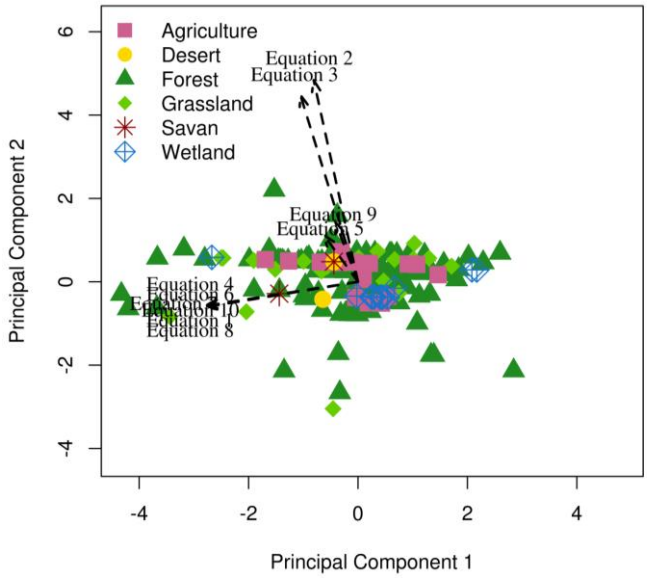


Figure 2: PCA analysis of the residuals

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.

4.2 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.

Papers

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.

PedoMathemagica

Answer to Pedomathemagica (3) in Pedometron issue 33

Bert is not amused. He has no intention of wasting time on puzzles, let alone buying all the Ruritanian lager. So he calls Kurt from the statistics department. Kurt turns up and looks at the list. He laughs. `Had you not noticed that the numbers against the six items are the first six prime numbers?' he says. Bert glares at him, so he carries on. `Every number larger than one is the product of a set of prime factors. It must be so, because, if any of the numbers in a factorization that you proposed was not prime, then you could factorize it into primes by definition. So, for example,

$$490 = 7 \times 7 \times 5 \times 2.$$

Now the order doesn't matter, of course, multiplication is commutative, but it must be the case that any set of prime factors that we propose corresponds to a unique number (their product) just as any number corresponds to a set of unique primes (its factors). That means that you can write any integer as:

$$2^{n_2} \times 3^{n_3} \times 5^{n_5} \times ...$$

where the integers in the sequence are the prime numbers and the powers are some integer value (which may be zero). Any integer corresponds to a unique set of values of n_2, n_3, n_5, ... and vice versa. That's the Fundamental Theorem of Arithmetic.' 'FTA!' says Bert, and looks hopeful. 'Let's guess at an underlying rule, and see if it gives sensible results in this particular case', says Kurt. ' I think that Alf's code is defined thus:

$$1486485000 = 2^{n_2} \times 3^{n_3} \times 5^{n_5} \times 7^{n_7} \times 11^{n_{11}} \times 13^{n_{13}}$$

Where n_2 is the number of augers, n_3 is the number of spades n_5 and so on, with n_{13} the number of GPS.' Bert is not impressed. 'We have one equation and six unknowns', he says. 'Alf will still win'. 'Not so fast', says Kurt. If you divide the code by 2^n for any n less than or equal to the number of augers, n_2, then you should have no remainder because your divisor should be a factor, but if you divide by 2^{n_2+1} it can't be a factor so you will end up with a remainder. Look!' and he scribbles the following down on the bonnet of the Landrover with a piece of chalk that he takes from behind his ear.

1486485000/2 = 743242500
743242500/2 =371621250
371621250/2 =185810625
185810625/2 =92905312.5

'You can divide the code by 2^3 without remainder. Divide by 2^4 and you get a remainder, so, if I am right, then Alf wanted three augers'. 'Shall we do the same with the other numbers?' says Bert. 'You could', said Kurt, but it would be neater to use the last value you got without remainder for augers, 185810625, and carry on. That way when you get to GPS your final quotient without a remainder will be 1, an extra check your arithmetic'.

The fact that the fundamental theorem of arithmetic allows a single number to code for a unique set of integers (provided you know the prime which corresponds to each integer in the set) still seems like magic to me, even though the theorem is not hard to follow. Kurt Gödel used this to produce unique numbers to code logical propositions. When I first wrote this problem I offered a prize for anyone who could provide a genuinely useful and non-trivial application of the fundamental theorem of arithmetic in pedometrics. However, before sending

the problem to Pedomathemagica I had already found such an application: the definition of unique codes to define teams of samplers given a code for each individual, and without having to specify the individual samplers in any particular order. This can be done by giving each sampler a unique prime code and forming the team code as the product of the codes of each constituent sampler. The desirable properties of the code follow from the FTA (uniqueness) and from commutativity (independence of order).