Computers & Geosciences 156 (2021) 104907

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Three-dimensional modelling using spatial regression machine learning and hydrogeological basement VES

Gastón M. Mendoza Veirana [a,b,*], Santiago Perdomo [c], Jerónimo Ainchil [a]

[a] Universidad Nacional de La Plata, Facultad de Ciencias Astronómicas y Geofísicas (National University of La Plata, School of Astronomical and Geophysical Sciences), Paseo del Bosque s/n, La Plata, B1900, Buenos Aires, Argentina
[b] Department of Environment, Faculty of Bioscience Engineering, Ghent University, Coupure 653, 9000, Gent, Belgium
[c] Universidad Nacional del Noroeste de la Provincia de Buenos Aires, Centro de Investigación y Transferencia del Noroeste de la Provincia de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (National University of Northwestern Buenos Aires, Research and Transference Centre of Northwestern Buenos Aires, National Scientific and Technical Research Council), Monteagudo 2772, Pergamino, B2700, Buenos Aires, Argentina

ARTICLE INFO

Keywords:
Extremely randomized forest
Vertical electrical soundings
Interpolation
Spatial regression
Geostatistics
Interserrana

ABSTRACT

In the last decade, machine learning algorithms have shown their superior performance in the spatial interpolation of environmental properties compared to classical interpolation models. In particular, the random forest ensemble model has provided the best adjustment. In this work, we compare the performance of support vector machines (SVM), simple trees (ST), random forests (RF) and extremely random forests (ERF), using discrete depths obtained by vertical electrical sounding (VES) from the hydrogeological basement of a sedimentary basin in Argentina; the coordinates are not gridded but almost aligned. On the other hand, in different artificial intelligence applications, the ERF algorithm has surpassed several methods of machine learning, including random forests. To the best of our knowledge, we hereby report the first spatial regression application of the novel ERF algorithm, which predicted—even better than RF—values it had not been trained for with an average R^2 score of 97.6%. This allowed us to obtain a satisfactory generalization of VES depths in the form of a three-dimensional approximation of the basement. The ERF algorithm also outperformed RF in computation time and smoothness of the surface generated. The primary significance of the results reported here lies in the relative independence that this technique has to offer, considering the area of application and gridding. Added to this, the nature of the method by means of which the discrete data are obtained is independent as well, as these could not only be derived from the VES technique, but also from well data or from different geophysical inversions.

1. Introduction

Statistical methods have a variety of uses in fields such as water resources, environmental sciences, agriculture or soil sciences, statistics and probability, ecology and civil and petroleum engineering (Li and Heap, 2014). In general, geostatistics is used when, based on discrete data, the aim is to know a spatial property in a continuous manner. In particular, the quantity and location of groundwater resources may be estimated through the knowledge of the confining stratum, which is commonly identified by means of discrete data from well observations (accurate) or geophysical inversions (estimated). In this case, spatial regression methods are essential to continuously determine the extent of an aquifer.

Based on the scarcity of lithological information from boreholes and geophysical explorations in this key productive area, a geoelectric campaign was planned and carried out. Difficulties such as accessibility to private fields and the relative remoteness of the area limited us to making observations along a route, as opposed to a regular grid. The aim of our work is to find the best algorithm to three-dimensionally model the hydrogeophysical basement studied, using machine learning (ML) regression techniques based on inverted vertical electrical sounding (VES) data obtained in the measurement campaigns. This issue will be addressed by delving into the prediction capacity of algorithms. One of the main uses of ML is to generalize predictions (Domingos, 2012), which is utilized here to obtain a close spatial extrapolation of the VES depths of the relatively underexplored basement. The main relevance of

* Corresponding author. Department of Environment, Faculty of Bioscience Engineering, Ghent University, Coupure 653, 9000, Gent, Belgium.
E-mail addresses: gaston.mendozaveirana@ugent.be (G.M. Mendoza Veirana), sperdomo@comunidad.unnoba.edu.ar (S. Perdomo), jero@fcaglp.unlp.edu.ar (J. Ainchil).
URL: https://github.com/MendoVeirana/ML-spatial-regression-testing (G.M. Mendoza Veirana).

https://doi.org/10.1016/j.cageo.2021.104907
Received 20 November 2020; Received in revised form 1 August 2021; Accepted 3 August 2021
Available online 18 August 2021

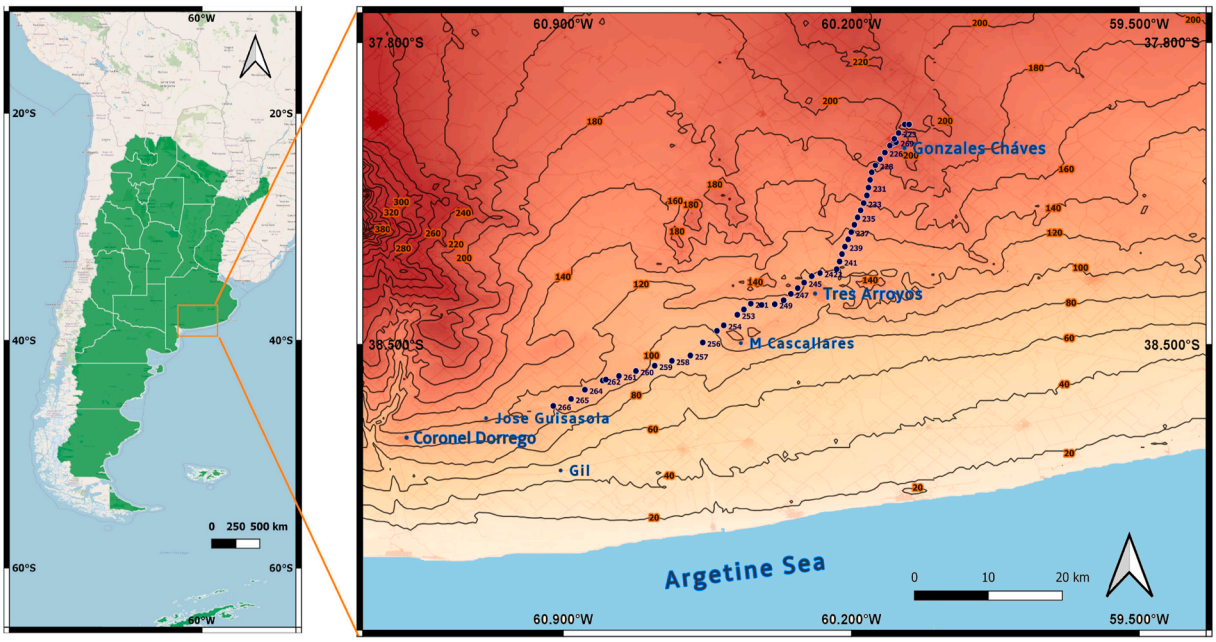

Fig. 1. Study area located in the Interserrana Basin. Each blue dot represents a VES. Level curves and distance to the coast of the Argentine Sea can be observed. [double column]. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

this work lies in the theory and application of the extremely randomized forest (ERF) algorithm. To the best of our knowledge, it has never been implemented in the field of spatial regression techniques. We discuss and show why this is the best option in comparison to RF, ST and SVM.

Spatial interpolation is defined as predicting the values of a primary variable at points within the same region of sampled locations, whereas predicting the values at points outside the region covered by existing observations is called *extrapolation* (Burrough and McDonnell, 1998).

Even though the ML methods used here are well-known and widely tested, a detailed revision of their underlying theory will allow us to understand why ERF outperforms RF, added to the reliability that the extrapolation of information has to offer. Due to the fact that such methods will be used in supervised learning mode, the problem may be thought of as its corresponding inverse: to propose a three-dimensional basement model that, when evaluated on the coordinates of the VES, returns known depths within a certain tolerable error. In this way, infinite models that correctly predict the known depths of the basement may be proposed. In order to face this problem inherent to every regression method, as suggested by Li et al. (2011) and Li and Heap (2014), the models are constrained by one main criteria: prediction error minimization. The secondary criteria to be taken into consideration can be the geological sense and smoothness or non-staggering of the generalized surface, minimization of calculation time, the simplicity in the implementation and the interpretability of their functioning.

1.1. State of the art of ML in spatial regression problems

The first contributions on the comparison of spatial regression ML applications for environmental studies can be seen in Prasad et al. (2006), Li et al. (2011), Li and Heap (2014) and, recently, Thessen (2016), Nussbaum et al. (2018), da Silva Júnior et al. (2019) and. Even though it is well known that there is no single interpolation method that is the best in all particular cases (Li and Heap, 2011, 2014; Li, 2016), all of the previously mentioned authors report the notable performance of RF compared to an extensive list of over forty classical methods, such as ordinary kriging (OK), inverse distance squared (IDS) and other ML methods, including ST, SVM regression and neural networks. The combination of RF and OK has only shown a slightly better adjustment than only RF in Li et al. (2011) for a particular case study. However, Li et al. (2011) takes the ten-fold cross-validation over all of the data available as performance criterion for the algorithms. But as reported in Hastie and Friedman (2009), Mehta et al. (2019) and the user's guide that was used in the algorithm implementation,[1] the use of the expected test (or out-of-sample) error, averaged over several partitions of the original data, is more appropriate as model test criterion. In this case, the algorithms predict values they are still untrained for, with cross-validation being entirely implemented within the training set. On the other hand, recent studies in geophysical applications (Elmousalami and Elaskary, 2020; Lawson et al., 2017) have found the ERF algorithm (Geurts et al., 2006) is the one that fits best, even superior to RF. In fact, the original work on ERF emphasizes the advantages of this algorithm compared to RF and other ensemble methods based on decision trees.

As described in Li and Heap (2008, 2014) and Domingos (2012), ML methods are differentiated by their capacity to use multiple variables, their deterministic character and the possibility of generalizing predictions. As regards the first characteristic, in Li et al. (2011) and Geurts and Wehenkel (2006) it can be noticed that the introduction of secondary variables may not improve the prediction capacity of the algorithm. On the other hand, in the specific case of Li et al. (2011), it has been found that for RF spatial regression in longitude and latitude it is optimal to use only these two variables, whereas the introduction of additional ones, even if they have a physical relationship with the variable to be predicted, does not necessarily improve the performance of the algorithm.

2. Study area

The study area (Fig. 1) is the Cuenca Interserrana (Interserrana Basin), which is located in an extensive area of 60,000 km^2 in a region of the Pampean plain of the province of Buenos Aires, Argentina. It has a smooth topography, with heights of 200 m above sea level (m. a. s. l.) in the northeastern sector and up to 100 m a.s.l. in the southwestern sector. Livestock and agricultural activities are carried out in this region and are mainly supplied by groundwater, which is also the source of water for human consumption. Due to the scarcity of exploratory wells and

[1] https://scikit-learn.org/stable/modules/cross_validation.html.

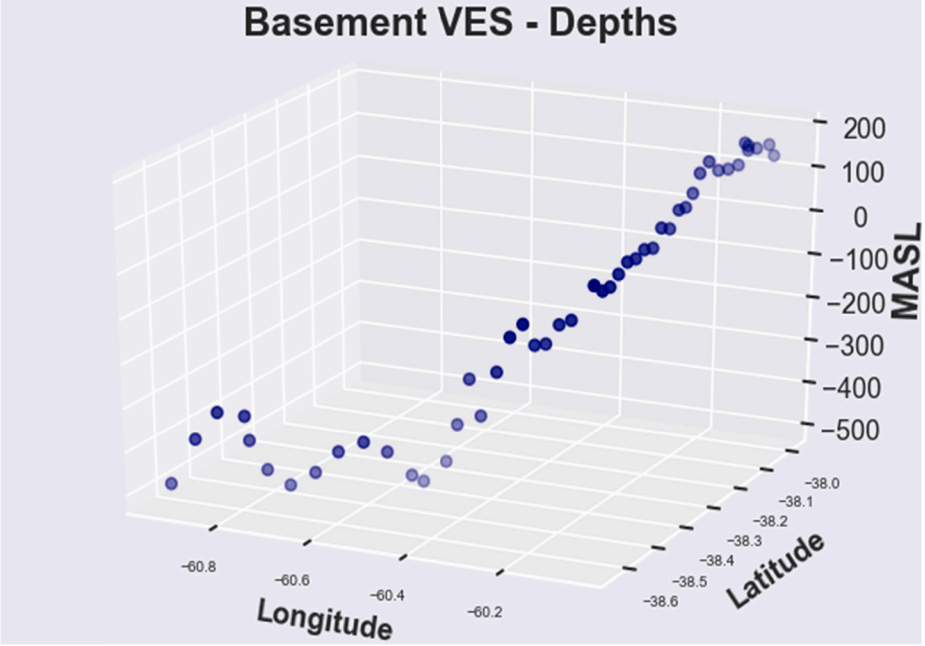

Fig. 2. Basement VES depths. Depth in meters above medium sea level were obtained by VES inversion. [single column].

hydrogeological studies in the area, it was decided to carry out VES measurements to estimate the thicknesses and hydrochemical characteristics of the different aquifer and aquiclude units. The sedimentary deposits of the basin are of eolian origin (Kruse et al., 1997; González, 2005) and overlie a Paleozoic basement composed of quartz schists. This basement, with impervious characteristics, has high values of electrical resistivity. Previous geophysical contributions on the geoelectric exploration of the Interserrana Basin are critically scarce, with only one previous work (Weinzettel and Varni, 2007) in the Claromecó Basin, which is part of the Interserrana Basin. As a result, a strong positive correlation can be observed between the deepening of the basement and the proximity to the Argentine Sea.

2.1. Vertical electrical soundings

Forty-eight VES with a Schlumberger array were measured along a 117-km route [37° 59′–38° 38′ south latitude and 60° 04′–60° 55′ west longitude, WGS84 system]. The distance between the observations was from 2 to 4 km. The maximum separation between electrodes (AB/2) was 500 m. The initial data processing was performed with Zohdy's algorithm (Zhody, 1989) and then the number of layers was reduced using the Dar Zarrouk parameters (Maillet, 1947) with a root mean squared error (RMSE) tolerance of 5%. Due to the geological literature and the resistivity values observed for the basement, its true resistivity value was set at 100 Ohm m. Taking these two considerations into account, all of the VES were processed with adjustment within the accepted tolerance, obtaining an average value of 3.78% across all the VES. In Fig. 2, the interpreted depth of the basement can be observed. The basement is superficial at the northeastern sector, beginning at 170 m a.s.l. and continuing with an irregular deepening that reaches −480 m a.s.l. in the southwestern sector. Extended details and analysis about the VES acquisition and processing are discussed in Appendix.

3. General considerations about machine learning

The aim of this short review section is to provide a solid foundation regarding the functioning of the ML methods, the main results concerning model selection theory and our implementation on such basis.

We have used and tested two successful algorithm families in this field: support vector machines and decision trees. The input data structure of these algorithms is a set of pairs $D = \{X,\ Y\} = \{(x_1, y_1),\ \ldots\ (x_N,\ y_N)\}$, where $N$is the number of samples, each sample x_i is a $1\ x\ P$ vector of P features and y is the unidimensional vector of depths calculated from VES. In our first test, $x_i\ = \{Lat_i, Long_i\}$; subsequently, in the implementation including secondary variables, $x_i\ = \{Lat_i,\ Long_i,\ Coast\ Distance_i,\ Topographic\ Height_i\}$. *Coast Distance* is the minimum distance to the coast of the Argentine Sea for each point. According to the literature (González, 2005; Kruse et al., 1997; Mendoza Veirana, 2019; Weinzettel and Varni, 2007) *Coast Distance* and *Topographic Height* features were chosen since they would have geological relationship with the basement depth. The distance to the coast has been characterized as a related indicator of basement depth in Claromecó Basin (see Weinzettel and Varni, 2007, Fig. 4) and topographic height was considered since structural control is not fully ruled out in Interserrana Basin. Pearson product-moment correlation coefficients between basement-depth and variables are 0.95 for *Latitude*, 0.9 for *Longitude*, 0.95 for *Coast Distance* and 0.89 for *Topographic Height*.

As required in supervised learning, set D is randomly divided in succession and shuffled into training sets D_{TRAIN} and testing sets D_{TEST} with an 80/20 ratio to train the models in the former and test them in the latter. It is the aim of this procedure to guarantee that the trained model can successfully predict the test data and then establish a reliable generalization. During the training stage, the successive complexities of the models are used to calculate the training error E_{TRAIN}, and then test the model in D_{TEST}, thus searching for the lowest test error E_{TEST}. As is easily seen, every randomly drawn D_{TRAIN} will probably be different from the others similarly formulated and, consequently, the trained model will predict different depth values also with different errors. In order to overcome the obstacles in the random selection of D_{TRAIN} and D_{TEST}, we averaged the errors over one hundred divisions of D, so as to obtain the expected value from the test error $\mathbb{E}[E_{TEST}]$ (Mehta et al., 2019). Errors will be calculated by mean squared error (MSE) and coefficient of determination (R^2 score). As is recommended in Li (2016), in the case of test errors, both parameters are the most appropriate to assess spatial interpolation predictive models.

3.1. Bias-variance tradeoff

One of the classical problems in supervised predictive modelling is the bias-variance tradeoff. A high bias may cause the loss of relevant relationships between the characteristics and the target variable, that is, underfitting. On the other hand, a high variance may cause small adjustment variations (overfitting), such as in random data noise and, therefore, in this case the generalization will inevitably be incorrect (Domingos, 2012). The ways of reducing tradeoff are by means of ensemble methods, such as bagging and RF (Hastie and Friedman, 2009). Overfitting problems are detected when the difference between the training and testing errors is large, or the testing is deficient. One of the most important and useful results in model selection theory in statistical learning is the squared error decomposition of an estimator.

Given the y-vector of depths calculated from VES data in D, it is possible to think that these values are composed of the actual depths added to noise:

$$y = F(x) + \varepsilon \quad \text{Eq. 1}$$

whereε is normally distributed with zero mean and standard deviation σ_ε.

Regardless of $j(x)$ being an estimator or an ensemble of them successively trained on particular D_{TRAIN} that adjusts (x_i, y_i) with$i = 1, 2 \ldots 0.2{*}N$ belonging to D_{TEST}, then:

$$\mathbb{E}\,[E_{TEST}] = \mathbb{E}\left[\sum_i (y_i - j(x_i))^2\right] = \sum_i (F(x_i) - \mathbb{E}[j(x_i)])^2 + \sum_i \mathbb{E}\big[(j(x_i) - \mathbb{E}[j(x_i)])^2\big] + \sum_i \sigma_{\varepsilon i}^2 \quad \mathbb{E}[E_{TEST}] = Bias^2 + Var + Noise \quad \text{Eq. 2}$$

where $\mathbb{E}$ is the expected value of the argument over all of the partitions of D. In the cases in which the variance of the estimator grows along with the complexity of the model, $\mathbb{E}[E_{TEST}]$ commonly has a minimum value between the simplicity and the maximal complexity of the model—even though, as we shall see, in the ensemble methods such as RF, the variance tends to zero as the randomness of the estimators and their number increase. A modern discussion on the classical U-shape in the bias-variance tradeoff may be found in Belkin et al. (2019). Therefore, in ML we will not seek perfection, i.e., $\mathbb{E}[E_{TEST}] = 0$, we will aim for the best tradeoff, that is, the model for which $\mathbb{E}[E_{TEST}]$ is minimal.

4. Machine learning methods

4.1. Support vector machine regression

SVM is a machine learning algorithm for classification and regression problems developed by Cortes and Vapnik (1995). The methodological procedure for an optimal implementation was taken from Hsu et al. (2003).

The SVR method aims at fitting a $(p + 1)$-dimensional hyperplane that tends to be flat. The constant $C > 0$ determines the fit between the flatness of the estimator f and the largest amount of tolerated deviation (Smola and Schölkopf, 2004). As recommended in Hsu et al. (2003), we have chosen the radial basis function (RBF) as kernel since linear and sigmoid kernels may be particular cases of RBF for certain parameters (C, γ) (Lin and Lin, 2003).

4.2. Single tree

A decision tree is a machine learning technique that may be described as a succession of instances of decision that is formed by a root node, nodes and leaves; we briefly describe here the split criterion of the regression tree (CART) (Breiman et al., 1984). The root is the first node, and the leaves are the final unpartitioned instances that provide final solutions. The depth of the tree is the number of steps from the root node to the actual instance m. In classical regression trees, splits are made through a case of greedy algorithm. Nodes are instances of binary division, and based on sample space R_m, new subspaces $R_{left}$and R_{right} are defined here in each partition. For both subspaces, constant values $\underline{y_{left}}$ and$\underline{y_{right}}$, which are the averages of y_i for each subspace, are fitted; then, in our case, a constant depth is assigned to each partition. The candidate to split the local subspace into node m is $(s,\ a)$, where s is the characteristic chosen among those K available and threshold a is a specific value of it, so that:

$$R_{left}\ (s,\ a) = \left\{\frac{X}{Xs} \le a\right\}\ \&\ R_{right}(s,\ a) = \left\{\frac{X}{Xs} > a\right\} \quad \text{Eq. 3}$$

Parameters s and a, which minimize the cost function for the division, are selected.

$$\text{Cost function: } min_{y_{left}} \sum_{x_i \varepsilon R_{left}(s,a)} \left(y_i - \underline{y_{left}}\right)^2 + min_{y_{right}} \sum_{x_i \varepsilon R_{right}(s,a)} \left(y_i - \underline{y_{right}}\right)^2 \quad \text{Eq. 4}$$

Then, in the following step, $m + 1$, another split could be made until the samples in the node are a single one (leaf) or the tree depth is pruned.

4.3. Random forest

A random forest is an ensemble method based on a sequence of decision trees developed by Breiman (2001). This method may be quickly described as a collection of uncorrelated M random trees and their average. A training set Θ_k is generated for each tree, randomly drawn on D_{TRAIN}, independently from the sets of other previous trees $\Theta_1, \Theta_2 \ldots \Theta_{k-1}$ but with repetition of samples and the same probability distribution. Due to the replacement, the size of Θ_k and D_{TRAIN} are similar, that is, 0.8 * N. This procedure is called bootstrap aggregating (bagging) (Breiman, 1996); it is a method used to reduce the variance of the forest and avoid overfitting problems (Hastie and Friedman, 2009). The samples that are out of Θ_k but into D_{TRAIN} are called out of bag (OOB) and they are used to calculate the OOB error (Biau and Scornet, 2016).

The tuning parameters are the pruning, the number of M trees in the forest, the amount of features considered in every node (K) and the minimum number of samples admitted for division n_{min}.

The vector of estimated depths in each tree trained on D_{TRAIN} is $J_k = T_k(X_{TRAIN};\ \Theta_k)$, and is averaged to obtain the final random forest model.

$$J_{RF}^M = \frac{1}{M\sum_{k=1}^M T_k(X_{TRAIN};\ \Theta_k)} \quad \text{Eq. 5}$$

The variance of an RF can be expressed depending on the trees that compose it:

$$Var\left(J_{RF}^M\right) = \rho(x)\sigma^2 + \frac{1-\rho(x)}{M}\,\sigma^2 \quad \text{Eq. 6}$$

where σ^2 is the sample variance of any tree drawn randomly and $\rho(x)$ is the sampling correlation between any pair of trees used in the forest. This last formula is the key to understanding the power of random sets. Taking into consideration that in using large ensembles, M tends to infinity, we can reduce the variance significantly. Besides, in the case of completely random sets where trees are not correlated $\rho(x) = 0$, we can suppress variance as much as possible and then minimize E_{TEST} (Louppe, 2014). As for the forest bias, it is similar to the bias of any of the individual trees. Since the introduction of randomness is a key point in reducing the variance of the ensemble, in section 4.4 we will see an extreme case of this.

4.4. Extremely randomized forest

Extremely random trees, or simply extra trees (ET), developed by

Table 1
Main results using primary variables. The features used here were Latitude and Longitude.

80/20 Ratio split	Train MSE [m^2]	Train score [%]	Test MSE [m^2]	Test score [%]	5-fold CV [%]	Variance [m^2]	Bias^2 + noise [m^2]	Para-meters	Smoothness index	Computa- tion time [ms]
SVR*	337	99.3%	1411.8	96.5%	95.4%	343.7	1068.1	C = 1000 γ = 3.16	1	5.2
ST*	0	100%	2252.3	94.7%	93.1%	550.6	1701.7	K = 2	0.01	2
RF*	1649.9	96.7%	1603.5	96.3%	–	162	1441.5	K = 1	0.71	772.8
ERF*	0.7	100%	1217.6	97.6%	96.3%	152.9	1064.7	K = 1	0.98	625.5

Table 2
Main results using secondary variables. The features used here were Latitude, Longitude, Coast distance and Topographic height.

80/20 Ratio split	Train MSE [m^2]	Train score [%]	Test MSE [m^2]	Test score [%]	5-fold CV [%]	Variance [m^2]	Bias^2 + noise [m^2]	Parame- ters	Smoot-ness index	Computa- tion time [ms]
SVR*	236.7	99.5%	2133.2	95.1%	93%	573.4	1559.8	C = 316 γ = 3.16	1	5.6
ST*	0.3	100%	3362	91.5%	90.8%	1104.1	2257.1	K = 2	0.01	2
RF*	1940.7	96.1%	1948.6	95.5%	–	151.5	1797	K = 2	0.68	764.2
ERF*	0	100%	1540.1	97%	95.9%	162.2	1377.9	K = 4	0.89	728.9

Geurts and Wehenkel (2006), are a simpler special case of decision trees that differ from classical trees in the way they are built. There are two main differences: first, instead of using bagging, all the data available in the training set are used to build every node. Secondly, the division criterion to separate the samples in a node into two groups is to draw the thresholds at random for each candidate feature, instead of looking for the most discriminatory thresholds. The best of these randomly generated thresholds is chosen as the division rule by means of the MSE criterion. In ERF, as well as in RF, extra trees are used several times to generate an ensemble model of M trees. The predictions of each tree are used to produce the final prediction, similar to the one in Eq. (5), by arithmetic average in regression problems.

From the viewpoint of bias-variance, the reason underlying the ERF is that the explicit randomization of the cut-off point and the attribute (s, a), combined with the ensemble average, should be capable of reducing the variance with greater strength than the weaker randomization schemes used in RF. In other words, we seek to minimize $\rho(x)$ (Eq. (6)). The use of the complete original training sample instead of the bagging replicas is motivated to minimize the bias. In calculation times, the growth process of ERF is of the order of 1.23 times faster than RF. Another important characteristic to be taken into consideration, as well as in RF, is the number of features (K) to be considered in each node. There are several criteria to relate K with P but, as recommended by Geurts and Wehenkel (2006) and Li et al. (2011), it is preferable to seek the optimum K testing for each value. A thorough analysis of the error, the parameters and the bias variation can be found in the original paper, which shows in several tests the benefits in comparison to ST and RF.

5. Results

In this section we show the results of the algorithms implemented. The main results for training with the variables Longitude and Latitude can be found in Table 1, and the same features, adding Coast distance and Topographic height, in Table 2. Therefore, we show here the performances and how the parameters were selected, whereas the discussions are presented in the following section. Once the best parameters were chosen, we proceeded to retrain the models on every D. Owing to the smoothness of the geology, the spacing between the VES and the good fit of the algorithms implemented, we have taken a 10 km wide generalization to obtain the three-dimensional models shown in Fig. 3. In the plots in Figs. 4 and 5, blue represents train errors; red, tests; cyan, cross-validation; yellow, variance and green, bias. Given that with the data available it is not possible to determine the noise term in Eq. (2) and, therefore, to calculate the bias directly, we shall refer to both of them simply as bias, which was calculated as the subtraction between $\mathbb{E}[E_{TEST}]$ and variance. The computation times were obtained by averaging 1000 training instances on random partitions of D. In the case of SVR*, data normalization was included due to the inherent need of the method.

5.1. SVR results

The algorithm was implemented using the Python programming language, the Scikit-learn library (Pedregosa et al., 2012) and the sklearn.svm.SVR[2] module for the application of $\varepsilon - SVR$. (Chang and Lin, 2011). Thus, the following steps were undertaken:

1. The data were scaled to [−1,1] with the aim of avoiding attributes in higher numerical ranges from dominating those in smaller numerical ranges (Hsu et al., 2003).
2. Successive C and γ were used to train the model on the 100 D_{TRAIN} partitions of D.
3. A grid search was applied for each parameter combination; the expected error values were calculated with MSE and R^2 to seek the optimum C and γ.
4. Once the parameters that minimize $\mathbb{E}[E_{TEST}]$ were found, a closer search around the first choice for (C, γ) was performed.
5. A random normal point cloud centered on the VES coordinates was generated to apply the final model and thus obtain the 3D structure of the hydrogeophysical basement by VES.

Fig. 4 visualizes the grid search showing the distribution of MSE (left) and R^2 (right) errors for C and γ combinations, in the training, testing and cross-validation stages. Visually, the smaller MSE test corresponds to the smaller red dot, while for R^2it is the larger red dot. In the case of the primary variables, C = 1000 and γ = 3.16 were chosen as tuning parameters for SVR*. Regarding the secondary variables, the optimal parameters were C = 316 and γ = 3.16.

[2] https://scikit-learn.org/stable/modules/svm.html#svm-regression.

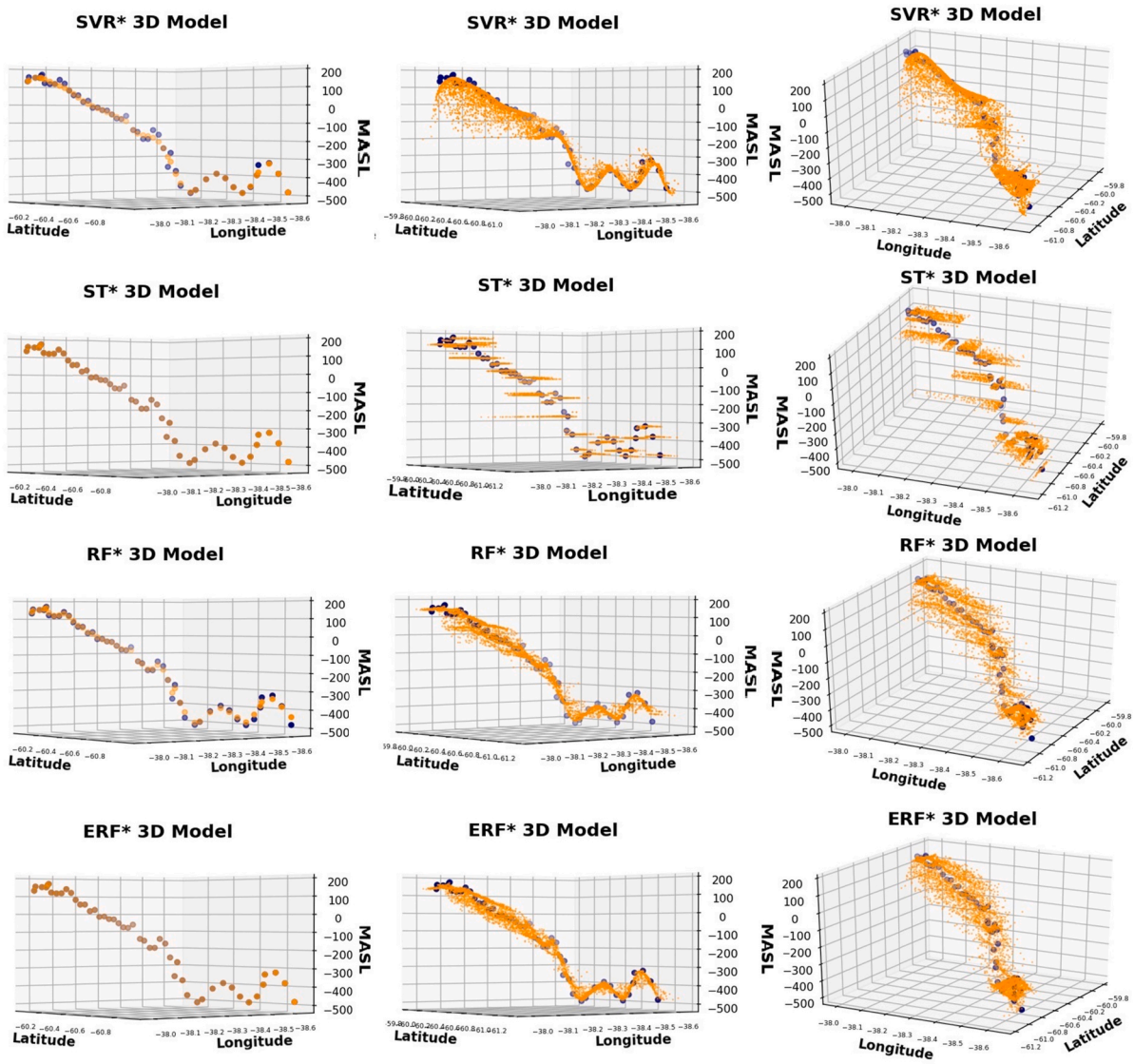

Fig. 3. Results of three-dimensional generalization of SVR* in the first row, ST* in the second row, RF* in the third row and ERF* in the fourth row. The first column shows a profile view of the predicted values of the model for the same coordinates of the VES. The second column shows the basement generalization using a Gaussian cloud of 4000 random points with a width of 10 km centered on the VES coordinates. Finally, the third column shows a more frontal view of the same generalization data. [double column].

5.2. Simple trees, random forest and extremely randomized forest

We implemented ST, RF and ERF in Python using the Scikit-learn library[3] (Pedregosa et al., 2012). The implementation steps were simple:

1. The input data format was as described previously in section 3.1.
2. One hundred partitions of D were defined in D_{TRAIN} and D_{TEST} to train the algorithms with trees with a depth between two and ten (unpruned).
3. For each partition, the implementation of the algorithms was similar: two minimum samples per node, successive K and 500 trees in the cases of RF and ERF.
4. E_{TRAIN} and E_{TEST} were calculated using MSE and R^2, estimating the expected errors by averaging over 100 values.
5. The depth of pruning and K that minimize $\mathbb{E}\,[E_{TEST}]$ were chosen to retrain each particular model, called ST*, RF* and ERF*, over all D.
6. The final step is similar to the last one in the implementation of SVM.

Fig. 5 shows the evolution of the three ensemble algorithms over increasing depth, from two to ten (unpruned) tree depths. The light lines show the performance (in MSE and R^2) of the algorithms for each data partition, while the dark curves show the average performance across all partitions for each depth. Note that there is no increase in test errors as the complexity of the algorithms increases.

6. Model selection and limitations

Based on the statistical model selection theory (Hastie and Friedman, 2009; Mehta et al., 2019), the algorithms implemented have been successful from a numerical standpoint. The training and test scores expected are not only very high, but also very close to each other for each model. This could be due to the relative simplicity of the statistical pattern to be recognized (Li and Heap, 2008) or the high correlation coefficients between features and the target. The best prediction capacity performance is the one of the ERF* algorithm using only primary variables.

[3] https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees.

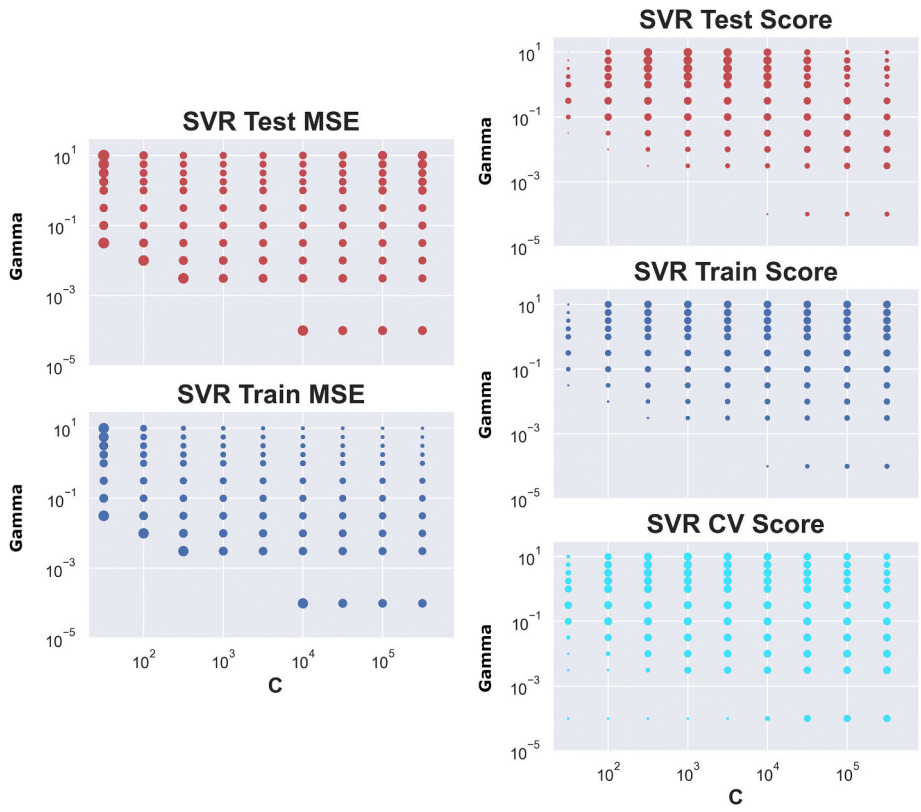

Fig. 4. Bilogarithmic plots of error distribution for MSE (left) and R^2 Score (right) related to C and γ for primary variables. As a first result, C = 1000 and $\gamma = 10$ were chosen; however, upon closer search, C = 1000 and $\gamma = 3.16$ have shown the best fit. Parameters have been ignored for which CV Score is negative. [1.5 columns].

As discussed in Li et al. (2011), a visual inspection of the surfaces generated is necessary, regardless of how good the fit was. Basement generalizations are convex in the cases of ST*, RF* and ERF*, whereas in SVR* it is non-convex. Despite the fact that the interpolation capacity of the algorithm is good, as well as the implementation of ST, the basement proposed does not make geological sense, as it does not fit the geological smoothness described in the literature (Kruse et al., 1997; González, 2005; Mendoza Veirana, 2019; Weinzettel and Varni, 2007). These contributions indicate a clear regional trend towards a gentle deepening in the direction of the Argentine Sea, contrary to the surface trend of the basement towards the west. The generalization of ST* is noticeable here, as the sample space was partitioned into 45 different subspaces, resulting in only 45 different values for the 4000 points evaluated. Therefore, it can be said that the smoothness index of ST* is $45/4000 \simeq 0.01$. On the other hand, ERF* generated a considerably smoother surface than RF* (see Table 1), which is also visually noticeable in Fig. 3. As shown in Geurts and Wehenkel (2006), this is due to the randomization of the split point, which makes it possible to generate more subspaces than in the case of RF. From the point of view of the bias-variance decomposition for primary variables, the ERF* algorithm improved the prediction error of RF*, reducing the variance and, even further, the bias. As can be observed in Geurts and Wehenkel (2006), the first source of error is minimized introducing randomness in the selection of the split point, whereas the bias is reduced by omitting the use of bagging. In fact, if we introduce bagging in the implementation of ERF* for primary variables, we obtain an increase in the bias from 1064.7 to 1375.5 m^2, while the variance remains stable.

As in the implementation in Li et al. (2011), the introduction of additional features that in principle physically relate the property to be predicted has only led to a reduction in the predictive ability. We believe that this fact, which could be counterintuitive in principle, may be due to the nature of ML, that is, to recognize statistical patterns and reproduce them. As discussed in Domingos (2012) and Geurts and Wehenkel (2006), the use of more predictor variables does not necessarily imply an improvement in predictive ability. As for tree ensembles spatial regression, it would seem that—as in the present case—two non-collinear spatial variables like Longitude and Latitude are enough to recognize the distribution pattern of the property and to predict it successfully. Additionally, as well as is highlighted in Li et al. (2011) it should be noted that in this study the projection utilized is not equal distance, which may affect the performance of the regression algorithms. Another issue is to define the extrapolation distance that may be reliable for the basement topography. Considering the fact that our geological area is not rugged, the VES were not measured strictly in a line but in a curve with a cross section of 20 Km, and that the spacing between them is approximately 3 Km, we proposed generalizing no further than 10 km to the closest VES coordinate. This close extrapolation is useful to estimate the depth of the basement obtained by VES instead of having to perform new expensive surveys, and also to evaluate how the algorithms perform outside of the training data domain. Nevertheless, it can be expected that meaningless behavior could occur in extrapolations far from the training data, since there are not measurements in directions perpendicular to the transect.

7. Conclusions

Despite the fact that spatial regression is a process that may give results that are at first unconfirmed with direct studies such as drillings, we trust that we have found out that—and explained why—, at least for our case, the ERF algorithm greatly outperforms RF, the most reliable one and widely recommended so far (Li et al., 2011; Nussbaum et al., 2018; Prasad et al., 2006; Thessen, 2016). This novel superiority is in its predictive ability and the smoothness of the surface generated. The minimization of the expected out-of-sample error was carried out reducing bias and variance. This could be attributed to the randomization of the cut-off point in nodes and the use of the whole training set to

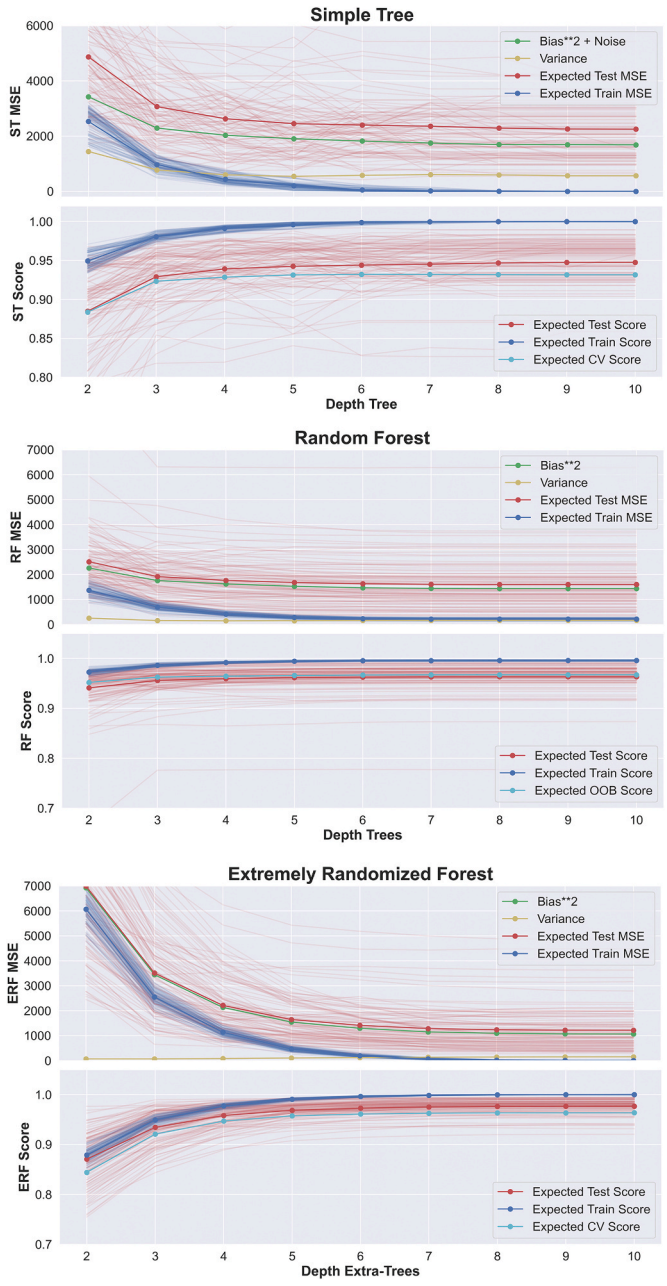

Fig. 5. Tree and ensemble complexities. One hundred partitions of D training and test sets were randomly created to train and test progressive depths of ST (top), RF (middle) and ERF (bottom) considering only primary variables. As can be seen, similarly as in Li et al. (2011), the minimum values of $\mathbb{E}[E_{TEST}]$ (in red) are for trees with no pruning. Take into consideration that due to the use of out-of-bag (OOB) estimates in RF, OOB and training errors are similar. [single column]. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

train every tree.

As is highlighted in Li et al. (2011) and Prasad et al. (2006), a visual examination of the data generated is essential regardless of the predictive ability of the algorithms. In this way, we may detect whether the generalized data make sense in the context of their physical characteristics. This has allowed us to reject the generalizations obtained with ST* and SVR* as being geologically inconsistent with the previous studies (Kruse et al., 1997; González, 2005; Mendoza Veirana, 2019; Weinzettel and Varni, 2007).

As stated in Li et al. (2011), the introduction of secondary variables deteriorated the performance of the algorithms, despite the fact that such variables are physically connected to the property to be predicted. This could be explained by the statistical nature of ML, which does not necessarily make use of features causally related to the property to be predicted.

The spatial regression procedure by ERF allowed us to delimit the extension of the aquifer with a cutting-edge precision. Moreover, the source of the basement depth data used for the model was irrelevant in the implementation of ERF and could originate from any reliable method other than VES, such as well observation or other geophysical inversions. Besides, data gridding, which in our case is almost linear, was no obstacle for the implementation of the algorithm. However, based on the limitations described, the hydrogeological basement in Interserrana basin should be further explored using a regular mesh as well as testing the performance of ERF in similar grids.

Finally, since the ERF algorithm generalizes from known data, a caution must be made as usual in any spatial regression method about the applicability of this procedure in rugged geological environments.

Computer code availability

Name: ML spatial regression testing.

Developer: Gastón Mendoza Veirana, University of La Plata, School of Astronomical and Geophysical Sciences. Paseo del Bosque s/n, La Plata (B1900), Buenos Aires, Argentina. Department of Environment, Faculty of Bioscience Engineering, Ghent University, Coupure 653, 9000 Gent, Belgium.

Mail: mendoveirana@gmail.com, cellphone: +32 0496908368. Available since November 2020. Code language: Python. Hardware minimum requirements: i3, two nucleus processor, 8 GB ram memory. Program size: 3.4 MB.

Open access source code: https://github.com/MendoVeirana/ML-spatial-regression-testing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

Dataset related to this article can be found at https://doi.org/10.17632/vjhy9gkzg3.1, an open-source online data repository hosted at Mendeley Data (Mendoza Veirana, 2020).

Authorship statement

Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing.

Resources, Supervision, Writing - Review & Editing.

Resources, Supervision, Founding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the reviewers for their valuable suggestions and criticisms. To Anna Kover and Magdalena Ponce for their precise language suggestions. To Juan Gabriel Gomila Salas for his valuable motivation and teaching in the use of machine learning. To Omar and Graciela for their indispensable support. Finally, to Pandemic for challenging us.

Appendix

A detailed literature review was conducted to include boundary conditions and thus limiting uncertainty in VES inversion. Only one exploration well was encountered 20 km from VES 266 at a locality called Gil (Fig. 1) (Dirección Nacional de Geología y Minería, 1970) where a quartz schist associated with resistive basement was encountered 427 m below the exploration point. Chemical analysis of water extracted at different depths of the borehole showed an increase in salt concentration with increasing depth. Due to the geological background, the observation of the borehole and the measured apparent resistivity contrasts, a value of 100 Ohm.m was established for the true resistivity of the hydrogeological basement. This value is between one and two orders of magnitude higher than the observed apparent resistivity of the sediments immediately above the basement.

All VES were performed using a Schlumberger array, with ten observations per decade at each point. Apparent resistivity was calculated for each VES using the following equation:

$$\rho_{oa} = \frac{\pi}{4*MN} * \left(AB^2 - MN^2\right) * \frac{\Delta V}{I} \qquad \text{Eq. Appendix 1}$$

where ρ_{oa} is the apparent resistivity observed in each measurement, MN is the distance between potential electrodes (ΔV), and AB is the distance between direct current electrodes of current *I*. Data processing and inversion were performed with SEV's software (Nigro and Perdomo, 2017), which implements the algorithm proposed in Zhody (1989). The initial true resistivity curve model has as many layers as observation points in the observed resistivity curve, and then this number of true layers is manually reduced using Dar Zarrouk's parameters (Maillet, 1947). Calculated apparent resistivity curves were computed by a digital filtering technique using Johansen's inverse filter (Johansen, 1975) of 139 coefficients:

$$\rho_{ca,\,w} = \sum_{u=1}^{139} b_u * T_{w-u} \qquad \text{Eq. Appendix 2}$$

where $\rho_{ca,\,w}$ is the apparent resistivity calculated at observation number *w,* and b_u are the Johansen's filter coefficients. *T* is the resistivity transform which is recursively calculated using the true depth and resistivities of each layer using Sunde's algorithm (Sunde, 1969). The maximum acceptable RMSE between the observed and calculated apparent resistivity curves was set at 5%.

Two representative examples are shown in Appendix Fig. 1, along with their observed and calculated resistivity curves (Appendix Table 1) and true resistivity curves (Appendix Table 2). VES 230 is placed in the NE sector while VES 257 in the SW sector of the basin.

Appendix Table 1
OARC and CARC of VES 230 and VES 257. The maximum AB/2 in VES 230 is 250 m, while in VES 257 it is 500 m.

AB/2	VES 230 OARC [Ohm.m]	VES 230 CARC [Ohm.m]	VES 257 OARC [Ohm.m]	VES 257 CARC [Ohm.m]
2	25,80	26,32	19,55	19,82
3	30,67	31,25	21,93	22,53
4	34,8	34,49	23,16	23,99
5	36,49	36,45	23,61	24,09
6	37,33	37,50	23,35	23,88
8	37,24	37,77	22,51	23,05
10	35,83	36,60	21,06	21,69
13	33,02	33,74	18,94	19,4
16	29,55	30,68	17,2	17,52
20	26,17	27,09	15,42	15,96
25	24,09	23,63	13,76	13,87
32	19,78	20,21	11,93	12,18
40	17,01	17,51	10,21	10,11
50	14,55	15,16	8,48	8,57
65	13,12	13,09	6,65	6,60
80	12,5	12,24	5,49	5,78
100	12,3	12,32	4,56	4,85
125	13,1	13,45	3,92	4,13
160	14,9	15,83	3,48	3,5
200	18,2	18,83	3,25	3,2
250	23	22,51	3,15	3,1
320			3,17	3,1
400			3,31	3,3
500			3,61	3,7

Appendix Table 2
TRC calculated for VES 230 and VES 257. VES 230 was modelled with 5 layers with resistivities from 7.6 to 100 Ohm.m while VES 257 was modelled with 6 layers with resistivities between 2.8 and 100 Ohm.m.

VES 230 TRC depth layers [m]	VES 230 TRC resistivity layers [Ohm.m]	VES 257 TRC depth layers [m]	VES 257 TRC resistivity layers [Ohm.m]
0.9	18	0.9	15
5.8	49	4.3	29.5
25	20	22.4	14

(*continued on next page*)

Appendix Table 2 (*continued*)

VES 230 TRC depth layers [m]	VES 230 TRC resistivity layers [Ohm.m]	VES 257 TRC depth layers [m]	VES 257 TRC resistivity layers [Ohm.m]
85	7.6	56.4	4.6
Inf	100	480	2.8
		inf	100

As a result, the final number of layers in the true resistivity curves ranges from 5 (as in VES 230, see Table 2 in the appendix) to 8 for the more complex observed apparent resistivity curves. Every VES was processed with a fit within the accepted tolerance, obtaining an average value of 3.78%. Finally, as can be seen in Fig. 2, basement depths were modelled from 170 in the NE sector to −480 m.a.s.l. in the SW.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2021.104907.

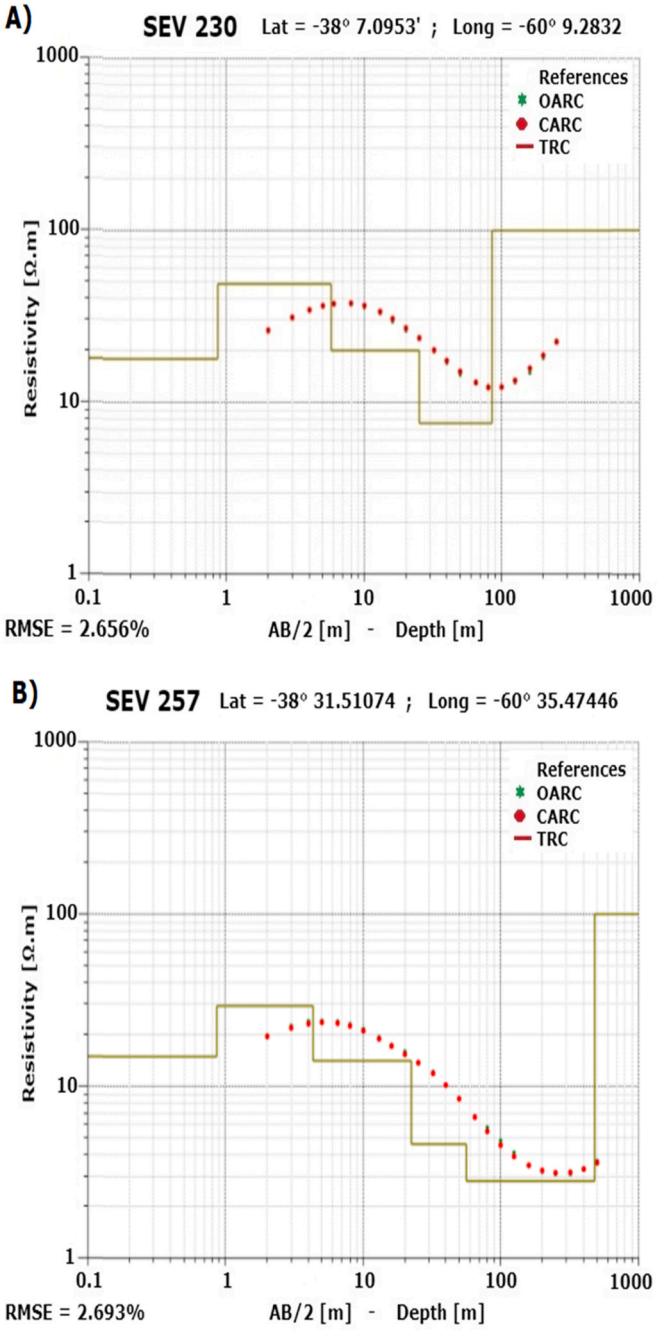

Fig. Appendix 1. Two VES proceed using SEV's software. A) In VES 230 the basement was modelled at a depth of 85 m with a RMSE of 2.656%. B) In VES 257 the basement was modelled at 480 m with a RMSE of 2.693%. The observed, calculated and inverted resistivity data are presented in Appendix Tables 1 and 2. [single column]

References

Belkin, M., Hsu, D., Ma, S., Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc. Natl. Acad. Sci. Unit. States Am., 201903070 https://doi.org/10.1073/pnas.1903070116.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25 (2), 197–227.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC press.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, Oxford.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 27. https://doi.org/10.1145/1961189.1961199. Article 27 (April 2011).

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018.

da Silva Júnior, J.C., Medeiros, V., Garrozi, C., Montenegro, A., Gonçalves, G.E., 2019. Random forest techniques for spatial interpolation of evapotranspiration data from Brazilian's Northeast. Comput. Electron. Agric. 166, 105017.

Dirección Nacional de Geología y Minería, 1970. Ministerio de Economía y Trabajo de la Nación. Perfiles de perforaciones periodo 1936-1945. Publicación 153, 146.

Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.

Elmousalami, H.H., Elaskary, M., 2020. Drilling stuck pipe classification and mitigation in the Gulf of Suez oil fields using artificial intelligence. J. Petrol. Explor. Prod. Technol. 1–14.

Geurts, D. Ernst, Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42.

González, N., 2005. Los ambientes hidrogeológicos de la Provincia de Buenos Aires. Relatorio del XVI Congreso Geológico Argentino, La Plata, pp. 359–374.

Hastie, T. R. Tibshirani, Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer, p. 745.

Hsu, C.W., Chang, C.C., Lin, C.J., 2003. A Practical Guide to Support Vector Classification.

Johansen, H.K., 1975. An interactive computer/graphic-display-terminal system for interpretation of resistivity soundings. Geophys. Prospect. 23 (3), 449–458.

Kruse, E., Deluchi, M., Laurencena, P., Varela, L., 1997. Caracterización de la red de drenaje para la evaluación hidrológica en la región interserrana (provincia de Buenos Aires). I Congreso Nacional de Hidrogeología y III Seminario Hispano – Argentino sobre temas actuales de hidrología subterránea. Bahía Blanca, pp. 133–145.

Lawson, E., Smith, D., Sofge, D., Elmore, P., Petry, F., 2017. Decision forests for machine learning classification of large, noisy seafloor feature sets. Comput. Geosci. 99, 116–124. https://doi.org/10.1016/j.cageo.2016.10.013.

Li, J., 2016. Assessing spatial predictive models in the environmental sciences: Accuracy measures, data variation and variance explained. Environ. Model. Software 80, 1–8.

Li, J., Heap, A., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Canberra, p. 13.

Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol. Inf. 6 (3–4), 228–241.

Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model. Software 53, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Software 26 (12), 1647–1659.

Lin, Lin, C.-J., 2003. A Study on Sigmoid Kernels for SVM and the Training of Non PSD Kernels by SMO-type Methods. Technical report. Department of Computer Science, National Taiwan University.

Louppe, G., 2014. Understanding Random Forests: from Theory to Practice. Ph.D. Dissertation. University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science, p. 211. arXiv preprint arXiv:1407.7502.

Maillet, R., 1947. The fundamental equations of electrical prospecting. Geophysics 12 (4), 529–556. https://doi.org/10.1190/1.1437342.

Mehta, P., Bukov, M., Wang, C.H., Day, A.G., Richardson, C., Fisher, C.K., Schwab, D.J., 2019. A high-bias, low-variance introduction to machine learning for physicists. Phys. Rep. 810, 1–124.

Mendoza Veirana, G., 2019. Determination of Hidrogeologic Basement on Interserrana Basin Using the Vertical Electrical Sounding Technique. M.Sc. Thesis. Facultad de Ciencias Astronomicas y Geofisicas. Universidad Nacional de La Plata. https://doi.org/10.3897/oneeco.1.e8621.

Mendoza Veirana, Gastón, 2020. VES-features. Mendeley Data, p. V1. https://doi.org/10.17632/vjhy9gkzg3.1.

Nigro, J., Perdomo, S., 2017. Desarrollo de software de inversión de datos 1D para sondeos eléctricos verticales Schlumberger. In: XXVIII Reunión Científica de la Asociación Argentina de Geofísicos y Geodestas (AAGG 2017) y Tercer Simposio sobre Inversión y Procesamiento de Señales en Exploración Sísmica (IPSES\'17)(La Plata, 17 al 21 de abril de 2017).

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M., Papritz, A.J., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4 (1), 1–22.

Pedregosa, Fabian, Varoquaux, Gael, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, Duchesnay, Edouard, Louppe, Gilles, 2012. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9 (2), 181–199.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Sunde, E.D., 1969. Earth Conduction Effects in Transmission Systems. Dover Publications, New York.

Thessen, A., 2016. Adoption of machine learning techniques in ecology and Earth science. One Ecosyst. 1, e8621 https://doi.org/10.3897/oneeco.1.e8621.

Weinzettel, P. y, Varni, M., 2007. Aportes al conocimiento del subsuelo de la cuenca del arroyo Claromecó, provincia de Buenos Aires. Taller de Geofísica Aplicada a la Hidrogeología. Universidad Nacional de Entre Ríos, ISBN 978-987-23936-0-1, 2007.

Zhody, A.R., 1989. A new method for the automatic interpretation of Schlumberger and Wenner sounding curves. Geophysics, USA 54 (2), 245–253.