Locally weighted support vector regression: a case study

Camilla Z. da Silva and Jeff Boisvert

Whenever a new methodology is proposed, is good practice to verify its applicability and performance relative to state of the art methodologies. This paper aims at demonstrating the use of LWSVR as an alternative algorithm to build trend model to be incorporated in the geostatistical framework. For that, the proposed workflow is applied on a well-known data set, the Walker Lake, and the results obtained are directly compared to Collocated cokriging with and elliptical radial basis function network interpolation, incorporated as auxiliary information. The results indicated that LWSVR is a practical alternative method to enhance predictions on estimated grade models.

Introduction

There have been a lot of discussion and research around the first and second order stationarity required by traditional geostatistical methods, given that the definition of stationary often is not straightforward due to the complexity of geological process. Machine learning (ML) methods arise as an alternative in this scenario since its premise is to learn complex relationships directly from the available data without being explicitly programed to do so and without the assumption of stationarity. Hence, its growing applicability on environmental data (Dowd and Saraç, 1994; Kapagerdis, 1999; Tahmasebi and Herzakahni, 2012; Dai et al., 2014; Gangappa et al., 2017, Maniar et al., 2018, Tomislav, 2018, Samson and Deutsch, 2018; Samson and Deutsch, 2019 Walch et al., 2019). Samson (2019) proposed the use of an elliptical radial basis function network (ERBFN) to generate an interpolated model that is used as an exhaustive secondary information on a collocated cokriging framework. Samson (2019) found that, the use of the ML method is able to significantly improve accuracy on estimated models.

On the other hand, most of ML models are built on the whole training set available and then the fit model is used to predict new instances. But, when data is not evenly distributed in the input space, the function fit globally, with the entire training set, could have the generalization capability compromised. That is because, in the machine learning context, models are built to extract general properties of the data and not specificities of individual training points. Hence, global fitting sometimes affects negatively the generalization capability of the model fit (Galvan et.al, 2011). Moreover, ML learning methods such as neural networks, support vector machine and so on have parameters that must be defined prior to the training stage, to which fine tuning is challenging. Generally, the parameter is selected through cross validation or grid search. However, with the complexity of geological data set, the irregularity and sparsity of data samples can lead to suboptimal parameters to the problem at hand.

So, this paper presents a case study where the methodology of LWSVR (Ellatar et al., 2010) is applied to the data set to obtain an interpolated model to be incorporated on the geostatistical framework. The proposed approach builds locally weighted models in which the parameters are set automatically within each neighborhood, avoiding the optimizing process prior to training.

Case study

The case study is presented as a comparison between models built through ERBFN (Samson, 2019), traditional SVR (Vapnik, 1995,1998) and LWSVR (Ellatar et al.2010). The data set used is the Walker Lake data set (Issaks and Srivastava, 1989) where the variable V, herein rescaled, is retained to obtain an estimated model. The full data set statistics is presented in Table 1.

Table 1: Walker lake full sample set statistics.

Variable	Num. of data	Mean (%)	Std. deviation
V clustered	470	4.35	2.99
V declustered	470	2.88	2.56

From Table 1 is seen that the sample strategy affects the data set statics. The declustered mean is reduced in 34% relative to the original value. The sampling strategy in this case has been determined according to the ore grade value, with preferential sampling over regions that present high grades. It has been shown in (Silva and Costa, 2018) that sampling strategy can lead to data sets that are not representative of the phenomenon, which ultimately affects the final model. The data set is split into two training sets with 235 samples each. The location map for the full data set and both training sets are presented in Figure 1.

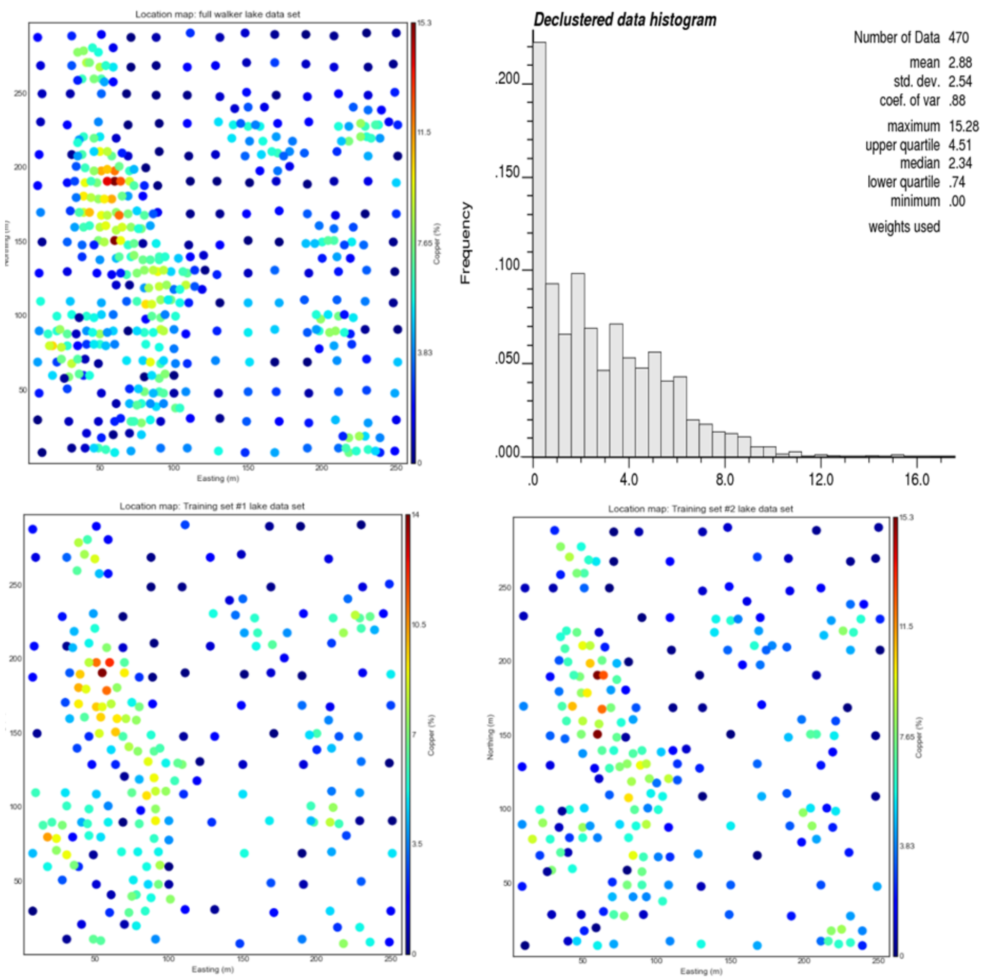

Figure 1 - Full data set (top left); training set for fold number 1 (bottom left); training set for fold number 2 (bottom right)

As it can be seen from Figure 1, the training sets present regions with very sparse data sampling, along with a great sample density on the west portion of the map, which corroborates the statistics drastic change when data declustering is applied. The full data mean value reduces 33% relative to the clustered value. Also, the data set presents outlier values, eight times greater than the sample data mean (2.88%). The goal is to interpolate the data on the training sets and evaluate the performance of the trend model generated. For that, LWSVR will be applied to the data and compared to two distinct global ML algorithms, ERBFN and traditional SVR.
The ERBFN algorithm proposed by Samson (2019) consists in an elliptical radial basis functions network applied in an ensemble context. One of the main parameters to be defined is the number of nodes for the hidden layer in the network, which greatly impacts model performance. To avoid over and under fitting different networks are trained with different parameters generating a different prediction relative to each network trained. The set of predictions is then ensembled to obtain a final model. The ERBFN algorithm applied to this case study trained four networks

with 25, 30, 40 and 50 nodes in the hidden layer respectively. The predictions of the 4 networks are ensembled to obtain the final model. Another important parameter in the ERNFN is the learning rate, which dictates the adjustment on the gradient descent function. A high learning rate does not allow to find an optimal minimum and a low learning rate may get the algorithm stuck in a local minimum. The learning rate defined is 0.01. Figure 2 presents the training process performed.

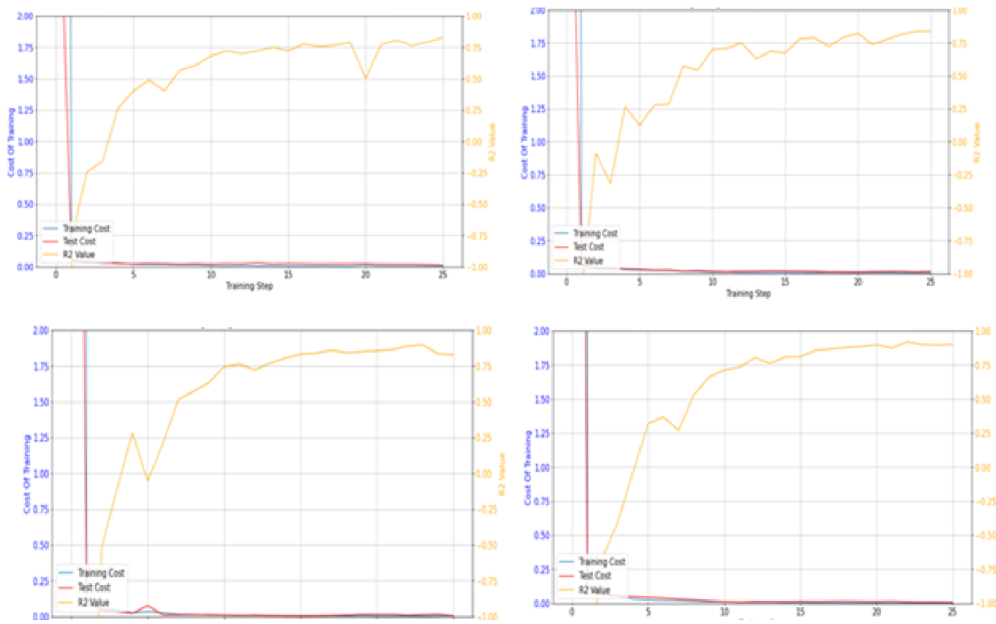

Figure 2 - Training performance of the ERBFN algorithm. Blue line: training cost; red line: test cost; yellow line: R^2 value

It can be seen from Figure 2 that the ERBFN training process is adequate for the data set, given that the training and test costs asymptotically go to zero while the R^2 value is over 0.75 for each network.

Also, the traditional setting of SVR is applied to the data set, hence a global SVR model is fitted to both training sets and evaluated. The SVR algorithm is applied with an RBF kernel, with a gamma value of 0.9 and $C = 20$. For the traditional SVR the parameters are defined through grid search. For details on the SVR algorithm refer to (Silva and Boisvert, 2020).

The LWSVR is applied with a search radius of 30m, with minimum of 2 samples per neighborhood. As for the kernel used on the local SVR models is also the RBF, the parameter C is defined dynamically inside each neighborhood according to the sample distance to the estimation location through the following system:

$$C_i = W_i C$$

With W_i obtained through

$$W_i = e^{-\left(\frac{d_{M_i}}{2\sigma}\right)^2}$$

σ represents the smoothing, which is also defined locally throughout the algorithm, by the following equation:

$$\sigma = \left(\frac{d_{E_{min}}(d_{E_{max}} - d_{E_{mean}})}{d_{E_{mean}}(d_{E_{max}} - d_{E_{min}})}\right)^2 + 1$$

Goodness of a trend model

Qu (2018) proposes as a metric of a trend goodness the mean squared error (MSER) obtained between the trend values and the sample data. Table 2 presents the mean squared error obtained on both folds and the overall average for each ML algorithm.

Table 2: MSER obtained on the trend models generated with ML algorithms over a 2-fold cross validation

MSER	SVR	LWSVR	ERBFN
Fold #1	10.51	6.1	5.6
Fold#2	7.79	5.3	5.8
Average	**9.15**	**5.7**	**5.7**

The results shown in Table 2 demonstrate a similar performance between ERBFN algorithm and LWSVR on average. ERBFN obtained 8% reduction on the MSER on the first fold, and obtained 9% higher MSER on the second fold. Both LWSVR and ERBFN outperformed SVR. It is important to highlight from this result the gain in performance that SVR obtained when applied locally to the data set. The SVR algorithm obtained an overall MSER of 9.15 while LWSVR obtained an average of 5.7, reducing the MSER in 38% relative to the global approach. Figures 3, 4 and 5 present the squared error (SER) location map for each methodology along with the validation set.

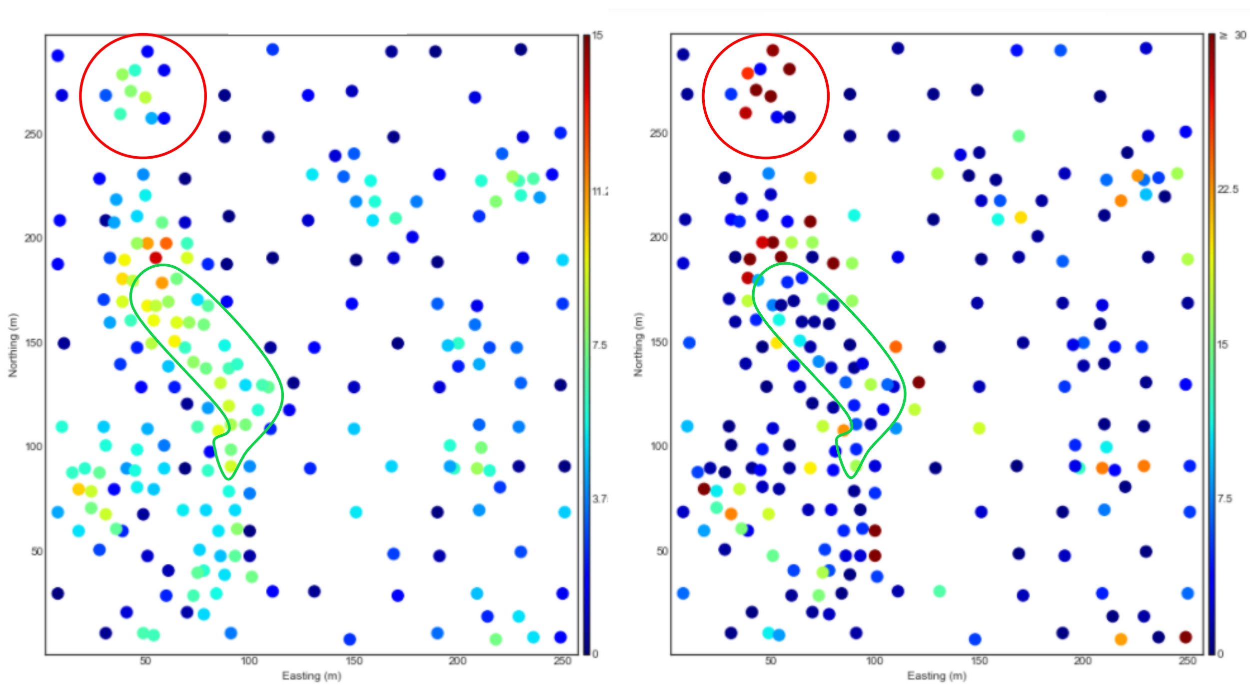

Figure 3: Validation set map (left); SER location map obtained from predicted values through SVR on validation set

Figure 3 shows validation set from fold number 2 and the SER obtained on prediction made though SVR on these locations. On the sample map (right) it is seen that the top North-West the portion highlighted in red, there is a cluster of samples with values that range from low to mid-grade values. The corresponding region on the SER map show high error values obtained. This is due to the heterogeneity on sample grades on this cluster, where the global model demonstrates to be suboptimal to the specificity of the subregion. In Contrast, on the center of the cluster on the west side, the portion highlighted in green, the sample values vary less than the before mentioned. As a result, on the SER map the errors obtained from the predicted values are lower. From the SER map, is seen that SVR is approximating local mean values, and therefore, obtaining poor predictions on the unsampled locations. According to Wang and Xu (2017), grid search is inefficient as an optimization algorithm due to the exhaustive search of the parameter space and computational time. Figure 4 presents the SER location for LWSVR.

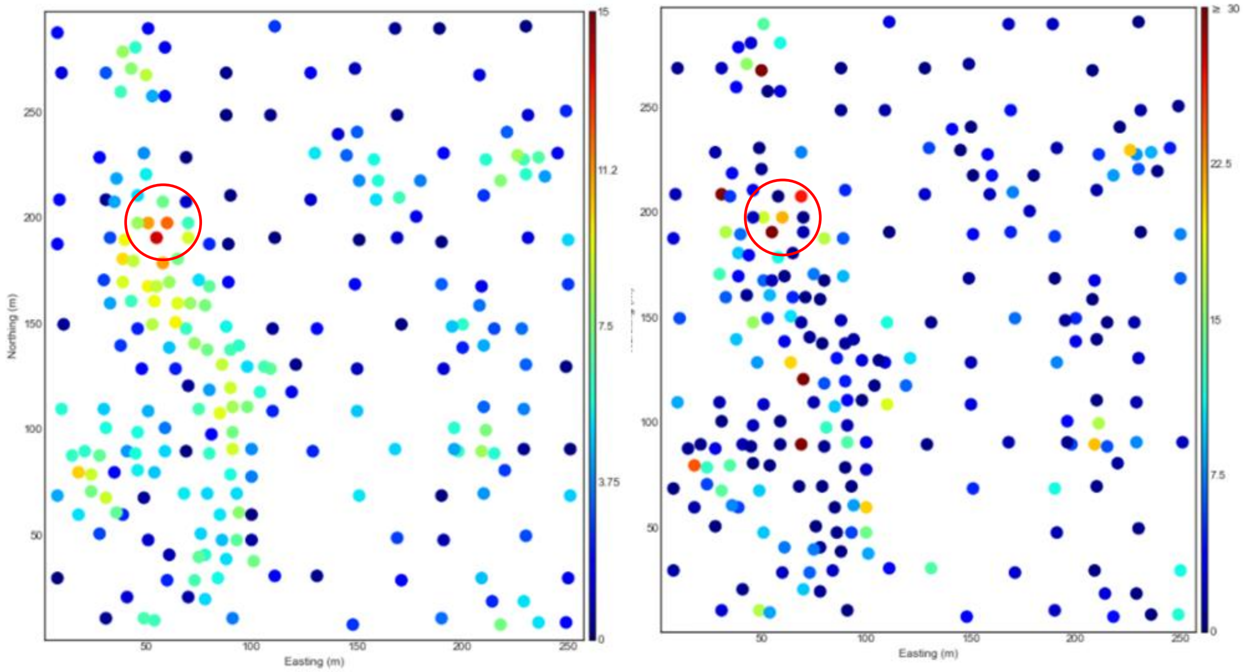

Figure 4: Validation set map (left); SER location map obtained from predicted values through LWSVR on validation set

In contrast to the behavior observed on SVR, LWSVR does not obtained such high errors related to neighborhood heterogeneity. From the SER map (right) it is possible to see that the highest errors obtained from LWSVR predictions are related to neighborhood outliers. For example, on the portion highlighted in red two samples are significantly different from that local neighborhood. There is a high-grade value over 12% concertation and a low-grade value, below 3% concentration. For both samples in the neighborhood the LWSVR obtained higher SER values. However, it did perform better than SVR on general sense. Figure 5 show the map obtained from the ERBFN algorithm.

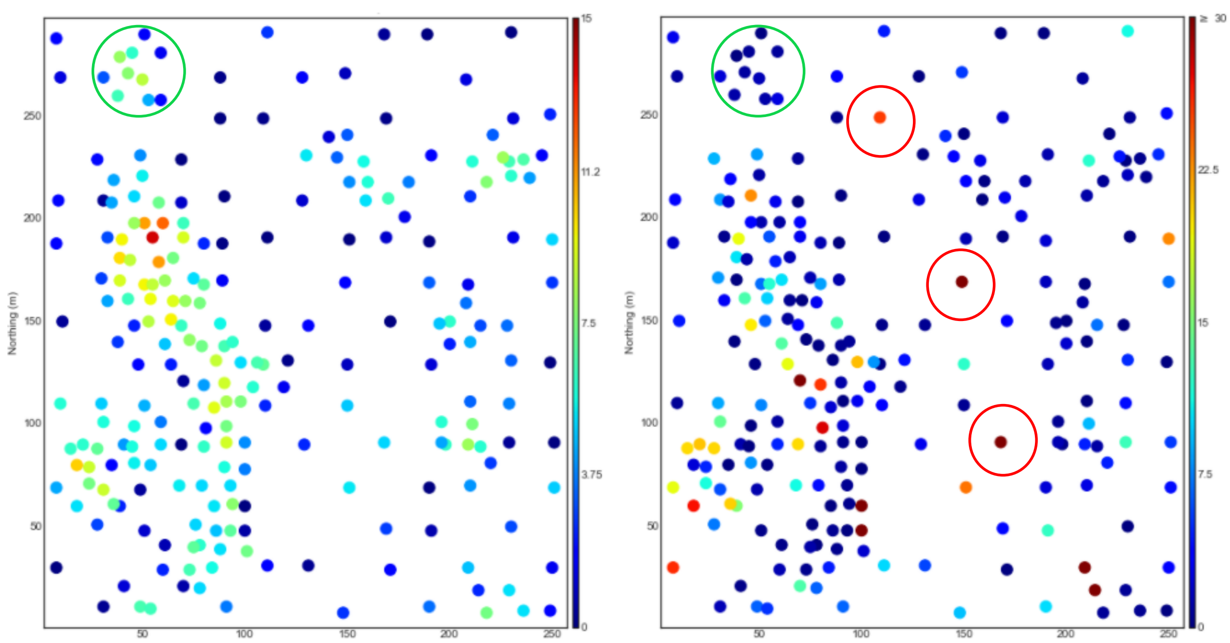

Figure 5: Validation set map (left); SER location map obtained from predicted values through ERBFN on validation set

For the ERBFN prediction on the validation set (Figure 5) is observed a good prediction pattern in a general sense. However, even though the global model is much more adequate than the obtained from SVR, in this case is possible to observe that sample sparsity has an important effect on prediction. In general, neural networks are data-greedy algorithms and this is translated on regions of the set where the sampling decreases significantly in number. The samples highlighted in red throughout the validation set obtained poor prediction due to the lack of information on those regions. In contrast to SVR the portion highlighted in green show that, even if the neighborhood is constituted by samples significantly different in grade values, with enough information the algorithm performs with a good prediction performance.

Another aspect of the goodness of trend is that the residuals be uncorrelated to the trend values at all locations. The residuals are defined as:

$$residual = z_i - z_i^*$$

Where z_i is the sampled value and z_i^* is the trend estimated value. Figures 6,7, and 8 show the cross correlogram obtained between the residuals and trend values.

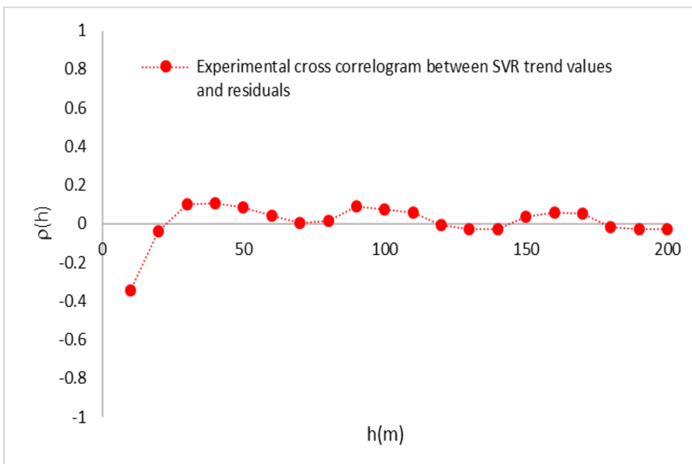

Figure 6: Experimental cross correlogram between the trend values obtained through SVR and the residuals

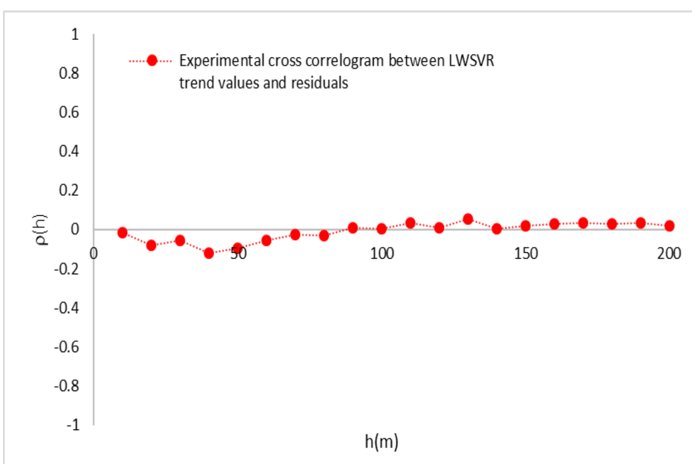

Figure 7: Experimental cross correlogram between trend models obtained through LWSVR and residuals

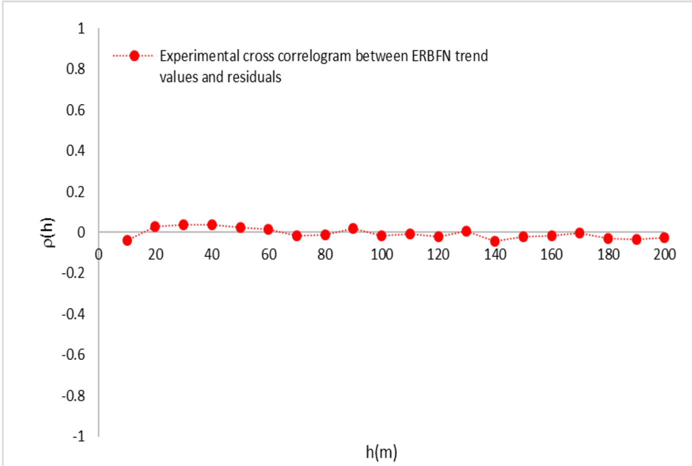

Figure 8: Experimental cross correlogram between trend values obtained through ERBFN and residuals

Figure 6 shows that the trend values obtained through SVR show a negative correlation with the residual values (-0.4) up to a 30m distance. From a radius of 30m the correlation is no longer observed. LWSVR (Figure 7) and ERBFN (Figure 8) obtained no correlation at all lags considered.

Collocated cokriging incorporating the trend models

Also, the performance of the geostatistical framework incorporating the interpolated values through ML algorithms is applied to evaluate the performance of the proposed approach in this context. Since ERBFN presented good results on previous studies (Samson and Deustch, 2019; Samson, 2019, Silva et al.,2020), LWSVR is incorporated through an analogous framework. To evaluate the applicability of LWSVR 2 models are built incorporating the trends obtained through ML (ERBFN and SVR). Also, an ordinary kriging model is generated for sake of comparison of the results obtained. The models are built on a 2-d grid, the cell size is 5.0m on X direction and 5.0 on Y direction. The entire grid comprises 52 cells on X and 60 cells on Y, covering in total 78.000m^2.

The spatial continuity for the V variable on the walker lake data set is presented on Figure 9. The experimental variogram is built using 20 lags of 10m, with a lag tolerance of 5m. Also, the anisotropy is explored along eight directions varying 22˚.

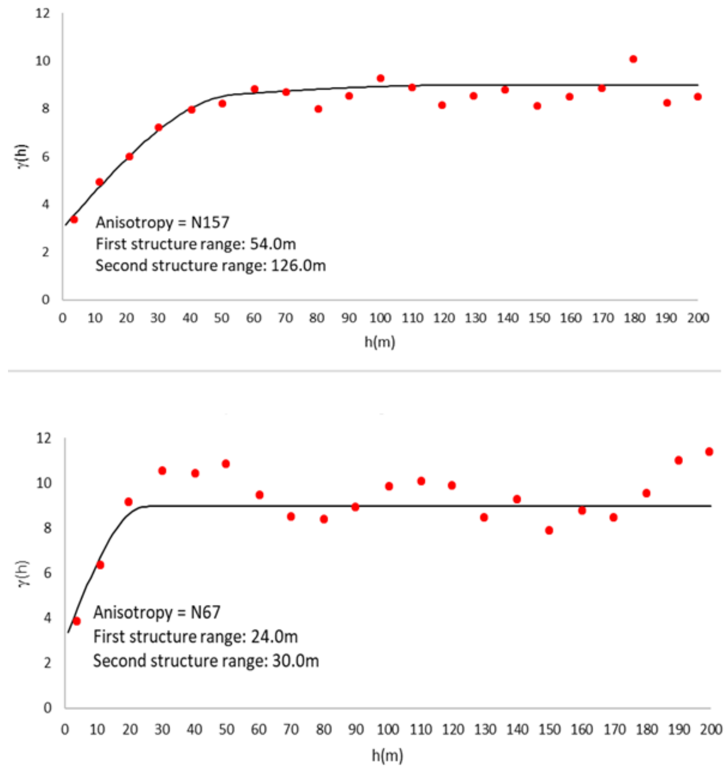

Figure 9: Experimental variogram for variable V. Top: Major anisotropy (N157); bottom: minor anisotropy (N67)

The ordinary kriging model requires the definition of a sample search strategy, which is shown in Table 3.

Table 3: Sample search strategy applied to estimated models

Maximum search radius	Minimum search radius	Minimum samples	Maximum samples	Maximum secondary if collocated
54.0	24.0	2	8	1

The two folds used on the previous section are also applied at this stage to evaluate each model performance through cross validation. Given the estimated grade models the performance of each methodology is presented on Table 4.

Table 4 - Comparison of the ranges of values obtained for some statistics in the 50 simulated models from each method and reference set, Walker Lake exhaustive.

	Mean	Minimum	Maximum	MSER fold#1	MSER fold#2	Average MSER
Original Data	2.8	0	15.2	------	------	---
OK	2.8	0	11.0 ↓27%	4.9	4.5	4.7
COK+ERBFN Trend	2.9 3.5%	0	12.38 ↓19%	4.5	4.5	**4.5**
COK + SVR Trend	2.9 3.5%	0	9.34 ↓38%	6.3	5.3	**5.8**
COK+ LWSVR Trend	2.9 3.5%	0	13.36 ↓12%	4.3	4.7	**4.5**

We see from the results obtained from cross validation of the estimated models, that the LWSVR obtained great improvement over the SVR model, reducing the MSER in 22%. Also, the LWSVR model obtained a comparative performance to the ERBFN algorithm. Also, it is seen that the SVR model obtained a smoother estimation reducing the maximum value of the distribution by 38%. However, the global declustered mean reproduction is the same obtained through every ML algorithm in collocated cokriging approach.

Is it important to highlight the following points in this case study: ERBFN is a known powerful ML, nonetheless its parameter setting is intrinsically related to the data complexity and do not have a clear physical interpretation. This characteristic imposes a challenge while setting the training phase of the algorithm, given that the final model and the generalization capability is greatly affected by the initial choice of parameters. Nonetheless, the results in Table 4 demonstrate the gain in incorporating this step throughout the geostatistical workflow.
SVR on the other hand, has not demonstrated globally the same prediction capacity as ERBFN, however, if the model is considered local, the results improve significantly. The LWSVR algorithm does not need the parameter setting, given that it adjusts dynamically inside each neighborhood. Based on the results from this case study, the local approach is attractive due to the fact that is produces a satisfactory model, with MSER lower than the obtained from ordinary kriging and equivalent to the one obtained through ERBFN with collocated cokriging, without the need to specifically set *a priori* the parameters for the ML model.

Conclusions

The case study has demonstrated the applicability of LWSVR on the geostatistical context, while maintaining a straightforward and simple workflow, with little user interference during the training stage of ML. The algorithm also has shown improvement on predictions with sparse data samples. Data sample sparsity often compromise the prediction performance of ML algorithms. Moreover, the approach of locally weghted ML has demonstrated in this case study that global traditional methods can benefit from the locality component when data is irregularly distributed over the input space. Local approaches can be adapted to a number of ML algorithms and will be further explored. In this case study, it has not been explored the influence of some parameters of SVR on local modeling such as, gamma in case of an RBF kernel. This is also is an aspect of further investigation.

References

Dai, F., Zhou, Q., Lv, Z., Wang, X. and Liu, G. (2014) Spatial prediction of soil organic matter content integrating artificial NN and OK in Tibetan Plateau. Ecological Indicators 45. https://doi.org/10.1016/j.ecolind.2014.04.003

Dowd, P. A. and Saraç, C. (1994). A NN Approach to Geostatistical Simulation. Mathematical Geology, 26(4).

Ellatar, E. E., Goulermas, J., Wu, Q.H.(2010) Electric load forecasting based on locally weighted support vector regression. IEEE Transactions on systems, man, and cybernetics—part c: applications and reviews. 40(4). https://doi.org/ 10.1109/TSMCC.2010.2040176

Galvan. I.M., Valls, J.M., Garcia, M., Isasi, P. (2011). A lazy learning approach for building classification models. International Journal of Intelligent systems 26(8). https://doi.org/10.1002/int.20493

Gangappa, M., Mai, C. K., Sammulal, P. (2017) Techniques for machine learning based spatial data analysis: research directions. International Journal of Computer Applications, 170(1). https://doi.org/10.5120/ijca2017914643

Kapageridis, I. K. (1999). Application of NNs systems to grade estimation from exploration data. (PhD). University of Nottingham.

Maniar, H., Srikanth, R., Kulkarmi, M.S., Schlumberger, A. A. (2018) Machine Learning methods in geoscience. Society of Exploration Geophysicists. International Exposition and 88th Annual Meeting. https://doi.org/10.1190/segam/2018-2997218.1

Qu, J. (2018). A Practical Framework to characterize non-stationary regionalized variables. PhD. Thesis. University of Alberta.

Samson, M. and Deutsch, C. (2018). Estimation with ML. CCG annual report 20. Paper 118.

Samson, M. and Deutsch, C. (2019). Elliptical Radial Basis Function vs. Radial Basis Functions. CCG annual report 21. Paper 110.

Samson, M. (2019). Mineral Resource Estimates with ML and Geostatistics. Master Thesis. Edmonton, Canada.

Silva, C. Z., Costa, J.F.C.L. (2018). The treatment of missing not at random geological data for ore grade modeling. Applied Earth Science. 128(1). https://10.1080/25726838.2018.157508

Tahmasebi, P. and Hezarkhani, A. (2012). A fast and independent architecture of artificial NN for permeability prediction. Journal of Petroleum Science and Engineering 86. doi: http://doi.org/10.1016/j.petrol.2012.03.019

Tomislav, H., Nussbaum, M., Wright, M. N., Heuvellink, G. B. M., Graler, B. (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ 6: e5518. https://doi.org/10.7717/perrj.5518

Vapnik, V. N. (1995) The nature of statistical learning theory. New York. Springer.

Vapnik, V.N. (1998) Statistical Learning Theory. New York. Wiley