SPS Model: a significant algorithm to reduce the time and computer memory required in geostatistical simulations

Behnam Sadeghi1,2

1. Earth Byte Group, School of Geosciences, University of Sydney, NSW 2006, Australia
2. Earth and Sustainability Research Centre, University of New South Wales, NSW 2052, Australia

Received 17 February 2020; accepted 24 July 2020

Abstract

In geochemical anomaly classification, different mathematical-statistical models have been applied. The final classified map provides only one scenario. This model is not certain enough since every model provides several thresholds which are almost different from each other meaning dissimilarity and spatial uncertainty of the classified maps. Spatial uncertainty of the models could be quantified considering the difference between the associated geochemical scenarios simulated (called: ‘realizations’) by geostatistical simulation (GS) methods. However, the main problem with GS methods is that these methods are significantly time-consuming, and CPU- and memory-demanding. To improve such problems, in this research, the method of “scaling and projecting sample-locations (SPS)” is developed. Based on the SPS theory, first of all, the whole sample-locations were projected (centralized) and scaled into a box coordinated between (0,0) to (150,0) and (0,0) to (0,100), for example (they can be equal though), with the cell-size of 1 m². Therefore, the time consumed and the memory demanded to generate a large number of realizations, for example, 1000 realizations based on the non-scaled/non-projected (NS/NP) and scaled/projected (S/P) sample locations per case-study were quantified. In this study, the turning bands simulation (TBSIM) were applied to geochemical datasets of three different case studies to take the area scales, regularity/irregularity and density of the samples into account. The comparison between NS/NP and S/P results statistically demonstrated the same results, however, the process and outputs of the S/P samples took a significantly shorter time and consumed a remarkably lower computer-memory. Therefore, experts are able to easily run this algorithm using any normal computer.

Keywords: spatial uncertainty, geostatistical simulation, scaling and projecting sample-locations (SPS) method, time, memory

1. Introduction

One of the most significant steps of geochemical exploration is recognition and classification of the geochemical anomalies, but as statistically certain as possible. One of the reasons why the certainty of the geochemical anomaly models generated in mining industry is quite important is the financial requirements. Considering this point, experts need to increase the certainty of the outputs (i.e., geochemical models) resulting in reduction of the mining and financial risks. Hence, to identify and classify the univariate geochemical anomalies, many mathematical/statistical methods have been applying such as traditional statistical (TS) methods (cf. Tennant and White 1959; Hawkes and Webb 1962; Sinclair 1974, 1983, 1991; Govett et al. 1975; Miesch 1981; Aucott 1987; Stanley and Sinclair 1989; Sinclair 1991; Grunsky et al. 1992; Harris et al. 1997, 1999; Kitanidis 1999; Chilès and Delfiner 1999, 2012; Davis, 2002; Mallet et al. 2005), exploratory data analysis (EDA) (Tukey 1977), weights of evidence (WoE) (Good 1950; Bonham-Carter et al. 1988, 1989; Agterberg et al. 1990), and fractal/multi-fractal and singularity models (cf. Mandelbrot 1983; Agterberg et al. 1990; Agterberg 1994, 2001; Cheng et al. 1994, 1999; Cheng 2007, 2012, 2015; Chen et al. 2007; Grunsky 2007, 2010; Zuo and Cheng 2008; Carranza 2009; Cheng and Agterberg 2009; Zuo et al. 2009, 2013; Carranza 2010a,b; Afzal et al. 2010, 2011, 2013, 2014; Zhao et al. 2011; Zuo 2011; Nazarpour et al. 2015a,b; Khalajnasumi et al. 2015, 2017; Sadeghi et al. 2012a,b, 2015, 2016, 2020; Wang et al. 2012; Xiao et al. 2012; Daneshvar Saein et al. 2013; Sadeghi et al. 2014; Sadeghi and Carranza 2015; Wang and Zuo 2015; Momeni et al. 2016; Sanchez and Sadeghi 2018; Agterberg 2018; Madani and Sadeghi 2019; Sadeghi 2020; Aliyari et al. 2020; Kourostabi et al. 2020; Solatani et al. 2020; Shamseddin Meigooni et al. 2020; Hajsadeghi et al. 2020; Pourgholam et al. 2021; Sadeghi and Cohen 2021). However, the outputs of all these models are only individual geochemical anomaly maps. If the outputs are generated using different methods such as the above-mentioned models, it turns out that none of the individual outputs are quite similar and they have differences in pixel values and thresholds. Simply, we can say even one pixel-difference could reflect a 100 m² or 1000 m² difference, for instance, considering the map scale. This fact demonstrates the spatial uncertainty of the individual geochemical anomaly models generated by each model. The aforementioned spatial uncertainty has significant sources (cf. Mann 1993; An et al. 1994; Kliir and Yuan 1995; Fisher 1999; Costa and Koppe 1999; Bárdossy and Fodor 2001, 2004; Walker et al. 2003; Kreuzer et al. 2008; McCuaig et al. 2007, 2009, 2010; Kiureghian and Ditlevsen 2009; Singer 2010; Singer and Menzie 2010; Caers 2011; Sheidt et al. 2018; Sadeghi 2020) such as...
sampling density, geochemical data analysis errors like interpolation errors, in addition to inappropriate geological and geochemical interpretations. In order to quantify the spatial uncertainty of geochemical anomalies recognized, geostatistical simulation (GS) methods, as one of the most robust methods, have been applying to generate more than one realization (i.e., geochemical anomaly maps in our field) and quantify the dissimilarity between them. The aim of this paper is not spatial uncertainty quantification, but one step ahead, i.e., simplification of the simulations to reduce the time and memory required to simulate a large of scenarios, known as realizations. Therefore, more realizations in a quite shorter time with smaller pixel-sizes (i.e., 1x1 m² instead of 100x100 m², for example) could be generated.

In summary, because the spatial GS methods, have been developed mainly to generate geochemical anomalies / concentrations and realizations, demonstrating the uncertainty of the phenomena (Caers 2011; Sheidt et al. 2018; Sadeghi 2020, 2021), their process would be remarkably time-consuming and memory demanding. Therefore, we need to look for a simpler shortcut to reduce the time and memory required. Considering this aim, in this paper “the scaling and projecting sample-locations (SPS)” method is proposed.

Fig 1. Sample locations in the study areas: a) Sweden till sample locations collected from almost 75% of the country area in a regular network, b) Moalleman stream sediment geochemical samples collected in a regular network, and c) Khooshab litho-geochemical samples collected in an irregular network.

2. Applied datasets
In this study, three different datasets have been studied to check if the SPS method is robust enough in different types of data with various densities, and regular and irregular sampling networks. The SPS method has been applied to the Cu element concentrations in all the three case studies although the method could be applied in multivariate analyses as well. However, in this research the focus is on the method, not the geochemical interpretations. The first dataset applied in this study has been provided by Geological Survey of Sweden (SGU) (scale: 1:100,000). The dataset includes the till samples collected from 75% of the whole country of Sweden, in a regular sampling network (Fig 1a). 2578 till samples have been collected by SGU, and the samples have been analyzed with the confidence level of 95% at the same laboratory. The density of the sampling is one sample per 150 km² with the distance of approximately 12.5 km between each two samples (Andersson et al. 2014). Although 66 elements had been analyzed using inductively coupled plasma mass spectrometry (ICP-MS), because in this research, the type of the elements is not the main issue, only the scale and sample locations have been taken into consideration.
The other datasets, which are provided by Geological Survey of Iran, are from the Moalleman and Khooshab 1:100,000 geological sheets. The former area is located in the Semnan Province (Central Iran), near to the Damghan city. In this area, 819 stream sediment geochemical samples have been collected in a regular sampling network, and Symmetrical sampling grid is based on taking a sample form a cell with 4 km2 area (Fig 1b). Around 70 elements have been analyzed using ICP-MS, although as mentioned above, we mainly focus on the sample locations in this study. The latter dataset belongs to the Khooshab area, which is situated in the North Khorasan Province (NE Iran). In this area, 230 litho-geochemical samples have been collected in an irregular sampling network (Fig 1c). Using these samples, 50 elements have been analyzed by ICP-MS. The maps have no coordinates because the method developed in this research is appropriate for spatial uncertainty quantification which could be calculated independently without using any coordinates, and only based on the dissimilarity of the realizations (Scheidt and Caers 2007; Scheidt et al. 2018).

3. The methodology of SPS
One of the advantages of the simulation methods is that they do not need the rasterized (regular or Cartesian) grids. In other words, simulation algorithms apply simulation on irregular grids such as point-sets. So, the conditioning data provided may or may not be on the simulation grid. If we apply the simulation algorithms on point-sets, it results in a performance penalty as the search for neighboring data, which is remarkably more important than on a Cartesian grid. It means, if the simulation grid is Cartesian, these simulation methods can relocate the conditioning data to the nearest grid node, then the simulation process time would be significantly decreased (Remy et al. 2009).

Given the fact discussed above, in the SPS method, we need to project the whole samples and their locations considering the origin coordinates (Fig 2). Then, it would be easier to calculate the distance between the realizations to quantify their statistical dissimilarity and after that the spatial uncertainty. It results in saving time and memory demanded for the geostatistical simulations. Now we need to transfer/standardize the whole sample locations to a smaller box within the coordinates such as (0,0) to (150,0), (0,0) to (0,100) and (150,100) (Fig 3 and Table 1). The maximum ranges could be anything small even the same. To do so, the general equations of 1 and 2 are proposed and applied to the dataset to project the available data into the new box.

\[
X_{\text{new}} = \frac{x_{\text{max}} - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \times (X_{\text{old}} - x_{\text{min}}) + x_{\text{min}} \quad (1)
\]

where \(x_{\text{max}}=150\) and \(x_{\text{min}}=0\); \(x_{\text{max}}\) and \(x_{\text{min}}\) are the maximum and minimum values of the whole initial values of the X coordinate, and \(X_{\text{old}}\) depicts each initial value of the X coordinate that is going to be scaled and transferred to the new coordinate \(X_{\text{new}}\).

\[
Y_{\text{new}} = \frac{y_{\text{max}} - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \times (Y_{\text{old}} - y_{\text{min}}) + y_{\text{min}} \quad (2)
\]

where \(y_{\text{max}}=100\) and \(y_{\text{min}}=0\); \(y_{\text{max}}\) and \(y_{\text{min}}\) are the maximum and minimum values of the whole initial values of the Y coordinate, and \(Y_{\text{old}}\) represents each initial value of the Y coordinate that is going to be scaled and transferred to the new coordinate, which is \(Y_{\text{new}}\).

![Fig 2. Schematic model of the sample locations. The model demonstrates how the real study area is scaled (i.e., 1 m² pixel size) and projected based on the SPS theory.](image)

4. Discussion
In geochemical data, the data distributions are mostly amorphous (i.e., with high entropy), so (semi)variogram-based methods (e.g., kriging interpolation and variogram-based simulations) are mainly applied to generate/simulate realizations (Chilès and Delfiner 1999, 2012; Remy et al. 2009). Given this point, in this study, to investigate that if the results of the SPS method are certain enough, both non-scaled/non-projected (NS/NP) and scaled/projected (S/P) outputs would be taken into account comparing their semi-variograms.

The semi-variograms demonstrate the similarity (i.e., variance) of the data points (y-axis) at the defined distances between each two points (x-axis). It means the variograms are the simplest but accurate statistical tools to indicate the relation between the uncertainty (i.e., dissimilarity of the data points’ simulated models) and the point distances in spatial analysis of the data samples (e.g., geochemical data samples) (Chilès and Delfiner 1999, 2012).

Considering the variograms generated based on the NS/NP and S/P data samples of the case study areas, the structures and specifically nugget effects and sills of the NS/NP and S/P data samples must be compared together to recognize if they are similar or close enough.
Fig 3. Projected and scaled sample locations: a) Sweden till samples, b) Moalleman stream sediment geochemical samples, and c) Khooshab litho-geochemical samples.

Table 1. The details of the NS/NP and S/P models of all the case studies.

<table>
<thead>
<tr>
<th></th>
<th>NS/NP</th>
<th>S/P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sweden</td>
<td>Moalleman</td>
</tr>
<tr>
<td>Max X-Coordinate</td>
<td>910838</td>
<td>317562</td>
</tr>
<tr>
<td>Min X-Coordinate</td>
<td>278302</td>
<td>272822</td>
</tr>
<tr>
<td>Max Y-Coordinate</td>
<td>7659696</td>
<td>3930649</td>
</tr>
<tr>
<td>Min Y-Coordinate</td>
<td>6140268</td>
<td>3886772</td>
</tr>
<tr>
<td>Length (X)= Max X-Coordinate - Min X-Coordinate</td>
<td>632536</td>
<td>44740</td>
</tr>
<tr>
<td>Coordinate</td>
<td>1519428</td>
<td>43877</td>
</tr>
<tr>
<td>Length (Y)= Max Y-Coordinate - Min Y-Coordinate</td>
<td>6325.36~</td>
<td>447.4~448</td>
</tr>
<tr>
<td>Coordinate</td>
<td>6326</td>
<td>6</td>
</tr>
<tr>
<td>Cell Size</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Number of Cells X= LX/Cell Size</td>
<td>6325.36~</td>
<td>447.4~448</td>
</tr>
<tr>
<td>Number of Cells Y= LY/Cell Size</td>
<td>15194.28</td>
<td>438.77~440</td>
</tr>
</tbody>
</table>
Table 2. Details of the semi-variograms generated based on the NS/NP and S/P data samples of Cu in the case studies.

<table>
<thead>
<tr>
<th></th>
<th>NS/NP</th>
<th></th>
<th></th>
<th>S/P</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sweden</td>
<td>Moalleman</td>
<td>Khooshab</td>
<td>Sweden</td>
<td>Moalleman</td>
<td>Khooshab</td>
</tr>
<tr>
<td>Number of lags</td>
<td>50</td>
<td>25</td>
<td>30</td>
<td>100</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Lag Separation</td>
<td>1000</td>
<td>1400</td>
<td>800</td>
<td>1</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Lag Tolerance</td>
<td>5000</td>
<td>700</td>
<td>400</td>
<td>0.5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Nugget effect</td>
<td>0.42</td>
<td>0.17</td>
<td>0.11</td>
<td>0.42</td>
<td>0.17</td>
<td>0.11</td>
</tr>
<tr>
<td>Sill 1</td>
<td>0.58</td>
<td>0.83</td>
<td>0.07</td>
<td>0.58</td>
<td>0.83</td>
<td>0.07</td>
</tr>
<tr>
<td>Max 1</td>
<td>255000</td>
<td>25550</td>
<td>2640</td>
<td>25</td>
<td>72.8</td>
<td>9.6</td>
</tr>
<tr>
<td>Sill 2</td>
<td>0.82</td>
<td></td>
<td></td>
<td>0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max 2</td>
<td>24000</td>
<td></td>
<td></td>
<td>2400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Details of the computer applied in this research.

<table>
<thead>
<tr>
<th></th>
<th>HP EliteDesk 800 G3 [1ME80PA] SFF Desktop PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand</td>
<td>Windows 10 Pro 64</td>
</tr>
<tr>
<td>Operating System</td>
<td>3.4 GHz</td>
</tr>
<tr>
<td>Processor Speed</td>
<td>7th Generation Intel® Core™ i5 processor</td>
</tr>
<tr>
<td>Processor Family</td>
<td>Intel® Core™ (5-7500 Processor (3.4 GHz, up to 3.8 GHz with Intel Turbo Boost, 6 MB cache, 4 cores))</td>
</tr>
<tr>
<td>Graphics</td>
<td>Intel® HD Graphics 630</td>
</tr>
<tr>
<td>Memory Slots</td>
<td>4 DIMM</td>
</tr>
<tr>
<td>Memory (RAM)</td>
<td>8 GB DDR4-2400 SDRAM (1 x 8 GB)</td>
</tr>
<tr>
<td>Standard Memory Note</td>
<td>Transfer rates up to 2400 MT/s</td>
</tr>
<tr>
<td>Internal Storage Type</td>
<td>256 GB SATA SSD Storage</td>
</tr>
</tbody>
</table>

Table 4. Time and memory consumed to generate 1000 realizations based on the Cu NS/NP and S/P data samples in the study areas.

<table>
<thead>
<tr>
<th></th>
<th>NS/NP</th>
<th></th>
<th></th>
<th>S/P</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sweden</td>
<td>Moalleman</td>
<td>Khooshab</td>
<td>Sweden</td>
<td>Moalleman</td>
<td>Khooshab</td>
</tr>
<tr>
<td>Time (s)</td>
<td>1,834,333,659,341</td>
<td>844,929,936</td>
<td>391,318,618</td>
<td>134,057</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory (KB)</td>
<td>(~21 days)</td>
<td>(~6.5 min)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moalleman area</td>
<td>4,916,846,453</td>
<td>1,925,193</td>
<td>369,816,554</td>
<td>148,951</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Khooshab area</td>
<td>5,942,991,925</td>
<td>971,216</td>
<td>744,585,849</td>
<td>119,164</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 and Figs. 4 to 6 demonstrate all the details obtained from the variograms of the NS/NP and S/P data sample per study area comparing to each other. In Sweden case study, based on the Figure 4 and its details (Table 2), the NS/NP and S/P data samples have spherical variograms including one structure and similar nugget effects of 0.42 and sills of 0.58. In addition, in the Moalleman area, both NS/NP and S/P geochemical data samples have also provided spherical variograms with one structure and the same nugget effects of 0.17 and sills of 0.83 (Table 2 and Fig 5). However, the NS/NP and S/P geochemical data samples of Cu in Khooshab provide spherical variograms with two structures. In this case, the semi-variograms of both NS/NP and S/P geochemical...
data samples have also the same nuggets and sills (Table 2 and Fig 6). Based on the variograms compared above, we can come up with this scenario that the common method of using the initial data samples for simulation and spatial uncertainty quantification of geochemical data could be replaced by the SPS method and could provide us with the same results and accuracy, but faster and by demanding less memory. For example, in this research, the Turning Bands Simulation (TBSIM) method, as one of the fastest and most accurate simulation methods (Chentsov 1957; Emery 2008b; Afzal et al. 2014; Emery and Lantuéjoul 2006; Paravarzar et al. 2015; Sadeghi et al. 2015; Sadeghi computer used for this research is a HP EliteDesk 800 G3 [1ME80PA] SFF Desktop PC (Table 3). 2020), was applied on the NS/NP and S/P formats of all the three datasets to generate 1000 realizations. The Based on Table 4, the time and memory consumed to generate the realizations based on the S/P data samples are significantly fewer than those of the NS/NP data sample realizations, no matter the datasets have regular or irregular sampling networks and the study areas are vast or limited.

Fig 4. Semi-variograms of Sweden samples, demonstrating the specific structures: a) NS/NP and b) S/P Cu samples.

Fig 5. Semi-variograms of Mouleman samples: a) NS/NP and b) S/P Cu samples.

Fig 6. Semi-variograms of Khooshab samples: a) NS/NP and b) S/P Cu samples.
5. Conclusions
In geostatistical simulations, the common and significant problem is the time required for the simulation process and the high memory demanded for complicated calculations during the simulation procedures. So, if a solution comes up with a simpler process but the same accuracy, it would be helpful to shorten the time and memory needed, which results in a faster simulation with even higher number of realizations. Therefore, even in the case of quantification of the uncertainty of the geochemical anomalies modelled, this solution could significantly accelerate the whole process. Considering the mentioned summary, the SPS method was developed in this research. SPS simplifies the spatial distribution of the samples mainly in terms of projecting sample locations considering the distance from the origin coordinates. In addition, because SPS scales the sample locations to a very smaller area (called ‘box’), the cell-sizes of the S/P sample locations would be 1 m², which is remarkably smaller and statistically more accurate than those of NS/NP sample locations.

Based on the results, the variograms of the NS/NP and S/P sample locations are similar in terms of their types, number of the structures, nugget effects and sills. It obviously demonstrates that the outputs of the NS/NP and S/P sample locations are statistically similar.

Considering the points mentioned above, the calculated time-consumed and consequently the memory demanded, we can apply SPS to different types of datasets, univariate and multivariate analyses, with either regular or irregular sampling networks in different scales, and consequently accelerate the simulation process, but keep the same accuracy. In addition, due to the lower memory required, the SPS outputs are readable/editable easily by any software to be visualized.

For the future studies, the author suggests developing and testing the SPS model in various conditions in 2D and 3D, e.g., in case of having anisotropy in regions, or having any block discretization.

References


Sadeghi B (2021) Evaluation of geochemical anomaly classification models based on the relevant uncertainties and error propagation per class to select the most robust model(s) for the follow-up exploration, EGU General Assembly: 19–30.


