SPECIAL ISSUE

Implementation of Geostatistical Algorithms

Clayton V. Deutsch[1]

Abstract The state of nature is uncertain. André G. Journel championed the concept that we must quantify our uncertainty, then make a risk-qualified decision. The theory for this concept applied to the earth sciences was developed by André and others over the past four decades. Theory is unquestionably important, but the implementation of theory in algorithms and practical workflows has always been an essential aspect of geostatistics. This motivated the geostatistical software library and other computer codes. The effort remains unfinished. The workflows for the quantification of geological uncertainty and risk-qualified decision making are achievable to a select few who are comfortable bending arcane software to their will. The routine calculation of probabilistic resources and risk qualified decision making requires a clarity of vision or, as André was fond of saying, an understanding of the essence. The essential steps required for constructing numerical models that effectively quantify our state of nature have emerged. The concepts for risk-qualified decision making are clear. Practitioners have technical solutions to quantify and manage geological risk, thus reducing the lost opportunity cost of low-quality geological models and inferior decisions

Keywords Parameter uncertainty · Geological uncertainty · Decision making · Sequential simulation

✉ Clayton V. Deutsch
cdeutsch@ualberta.ca

1   Centre for Computational Geostatistics, University of Alberta, 9211-116 Street, Edmonton, AB T6G 1H9, Canada

1 Introduction

This is a technical paper with a clear message to practitioners of numerical geological modeling for resource and reserve evaluation. This is also a personal viewpoint that reflects the author's experience. An important text that consolidated developments in the mid 1980s was Geostatistics in Five Lessons (Journel 1989): *It is my opinion that the main contribution of geostatistics has been and still is* implementation*, an essential follow-up step much too often forsaken by theoreticians.* The geostatistical software library and user's guide (GSLIB) (Deutsch and Journel 1992) was written to showcase the essence of geostatistical implementation. The efforts did not start there. The Fortran code from Mining Geostatistics (Journel and Huijbregts 1978) was punched in by many early practitioners including the author of this paper. Many others have advanced André's vision for implementation including SGeMS (Remy et al. 2011). The probabilistic view to geological uncertainty has extended to many disciplines including geophysics (Azevedo and Soares 2017). André's vision was an integrated theoretical/practical framework that leads to an understanding of geological variability, estimates of resources, a quantification of uncertainty at all relevant scales and support for decisions with a clearly specified position on risk.

André's vision is not fully implemented. Certain aspects were implemented almost immediately and the fingerprint of GSLIB has been found on many geostatistical studies; however, there were gaps in theory and practical understanding. The importance of parameter uncertainty was not understood; it is essential in practice. Without parameter uncertainty small scale variations cancel out and large-scale uncertainty is severely underestimated. The importance of a hierarchical workflow without branching in the simulation workflow was not understood. This criticism is not entirely fair; the one-for-one notion of sequential sampling was understood. Nevertheless, the interaction between parameters, data, surfaces, boundaries, categories and multivariate continuous properties was not formulated clearly. The challenge of decision making in the presence of uncertainty was underestimated. The use of a single kriged, localized or P50 model runs rampant decades after it should have been extinct. Finally, the required software environment for defendable and actionable models was not understood. A collection of student-driven incompletely-tested software should not be sent to the front line of modeling.

This paper consolidates selected practical developments since the 1980s that make it possible to realize André's vision. Quantifying geological uncertainty and making a risk-qualified decision are possible. Future theory will be developed and novel computational platforms will emerge, but a consolidated framework is available for practitioners. Future trends are likely to include extensive automation, implicit modeling, and real time updating with a wide variety of data. Machine learning and other data-driven quantitative tools developed in other fields will be adopted, but aspects of model-driven geostatistics will remain given (1) the wide spacing of drilling in most resource estimation projects, and (2) our general understanding of geological processes and analogue information from similar geological settings.

Some longstanding concepts of uncertainty quantification are recalled and presented in a modern geostatistical context. Implementation details for an accurate and precise quantification of uncertainty are reviewed. The disclosure of resource uncertainty is

clarified. Decision making in presence of uncertainty is reviewed. The thinking in the 1980s, led by André, was that the conversion to a probabilistic world view was just around the corner. The practical implementation of the required tools took longer than expected.

2 Concepts

Geological uncertainty is a consequence of geological variability at all scales and sparse sampling; it is inevitable. This uncertainty does not depend on the importance of any decision, but it can be reduced by careful consideration of all general and site-specific information. Predicting the uncertainty of one-offs with no context or history is challenging. Every deposit is unique and, in some sense, a one-off; however, our understanding of geological processes, availability of multiple drill data, and reconciliation with actual results makes the quantification of geological uncertainty more akin to a repeated event. The procedures to quantify geological uncertainty are understood and practical.

Many probabilistic predictions are for single events such as a politician being elected. A geological resource consists of many variables at many locations. The total number of random variables easily exceeds a million. The transfer of multivariable multilocation uncertainty through to resource uncertainty must be done with simulation, that is, the sampling of realizations. A large enough number of realizations are created and processed through resource calculations. The relationships between all variables and all locations are reproduced by the realizations. Simulation is the only viable approach to transfer geological uncertainty through to larger practical scales of relevance. The full combinatorial space of uncertainty is too large to consider. Between 50 to 10000 realizations are often considered to quantify uncertainty; 200 is suggested for large scale resource modeling.

Checking uncertainty is essential. Resource and reserve estimates may form the basis of decisions with large economic value. Checking should be done in hindsight as new data and production becomes available; however, checking of local uncertainty is possible with a "leave some out" approach. K-fold validation is commonly used. The data are divided into K folds (typically 5 or 10), then modeling proceeds K times leaving one fold of data out at a time. Good uncertainty is accurate and precise. Accuracy relates to correctness or closeness to the truth. For example, there should be 50% of the true values within the predicted interquartile range. Precision relates to the repeatability or narrowness of the predicted probability distributions. This could be assessed, for example, by variance or entropy. Once uncertainty is calculated and checked, a final set of realizations conditioned to all available data is constructed and passed forward.

Considering multiple realizations in all calculations is straightforward. There are two main applications (1) the calculation of resource and reserve estimates, and (2) optimization. Resource or reserve calculations are repeated on each realization. The expected value is taken at the end. The resource estimate is a single number that can be reported with the correct significant digits. A measure of uncertainty can accompany the resource estimate, but that is not recommended (more on this provocative com-

ment later in this paper). The resource should be classified with the relevant regulatory requirements and disclosed appropriately. Regarding optimization, the expected value over multiple realizations should enter the objective function. Risk could be considered with a loss or utility function. Many optimization engines are designed for one realization, but implementations are emerging that consider multiple realizations.

3 Geological Uncertainty

Rock properties are variable at all scales. A hierarchical modeling workflow is formulated for each site. The framework itself could be uncertain and alternative scenarios may be considered to account for this. Often, there is enough room within a well-crafted workflow to permit accurate and precise uncertainty quantification. The modeling workflow is a hierarchy that includes a combination of boundaries (solid or wireframe models), surfaces (subsurfaces and faults), categories (facies and rock types) and multivariate continuous rock properties. The techniques used at each step in the hierarchy require conditioning data, parameters and a sequence of random numbers to sample from uncertainty. The final model consists of a set of realizations. Each realization provides a unique set of rock properties at all locations within the study volume. Some special topics related to the quantification of geological uncertainty are reviewed here.

3.1 Parameter Uncertainty

Each technique in the specified hierarchical modeling workflow requires modeling parameters such as distributions, variograms, training images, and correlations. These parameters are used to compute uncertainty, but are themselves uncertain. Assessing and transferring such parameter uncertainty through the modeling workflow is more important than imagined. The GSLIB book and related texts recommended against an open-ended assessment of parameter uncertainty that would neither make the results more objective nor be data-driven. Experience has shown that parameter uncertainty is essential for accurate large scale resource uncertainty. First order parameters such as proportions, mean values and histograms are particularly important. Parameters that affect continuity may be important for flow-based responses. Constructing multiple realizations with fixed parameters leads to an unrealistically narrow distribution of resource uncertainty. Areas of high and low values tend to cancel out and resources calculated over a large volume are very similar.

The original bootstrap procedure is not applicable for spatial data; in almost all cases the data values from geological sites are correlated and cannot be considered independent. The spatial bootstrap was proposed shortly after the bootstrap, but counter intuitive results were distracting. Greater spatial correlation leads to larger uncertainty, yet one would expect less uncertainty because the data are more informative. Also, the results of the spatial bootstrap are independent of the study volume or domain limits, yet we expect more uncertainty for a large volume, provided everything else remains the same. A combination of research, examples with very large data sets, and case studies have shown that the results of a two-step procedure are correct (Journel and Bitanov 2004; Rezvandehy and Deutsch 2017). The first step is to calculate parameter

uncertainty with the spatial bootstrap that considers neither conditioning data nor the domain boundaries. The second step is to simulate within the domain boundaries with all available conditioning data and realizations of the input parameters coming from the spatial bootstrap. The second step reduces the spatial bootstrap uncertainty, but in a way that gives correct final results.

The implementation details of quantifying and considering parameter uncertainty have become established. Multiple realizations of parameters are simulated—one for each realization being simulated; each simulated geological model has a different set of input parameters. The reproduction of statistics is checked for the ensemble of final realizations. If any bias appears in the checking, then multiple realizations with fixed input parameters are checked. If problems remain, then unconditional realizations are checked. Problems are almost always attributable to non-stationary trends or complex spatial clustering of the data.

3.2 One for One Model Setup

The concepts of probability trees and decision trees are very useful, but they can promote a misunderstanding about the quantification of uncertainty. Branching the sampling of uncertainty is not a good idea—the combinatorial of models increases uncontrollably for no good reason (Deutsch 2017). A decision tree is completely valid and very useful; however, the quantification of geological uncertainty is a different problem that cannot be approached with a tree-branching concept. Given the tree concept, for each realization of surfaces, there could be multiple realizations of categorical variables; for each categorical variable realization there could be multiple realizations of the primary continuous variable, and so on. If the tree-branching concept is taken to an extreme, then for each location simulated, there could be multiple realizations for the next location. This would lead to an impossibly large number of realizations. André's pioneering concept of sequential simulation is much more general than for indicators and Gaussian regionalized variables. The concept applies to the entire workflow of model setup and simulation.

Additional realizations could be added at any time, but it is convenient to choose the number of realizations at the start. This facilitates simplicity and the parallelization of calculations. Conceptually each realization consists of one set of parameters, one set of data, one boundary model, one surface model, one categorical variable model and one model for each continuous variable. In practice, the parameters, data, boundaries, surfaces, categories and continuous variables are simulated one after another. The workflow is scripted and ordered so that each realization of each regionalized variable is simulated with the correct parameters and framework created by regionalized variables earlier in the model setup. The simulation is truly sequential with no branching. In model post processing and decision making, branching is included in decision trees as appropriate.

The one-for-one modeling of geological variables is straightforward to implement. Different coordinate systems are used at different steps to conform the directions of continuity to local geological conditions. This is particularly true for tabular stratigraphic or vein-type deposits. At the end of the modeling procedure all realizations

of the geological model are represented as a block model in real-world coordinates. The block model may include subblocks for improved representation of complex geological boundaries. All post processing calculations are performed on all realizations. For example, the probability of high quality resources could be computed by counting the number of realizations that are high-quality and dividing by the total number of realizations. This probability may not achieve 100% away from the drill holes in presence of surface and boundary uncertainty. Engineering design decisions would accommodate flexibility for this uncertainty.

The current approach to one-for-one modeling includes data uncertainty. There are two approaches to transfer data uncertainty through geostatistical modeling. The first approach is closely related to data imputation, that is, realizations of the underlying true data values are simulated for all available data, then they are used as inputs to subsequent steps. The second approach is to have the geostatistical algorithm consider multiple data types at the same time. If there are different types of drill data, the first imputation framework has proven itself. If there are near exhaustive secondary data coming from, say, seismic, then the second approach has established itself. In any case, quantifying and transferring data uncertainty through the modeling workflow through to resource and reserve uncertainty can be important.

The implementation details for a full one-for-one modeling workflow are well understood. Data imputation requires knowledge of the geostatistical modeling parameters; therefore, parameter uncertainty is assessed first. Data are imputed for missing values and/or data with error. Cokriging can be formulated to avoid data imputation and to consider data with error; however, the decorrelation techniques often used multivariate modeling are applicable only for equally sampled data sets. Imputation is a practical solution. The hierarchical workflow is applied where each simulated geological model has a different set of input parameters and data. A geological model typically consists of a boundary model, constraining surfaces, categorical variable realizations within different zones and multiple variable realizations within each category.

3.3 Useful Algorithms

As described above, the overall framework for probabilistic resource modeling is established; however, different techniques, algorithms and software tools could be used for different tasks. The chosen technique will depend on site specific geological conditions, available data, expertise, available software and the scope of the project. Three recent developments are noteworthy for their usefulness and rapid adoption in practice: (1) the hierarchical truncated pluriGaussian (HTPG) technique for categorical variable modeling, (2) trend modeling and the stepwise conditional transformation with a Gaussian mixture model (SCT-GMM) for modeling with a trend, and (3) the projection pursuit multivariate transformation (PPMT) for multivariate modeling.

Categorical variable modeling within volumes delimited by bounding surfaces and implicit boundary models are important. Sequential indicator simulation (SIS) cannot reproduce complex relationships and is prone to bias in the simulated proportions. Multiple point statistics (MPS) requires an elusive training image and tuning parameters such as multiple grid searches and correction parameters for proportions. The

family of truncated Gaussian (TG) techniques are mathematically consistent, but can be difficult to parameterize. SIS, TG and MPS have their place, but the hierarchical truncated pluriGaussian (HTPG) (Silva and Deutsch 2019) technique has emerged as a flexible and robust approach. A hierarchical tree structure is easy to understand and permits geologically realistic transitions and cross cutting relationships to be considered. Many Gaussian variables (five to fifteen or even more) can be used with unique spatial structure and accessible truncation rules for realistic modeling. Parameter and data uncertainty are naturally included in the HTPG modeling workflow.

Continuous variable modeling is almost always required within the categories or domains that are deemed stationary. Trends introduce challenges in workflows for probabilistic resource estimation. There are rarely enough data to enforce the large-scale trend-like features in simulated realizations; high and low values will overly influence locations where they do not belong. Kriging for the trend has not been successful because kriging aims for data reproduction and tends to the global mean at the margins of a domain. Geological variables are rarely amenable to a simple polynomial or functional trend shape. A weighted moving window average has proven effective. Some important implementation details to consider: (1) a length scale for the moving window specified for the primary direction of greatest continuity, (2) a Gaussian shape to the weighting function, (3) anisotropy in the kernel length scale somewhat less than that of the regionalized variable—often a square root of the anisotropy ratio to the maximum direction of continuity, (4) the weight to each data is the kernel weight multiplied by the declustering weight, and (5) a small background weight to all data of, say, one percent. The only free parameter is the length scale in the primary direction. Despite some worthwhile attempts to automate the calculation of this parameter it is set by experience and the visual appearance of the final model. A value one third of the domain size may be reasonable. Once the trend is modeled, we must simulate with the trend.

Creating a residual as $\{R(\mathbf{u}) = Z(\mathbf{u}) - m(\mathbf{u}), \mathbf{u} \in A\}$ is not good practice. Z and m are related in complex ways causing R and m to be dependent. If R is modeled independently then artifacts will be introduced in the $R+m$ back transform. A stepwise conditional transform (SCT) of the original variable conditional to the trend has proven effective to completely remove the dependence on the trend. Independent modeling proceeds and the reverse transform introduces the dependency between the original variable and the trend. The SCT considers a fitted Gaussian Mixture Model (GMM) between the normal scores of the original variable and the normal scores of the trend (Silva and Deutsch 2016). These normal score transforms are an intermediate step and are easily reversed. The use of a Gaussian mixture model makes the stepwise transform entirely bin-free and artifact-free. The workflow of trend modeling, data transformation, simulation and back transformation can be largely automated.

Multiple continuous variables perhaps detrended with the SCT-GMM technique are to be simulated simultaneously within multiple categories/domains. Cokriging based techniques account for correlations and linear relationships; however, real data almost always show complex relationships such as non-linear features, constraints and the proportional effect even after univariate normal score transformation. The projection pursuit multivariate transformation (PPMT) (Barnett et al. 2013) has emerged as a useful pre- and post-processor for simulation. The collocated multivariate complexity

between multiple variables is removed in a pre processing step. Conventional Gaussian simulation techniques are used for each independent factor. Then, the original units and complex collocated features are restored in the post processing back transformation.

In combination with many other techniques, the three useful algorithms mentioned in this section (HTPG, SCT-GMM, and PPMT) are used routinely in practice. The construction of multiple realizations accurately and precisely representing geological uncertainty is possible. The concepts, one-for-one workflow and useful tools are part of the modeling effort. Considering the resulting realizations in resources/reserves reporting and decision making comes next.

4 Risk Qualified Decision Making

Resources and reserves are directly reported from the realizations that quantify geological uncertainty. It is not good practice to have one deterministic model for base case resource estimates, then consider simulated realizations to provide a measure of uncertainty. There is a need, however, for a single resource number to report. The single resource estimate is the expected value of the resource computed across all realizations. There is no computational complexity in calculating the resources on, say, 200 geological models. The variation of the calculated resources provides a measure of geological uncertainty; the expected value provides the best base-case single estimate. This best estimate is not traced to a single model, but the underlying realizations are unique and the resource estimate can be validated from the unique set of realizations that quantify the true uncertain state of nature.

The best estimate of resources and reserves will not likely be achieved. The actual outcome could be better or worse than expected. There are consequences for better or worse outcomes. This does not mean that lower than expected resource and reserve numbers should be reported. Regulatory guidelines for disclosure require classification based on geological confidence and other mitigating factors; this should be done considering the quantified uncertainty. The public disclosure of uncertainty is not required; classification is required. A company that discloses a range of resources could be at a disadvantage in the marketplace. Most people are naturally risk averse and would value an asset by the lowest number that is reported. Until uncertainty is reported as an industry wide standard, best practice will therefore quantify geological uncertainty for expected resources and to support classification decisions. The uncertainty quantified in the geological model should also be used in decision making.

There are different types of decisions. Simple decisions are binary, that is, to participate or not in a venture. The venture may be to invest in a project, drill a well or send a truck load of rock to the mill as ore. The binary view is (1) positive, that is, an irrevocable commitment of resources in favor of the project, or (2) negative, that is, giving up an opportunity to participate in a project. Risk is the unfavorable consequences of an inadequate return after a positive decision or a significant lost opportunity after a negative decision. The asymmetric consequences of an inadequate return versus a lost opportunity leads a decisionmaker to take a risk averse or an opportunity seeking position on risk. There could be a list of decisions to choose from or the space of possible decisions could be combinatorially large, that is, the possible locations of

multiple wells or the sequencing of a mining operation. The position on risk will be discussed before some comments on risk qualified decision making.

4.1 Position on Risk

Taking a risk neutral stance will always lead to the best decision in expected value, that is, if many similar decisions are aggregated together. Risk aversion leads to decisions that avoid low value situations, but at the expense of capitalizing on high-value high-variance outcomes. Opportunity seeking leads to decisions that prefer high value situations, but at the expense of low value outcomes. A neutral position on risk is preferable for repeated similar decisions. There will be unfavorable outcomes, but the consequences average out and the decisionmaker will realize the greatest value. In situations of infrequent or one-off decisions a rational decisionmaker would be risk averse for medium-term decisions and opportunity seeking for exploration or long-term decisions.

The two main approaches to encode a non-neutral position on risk are to minimize expected loss or maximize expected utility. These are not equivalent and not directly related to each other. Loss functions were introduced in geostatistics by André G. Journel in the mid 1980s. Loss functions are suitable for choosing a value from a distribution. The concept of expected utility, however, is amenable to choosing from a list of options and for optimization (Zakamouline 2014). In many decision making settings, the distribution of uncertainty changes throughout an optimization process. The design parameters are chosen to optimize the distribution of uncertainty. This makes the use of utility more intuitive and straightforward.

The value of a decision could be expressed in terms of the net quantity of a commodity produced or the net present value. A utility function converts the value to utility to quantify (1) a preference to avoid low values (risk aversion) or (2) a preference to seek high values (opportunity seeking). A risk averse exponential utility function of the form $u(v) = 1 - exp(-av)$ where u is utility, v is value and a is a constant that represents our position on risk. When the parameter of risk aversion a or the form of the utility function is unclear, decision making should proceed with alternatives including a risk neutral decision. Then, the alternative optimized decisions can be reviewed. Decision making would consider expected utility and not expected value. Care must be taken with negative value outcomes ($v < 0$) and for distributions of value that span several orders of magnitude. In practice, the value for all possible decisions is scaled to be within a fixed positive range.

4.2 Decision Making

Selecting from a list of millions (or less) of alternative decisions is straightforward: compute the expected utility for each possible decision, perhaps considering alternative utility functions as a sensitivity, then choose the decision with the maximum expected utility. Decisions that are not final should consider a risk neutral option. A trained professional should review the optimized decision since there are often considerations not easily quantified in a simple utility function.

Many decisions have a solution space that is combinatorially large. In petroleum, the number of possible well locations and completion strategies is essentially infinite. In mining, the possible pit designs or stope sizes and sequencing is also essentially infinite. These type of decisions call for optimization, but the objective function should be expected utility and not expected value (Acorn and Deutsch 2018). There are other ways to embed risk in the decision, but directly optimizing expected utility or directly penalizing risk is the best alternative. The software implementations for this are limited.

5 Thoughts on Implementation

Open source code is essential for researchers and advanced practitioners to understand algorithms, consider alternatives, and explore new ideas. Such open source code could form the basis for commercial implementations, provide comparative tests and document implementation problems. Open source code provides a basis for detailed debugging of algorithms and some assurance of reproducibility and longevity for important algorithms.

GSLIB and other open source software initiatives that André G. Journel championed have been influential in numerical geological modeling, uncertainty quantification and decision making. Public domain student-driven code should be used with great care for commercial applications. In general, the coding standards, version control, reproducibility, flexibility, testing and support are inadequate. Some practitioners have the experience to include such research codes in a commercial workflow, but that is dangerous. The risk that buggy code causes untrapped errors is too high. The use of GSLIB-like code may be required for a period of time; software vendors are reluctant to commercialize unproven techniques. Even the best research ideas would remain unproven without some adventurous practitioners applying research codes.

Commercial software is recommended and preferred for commercial resource estimates, reserves optimization and decision making. Platforms for combined open-source and commercial applications may emerge. A key is to have highly qualified personnel applying the software and validating the results with a critical eye.

6 Conclusions

The important missing links reviewed in this paper are parameter uncertainty, a hierarchical one-for-one modeling workflow and optimization of decisions with expected utility. Some useful algorithms that grew from the seeds of early geostatistics are also described. This paper is a testament to the vision of André G. Journel for geological modeling and resource/reserve decision making. Appreciation for the complexity of implementation was incomplete in the 1980s and early 1990s. The essence is better understood and a computing environment is available for present day application. The room for future developments is vast, but we are, at least, able to provide accurate and precise estimates of resource uncertainty and make risk qualified decisions.

References

Acorn T, Deutsch CV (2018) Optimizing pit shells in the presence of geologic uncertainty with a heuristic algorithm. Trans Soc Min Metal Explor 344:38–47

Azevedo L, Soares A (2017) Geostatistical methods for reservoir geophysics. Springer International Publishing, New York

Barnett RM, Manchuk JG, Deutsch CV (2013) Projection pursuit multivariate transform. Math Geosci 46(3):337–359

Deutsch CV (2017) All realizations all the time. In: Sagar BSD, Cheng Q, Agterberg F (eds) Handbook of mathematical geosciences: fifty years of IAMG. IAMG, Houston, pp 131–141

Deutsch CV, Journel AG (1992) GSLIB: geostatistical software library and user's guide. Oxford University Press, New York

Journel AG (1989) Fundamentals of geostatistics in five. Volume 8 short course in geology. American Geophysical Union, Washington, DC

Journel AG, Bitanov A (2004) Uncertainty in n/g ratio in early reservoir development. J Petrol Sci Eng 44:115–130

Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic Press, New York

Remy N, Boucher A, Wu J (2011) Applied geostatistics with SGeMS: a user's guide. Cambridge University Press, New York

Rezvandehy M, Deutsch CV (2017) Geostatistical modeling with histogram uncertainty: confirmation of a correct approach. Nat Resour Res 18:7

Silva D, Deutsch CV (2016) Multivariate data imputation using gaussian mixture models. Spat Stat 20:1012

Silva D, Deutsch CV (2019) Multivariate categorical modeling with hierarchical truncated pluri-gaussian simulation. Math Geosci 51(1):527–552

Zakamouline V (2014) Portfolio performance evaluation with loss aversion. Quant Finance 14:699–710