2023

Multivariate cross-validation and measures of accuracy and precision

Ute Mueller
Edith Cowan University

Sangga Rima Roman Selia

Raimon Tolosana-Delgado

10.1007/s11004-022-10040-y
Mueller, U., Selia, S. R. R., & Tolosana-Delgado, R. (2023). Multivariate cross-validation and measures of accuracy and precision. *Mathematical Geosciences*, 55, 693-711.
https://doi.org/10.1007/s11004-022-10040-y

Mathematical Geosciences
https://doi.org/10.1007/s11004-022-10040-y

SPECIAL ISSUE

Multivariate Cross-Validation and Measures of Accuracy and Precision

Ute Mueller[1] · **Sangga Rima Roman Selia**[2,3] · **Raimon Tolosana-Delgado**[2]

Received: 30 August 2022 / Accepted: 1 December 2022

Abstract
Cross-validation and performance measures are standard components in the evaluation of a geostatistical model. These are well established in the univariate case, but measures for multivariate geostatistical modeling have not received as much attention. In the case of a single target variable, the univariate approaches remain valid, but in the fully multivariate case where a vector of variables needs to be estimated, the evaluation needs to be based on all estimates simultaneously. An extension of cross-validation and associated performance measures to the fully multivariate case is presented and discussed for the case of regionalized compositions. The method is demonstrated by validating geostatistical models for two case studies: a sample drawn from a geochemical survey data set estimated with cokriging, and an application of direct sampling multiple-point simulation.

Keywords Geostatistical simulation · Model validation · Compositional data

1 Introduction

Cross-validation and jackknifing are established methods for validating statistical models. In a geostatistical context, the model is based either on geostatistical estimation via kriging or on spatial simulation. The standard outputs include scatterplots and correlation coefficients of estimated against true values, and (possibly standardized) estimation errors against estimates, accompanied by error statistics (Webster and Oliver 2007). These are widely used to validate and optimize model parameters

✉ Raimon Tolosana-Delgado
r.tolosana@hzdr.de

1 Edith Cowan University, Joondalup, WA 6027, Australia

2 Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resources Technology, 09599 Freiberg, Germany

3 Geophysical Engineering Department, Universitas Sylah Kuala, 23111 Banda Aceh, Indonesia

Published online: 12 January 2023

or to determine the most suitable model from a set of competing models. In addition, scatterplots of coverage probabilities versus theoretical values may be used to check the quality of the posterior distributions derived from the model (Deutsch 1997; Olea 2012). These approaches were initially formulated for the estimation/simulation of univariate random functions or for cases where a clear primary variable is to be modeled, with one or more covariates which are of secondary importance. However, in the case of fully multivariate data such as directional (van den Boogaart and Schaeben 2002a, 2002b) or compositional data (van den Boogaart and Tolosana-Delgado 2013; Pawlowsky-Glahn and Egozcue 2020), the entire regionalized vector is seen as an entity which needs to be modeled rather than just its component parts. As a consequence, geostatistical estimation and simulation need to be treated as fully multivariate, and any appraisal of the quality of the geostatistical model needs to take this aspect into account. This concerns all statistical results mentioned before: error statistics, correlation between predictions and observations, and accuracy of estimated intervals.

In this contribution, a generalization of accuracy in the sense of Deutsch (1997) is proposed for the multivariate setting. The proposal is analogous to the method described by Olea (2012) for quantifying the quality of the estimated distribution. Of specific interest is the evaluation of the suitability of a geostatistical estimation or simulation model in the compositional framework, namely, where each variable is non-negative, and its values inform of the relative abundance of a certain component forming the system (Tolosana-Delgado et al. 2019).

After this introduction, four further sections follow. In Sect. 2 the fundamentals of compositional data analysis are recalled briefly along with their implications in geostatistics. Section 3 reviews the existing proposals for univariate validation, with methods, diagrams, and statistics commonly used for this task. In Sect. 4 a fully multivariate approach to cross-validation is proposed for vector-valued random functions specified on the example of compositional data, both for cokriging outcomes (Sect. 4.1) and cosimulation (Sect. 4.2). Two case studies are presented in Sect. 5 illustrating these two sets of techniques. Conclusions are provided in Sect. 6.

2 Regionalized Compositions and Their Geostatistical Treatment

A regionalized composition is a set $\{\mathbf{z}(u) = [z_1(u), \ldots, z_D(u)] : z_k(u) \geq 0, k = 1, \ldots, D; \sum_{k=1}^{D} z_k(u) = c, u \in \mathcal{A}\}$ of compositional data defined on some study region $\mathcal{A}$, where $u \in \mathcal{A}$ denotes a location in $\mathcal{A}$, and c is an arbitrary but fixed constant. To avoid problems arising out of the fact that compositional data are closed to that constant sum and formed by non-negative components, compositions are usually transformed prior to any geostatistical or statistical treatment. Several logratio transformations are commonly used. These include the centered (clr; Aitchison 1986), additive (alr; Aitchison 1986) and isometric (ilr; Egozcue et al. 2003) logratio transforms. However, the choice of logratio transformation does not impact the final results, because the geostatistical techniques discussed here are affine-equivariant (Filzmoser and Hron 2008; Tolosana-Delgado et al. 2019). Affine equivariance implies that $\mathbf{m}(\mathbf{Z}B) = \mathbf{m}(\mathbf{Z})B$ and

$$S(\mathbf{Z}B) = B^T S(\mathbf{Z}) B, \tag{1}$$

where $\mathbf{m}$ denotes the mean, S a covariance matrix, and B a linear transformation. In spite of this invariance, it is mathematically convenient to use the ilr transformation in the geostatistical workflow (Pawlowsky-Glahn and Egozcue 2020). The corresponding regionalized composition of ilr-transformed variables will be denoted by $\{\boldsymbol{\zeta}(u) = \big[\zeta_1(u), \ldots, \zeta_{D-1}(u)\big] : u \in \mathcal{A}\}$. The image space of the ilr transformation is $(D-1)$-dimensional Euclidean space.

The standard workflow then is known as the principle of working in coordinates:

1. Transform the regionalized composition to logratios using a suitably chosen ilr transformation (Egozcue et al. 2003; Tolosana-Delgado and Mueller 2021) $\boldsymbol{\zeta}(u) = \ln \mathbf{z}(u) V$, where V is a $D \times (D-1)$ matrix with $V^T V = I_{D-1}$ and $V V^T = I_D - \frac{1}{D} 1_{D\times D}$.
2. Apply the geostatistical technique to the logratios.
3. Backtransform the geostatistical estimate or realization of the logratio scores, $\boldsymbol{\zeta}^*(u)$, to the compositional space via $\boldsymbol{z}^*(u) = \mathcal{C}(\exp(\boldsymbol{\zeta}^*(u) V^T)$, where $\mathcal{C}(\cdot)$ denotes the closure operation defined as

$$\mathcal{C}(\boldsymbol{x}) = \frac{c}{\boldsymbol{x} 1_D^T} \boldsymbol{x}.$$

The Aitchison geometry version of the Mahalanobis distance is of fundamental importance for the methods introduced in this paper. With a covariance matrix S as defined above, the (square) Mahalanobis distance between two compositions in the Aitchison geometry is

$$d^2_{AM}(\mathbf{z}_\alpha, \mathbf{z}_\beta | S) = \big[\mathrm{ilr}(\mathbf{z}_\alpha) - \mathrm{ilr}(\mathbf{z}_\beta)\big] S^{-1} \big[\mathrm{ilr}(\mathbf{z}_\alpha) - \mathrm{ilr}(\mathbf{z}_\beta)\big]^T, \tag{2}$$

analogous to the (square) Aitchison distance

$$d^2_A(\mathbf{z}_\alpha, \mathbf{z}_\beta) = \big[\mathrm{ilr}(\mathbf{z}_\alpha) - \mathrm{ilr}(\mathbf{z}_\beta)\big]\big[\mathrm{ilr}(\mathbf{z}_\alpha) - \mathrm{ilr}(\mathbf{z}_\beta)\big]^T = d^2\big(\mathrm{ilr}(\mathbf{z}_\alpha), \mathrm{ilr}(\mathbf{z}_\beta)\big). \tag{3}$$

Note that, although d_{AM} is more comfortably defined in terms of the ilr-transformed scores, it is an affine-equivariant quantity, hence intrinsic to the composition and the covariance S, and not dependent on the actual logratio transformation being used. One must merely represent S in the same transformation used for the composition, according to Eq. (1). This is not true of the Aitchison distance (Eq. 3), which holds only in terms of the ilr or clr transformations. With the Aitchison–Mahalanobis distance, one can define the additive logistic normal distribution (ALN) as the probability model with density proportional to

$$f_Z(\mathbf{z}|\mathbf{m}, S) \propto \left(\prod_{i=1}^{D} z_i\right)^{-1} \det(S)^{-1} \exp\left(-\frac{d^2_{AM}(\mathbf{z}, \mathbf{m}|S)}{2}\right). \tag{4}$$

This probability density function is, as required, also affine-equivariant, owing to the determinant being one of the invariants of S.

3 Cross-Validation, Accuracy and Precision in the Univariate Case

Leave-one-out cross-validation is a well-established tool used to assess a given geostatistical model and to determine the best model from a set of competing models. It is based on kriging, and given a sample data set at each location u, a kriging estimate $z_K^*(u)$ and corresponding error variance $\sigma_K^{2*}(u)$ are derived by removing the location from the set and estimating the value of the variable of interest based on the neighboring data. In jackknifing, a separate validation set is assumed to be available, and the sample data are used to provide estimates or simulated values at the jackknife locations. A more general cross-validation approach, known as n-fold cross-validation, is to partition the data set into n disjoint subsets and then apply jackknifing on each of the subsets based on the remaining sample data. Independently of the validation method actually used, one finally has a paired list of observed and estimated values, as well as their kriging variance.

If the true value at u is $z(u)$, then the error is given by $e(u) = z(u) - z_K^*(u)$, and the squared deviation ratio is defined as

$$\mathrm{sdr}(u) = \frac{\left(z(u) - z_K^*(u)\right)^2}{\sigma_K^{2*}(u)} = \frac{e^2(u)}{\sigma_K^{2*}(u)}.$$

Averaging these quantities over all sample locations results respectively in a mean error (ME) and mean squared deviation ratio (MSDR). The former and the associated mean square error (MSE) should be close to 0 and the latter close to 1 if the geostatistical estimator is adequately defined (Webster and Oliver 2007). Typical diagnostic plots are a scatterplot of the true values against the estimates, a histogram of the standardized errors and a scatterplot of the standardized errors against the estimates. Analogously to the linear model, here one expects a tight scatter between estimates and true values, a symmetric histogram of standardized errors with mean close to 0 and variance close to 1 (as a weakened version of standardized normality), and the scatter between standardized errors and estimates showing no correlation (Chilès and Delfiner 2012; Webster and Oliver 2007). Cross-validation is routinely performed in geostatistical practice and used for the appraisal of the variogram and trend model, the local neighborhood and parameters associated with the simulation method applied.

Performance measures in addition to ME, MSE and MSDR concern the quality of the local posterior distributions. Their assessment was first discussed in the context of univariate geostatistical simulation by Deutsch (1997) and is based on the coverage of the local distributions. If the local distribution is F_u with mean $\mu(u)$ and standard deviation $\sigma(u)$, then the indicator function

$$i(u, p) = \begin{cases} 1 \text{ if } z(u) \in \left[F_u^{-1}\left(\frac{1-p}{2}\right), F_u^{-1}\left(\frac{1+p}{2}\right)\right] \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{5}$$

defined for $p \in (0, 1]$ allows measurement of the closeness of the true value to the local mean in the context of a symmetric target distribution. If the simulation algorithm is Gaussian, then the parameters of the local distribution are derived via kriging; otherwise, the local distribution is inferred via the generation of a family of simulated values at the sample location leaving the actual value out. The coverage $\pi(p)$ is set equal to the average of $i(u, p)$ over all available locations u.

To determine the accuracy of the model, a further indicator variable $a(p)$ is introduced which is set to 1 if the proportion $\pi(p)$ of locations falling into the p-interval exceeds p, and to 0 otherwise. The integral $A = \int_0^1 a(p)dp$ then provides a measure of accuracy. That is, A measures the proportion of exact or over-pessimistic confidence intervals around the estimates. A useful means for appraising the accuracy of the model is a plot of $\pi(p)$ against p. An accurate model will result in a plot where the pairs of points $(p, \pi(p))$ fall above the bisector line. Two measures of precision are defined in Deutsch (1997), one restricted to pairs of points $(p, \pi(p))$ falling above the bisector line, called precision and defined as $P = 1 - 2\int_0^1 a(p) \cdot (\pi(p) - p)dp$, and the other, called goodness, given by $G = 1 - \int_0^1 (3a(p) - 2)(\pi(p) - p)dp$. For any value of p for which the point $(p, \pi(p))$ lies below the bisector, the departure of $\pi(p)$ from p is penalized in this definition, and high values of G correspond to precise models. Thus, an accurate and precise model has values of A, P and G close to 1. It is important to note that for high accuracy, it suffices that the actual coverage is larger than the nominal one (i.e., $p < \pi(p)$), while precision and goodness also reward proximity to the bisector (i.e., $|p - \pi(p)|$ small). The precision measure P only makes sense in the case of accurate models, while goodness G is more generally useful.

An alternative approach for assessing the quality of the local posterior distributions was introduced by Olea (2012). In his approach, the symmetric p-interval used in Eq. (5) is replaced by the unilateral interval $\left(-\infty, F_u^{-1}(p)\right)$. For each $p \in (0, 1)$ and each u, an indicator variable is defined by putting

$$i_O(u, p) = \begin{cases} 1 \text{ if } F_u^{-1}(p) > z(u) \\ 0 \quad \text{otherwise} \end{cases}.$$

For each p, the empirical probability $p^*(p)$ is set equal to the average of $i_O(u, p)$ over all available locations u and plotted against p. The modeling is optimal if the pairs $(p, p^*(p))$ fall on the bisector line, and according to Olea (2012) indicates "perfect global agreement between the modeling of uncertainty and the limited amount of information provided by the sample." In practice, however, there will be deviations from the bisector, and they can be quantified via the maximum absolute deviation and the sum of absolute deviations between the empirical and theoretical probabilities. It should be noted that the functions $\pi(\bullet)$ and $p^*(\bullet)$ are related by $\pi(1 - \alpha) = p^*\left(1 - \frac{\alpha}{2}\right) - p^*\left(\frac{\alpha}{2}\right)$.

4 The Compositional Case

Here, the implementation depends on whether or not cokriging is applied to derive the parameters of the local distribution. In this case it is assumed that the composition follows an additive logistic normal distribution as expressed by Eq. (4), at least locally, that is, conditional on the estimated composition. As usual in geostatistics, this assumption cannot be formally tested, owing to the presence of spatial dependence, and is to be considered a modeling choice. Otherwise, if conditional additive logistic normality is deemed inappropriate, one should use either multipoint methods or else some form of multivariate transformation to normality (e.g., Barnett et al. 2014; van den Boogaart et al. 2017; Sepulveda et al., under review) followed by Gaussian cosimulation, in which cases Sect. 4.2 applies.

4.1 Cross-Validation, Accuracy and Precision via Cokriging

For compositional data sets, the implementation of the cross-validation procedure via cokriging is straightforward, even if not implemented in most software packages (Tolosana-Delgado and Mueller 2021). At a location to be cross-validated, the entire compositional vector is removed and the surrounding compositions are used to estimate the composition at the sample location via cokriging. As the cokriging is performed in logratio coordinates, the error measures are also calculated in terms of logratios. The mean error is given by

$$\mathbf{ME} = \frac{1}{N}\sum_{\alpha=1}^{N}\left(\boldsymbol{\zeta}(u_\alpha) - \boldsymbol{\zeta}^*_{CK}(u_\alpha)\right) = \frac{1}{N}\sum\nolimits_{\alpha=1}^{N}\ln[\mathbf{z}(u_\alpha)/\mathbf{z}^*_{CK}(u_\alpha)]V^T,$$

and the associated mean square error is $\mathrm{MSE} = \frac{1}{N}\sum_{\alpha=1}^{N}\|\boldsymbol{\zeta}(u_\alpha) - \boldsymbol{\zeta}^*_{CK}(u_\alpha)\|^2$, which corresponds to the average Aitchison distance (Eq. 2) between estimates and observations. There are two ways to generalize the mean square deviation ratio to a multivariate quantity (Tolosana-Delgado and Mueller 2021), namely

$$\begin{aligned}\mathrm{MSDR}_1 &= \frac{1}{N}\sum_{\alpha=1}^{N}\left(\mathbf{1}(u_\alpha) - \mathbf{1}^*_{CK}(u_\alpha)\right)\Sigma^{-1}_{CK}(u_\alpha)\left(\mathbf{1}(u_\alpha) - \mathbf{1}^*_{CK}(u_\alpha)\right)^T \\ &= \frac{1}{N}\sum_{\alpha} = 1^N \ln\left(\mathbf{z}(u_\alpha)/\mathbf{z}^*_{CK}(u_\alpha)\right)V^T\Sigma^{-1}_{CK}(u_\alpha)V\ln\left(\mathbf{z}(u_\alpha)/\mathbf{z}^*_{CK}(u_\alpha)\right)^T\end{aligned} \tag{6}$$

and

$$\mathrm{MSDR}_2 = \frac{1}{N(D-1)}\sum_{\alpha=1}^{N}\sum_{i=1}^{D-1}\|\zeta_i(u_\alpha) - \zeta^*_{CK,i}(u_\alpha)\|^2/\sigma^2_{ii}(u_\alpha). \tag{7}$$

In the equations above, $\mathbf{z}(u_\alpha)$, $\boldsymbol{\zeta}(u_\alpha)$, $\boldsymbol{\zeta}^*_{CK}(u_\alpha)$ and $\mathbf{z}^*_{CK}(u_\alpha)$ denote the true compositional vector, its logratio image, the logratio estimate and the corresponding backtransform at location u_α. The expression $\ln\left(\mathbf{z}(u_\alpha)/\mathbf{z}^*_{CK}(u_\alpha)\right)$ is an abbreviation of $\left[\ln\left(z_1(u_\alpha)/z^*_{CK,1}(u_\alpha)\right), \ldots, \ln\left(z_D(u_\alpha)/z^*_{CK,D}(u_\alpha)\right)\right]$. The matrix $\Sigma_{CK}(u_\alpha)$ denotes the cokriging error variance–covariance matrix at location u_α, and $\sigma^2_{ii}(u_\alpha)$ are its diagonal elements.

Only the mean error **ME** is a vectorial quantity. All other quantities (MSE, MSDR_1 and MSDR_2) are scalars. The measure MSDR_1 is nothing other than the average over the square Aitchison–Mahalanobis distances (Eq. 2) $d^2_{AM}(\mathbf{z}(u_\alpha), \mathbf{z}^*_{CK}(u_\alpha)|\Sigma_{CK}(u))$ between $\mathbf{z}(u_\alpha)$ and $\mathbf{z}^*_{CK}(u_\alpha)$ with respect to the cokriging error variance–covariance matrix. Its target value is equal to $D-1$. Moreover, under the hypothesis of additive logistic normality of the $D-$ component compositional random function, the square Aitchison–Mahalanobis distance follows a $\chi^2(D-1)$ distribution. The version of MSDR in Eq. (7) is the average over the univariate MSDR values for the components of the logratios. In contrast to MSDR_1, this measure does not have the equivariance property.

The relationship between the Aitchison–Mahalanobis square distances and the χ^2- distribution gives rise to a diagnostic tool that may be used in place of the histogram of standardized errors of the univariate case. This is a qq-plot of the observed quantiles of the Aitchison–Mahalanobis square distances against the quantiles of the χ^2- distribution with $D-1$ degrees of freedom. As in the univariate case (where the comparison is against the standard normal distribution), one expects the qq-plot to be close to the bisector. Even if multivariate additive logistic normality does not hold locally for the compositional random function, this diagram will still provide a means to rank competing geostatistical parameter setups (mostly variogram models or kriging neighborhoods) in the sense of their approximation to this distributional assumption, exactly in the same way as for the univariate case.

Other common diagnostic plots are scatterplots of the individual components against their estimates, or of the estimation errors versus the estimates. Compositional versions of these plots are formed by the set of scatterplots between pairwise logratios of estimates against pairwise logratios of true values, and a set of scatterplots of logratio estimation errors against logratio estimates (Tolosana-Delgado and Mueller 2021); both can be conveniently presented in a $(D \times D)$ matrix of scatterplots.

Compositional analogues of accuracy and goodness described in Sect. 3 are also based on the square Aitchison–Mahalanobis distance of estimates and true values with respect to the estimation error covariance. The definition of the indicator variables required for the calculation of coverage is subject to a certain arbitrariness, because there is no natural ordering of vectorial quantities. In general, any one-dimensional summary of the random composition can be used to generate coverage indicators, as long as the probability distribution of this target quantity is known. A reasonable requirement is for these quantities to be affine-equivariant. The square Mahalanobis distance (Eq. 2) arises naturally as the best option: for each location u and each $p \in (0, 1]$, the indicator function $i_{AM}(u, p)$ is defined as

$$i_{AM}(u,\,p) = \begin{cases} 1 \text{ if } \chi^2(d^2_{AM}\big(\mathbf{z}(u), \mathbf{z}^*_{CK}(u)|\Sigma_{CK}(u)\big), D-1) \le p \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases},$$

analogously to Olea's (2012) proposal. As in the univariate case, the coverage $\pi_{AM}(p)$ is defined as the average over all sample locations, and the indicator variable $a(p)$ is equal to 1 for $\pi_{AM}(p) > p$ and 0 else. The metrics A, P and G then have the same definitions as previously, and an accurate and precise model has values of A, P and G close to 1. Other one-dimensional summaries can also be of use: It should be noted that the univariate measures discussed in Sect. 3 may be computed for each relevant logratio variable, each one of them being univariate summaries of the random composition. Even the original variables could be considered in this sense appropriate univariate summaries, if one were ready to obtain the confidence intervals in Eq. (5) by means of Hermite quadrature, as explained in Pawlowsky-Glahn and Olea (2004).

4.2 Cross-Validation, Accuracy and Precision via Simulation

When cross-validation or jackknifing needs to be based on simulation at the sample locations, the definitions of the errors and also the local distributions are based on the simulation results. As before, the errors are in the first instance calculated in logratio coordinates. For L realizations $\big\{\boldsymbol{\zeta}^\ell(u_\alpha)|\ell = 1, \ldots, L, \alpha = 1, \ldots, N\big\}$, we define the local mean

$$\overline{\boldsymbol{\zeta}}(u_\alpha) = \frac{1}{L}\sum\nolimits_{\ell=1}^{L}\boldsymbol{\zeta}^\ell(u_\alpha)$$

and the local covariance as

$$\widehat{\Sigma}(u_\alpha) = \frac{1}{L^2}\sum\nolimits_{\ell=1}^{L}(\boldsymbol{\zeta}^{\updownarrow}(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha))^T(\boldsymbol{\zeta}^{\ell}(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha)).$$

Then, analogously to the cokriging case, one has the mean error given as

$$\mathbf{ME} = \frac{1}{N}\sum_{\alpha=1}^{N}\big(\boldsymbol{\zeta}(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha)\big)$$

and

$$\mathrm{MSDR}_{1,sim} = \frac{1}{N}\sum_{\alpha=1}^{N}\big(\boldsymbol{\zeta}(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha)\big)\widehat{\Sigma}^{-1}(u_\alpha)\big(\boldsymbol{\zeta}(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha)\big)^T,$$

$$\mathrm{MSDR}_{2,sim} = \frac{1}{N(D-1)}\sum_{\alpha=1}^{N}\sum_{i=1}^{D-1}\frac{\big(\zeta_i(u_\alpha) - \overline{\zeta}_i(u_\alpha)\big)^2}{\widehat{\Sigma}_{ii}(u_\alpha)}.$$

In contrast to Eq. (6) where the target value was $(D-1)$, the value of $\mathrm{MSDR}_{1,sim}$ for a good model should be slightly greater than this quantity, owing to the fact that $\widehat{\Sigma}^{-1}(u_\alpha)$ is estimated from a set of L realizations. A reasonable target quantity can be derived from the expected value of $\mathrm{MSDR}_{1,sim}$ under the assumption that $\boldsymbol{\zeta}(u_\alpha)$ is normally distributed conditionally on $\overline{\boldsymbol{\zeta}}(u_\alpha)$, which produces a Hotelling's T^2 distribution for the Mahalanobis distance, with parameters $(D-1)$ and $(L-1)$. This gives an expected value of

$$E\left[T^2_{(D-1,L-1)}\right] = \frac{(D-1)(L-1)}{L-D+1} E\left[F_{(D-1,L-D+1)}\right] = \frac{(D-1)(L-1)}{L-D+1} \times \frac{L-D+1}{L-D-1} = (D-1)\frac{L-1}{L-D-1},$$

thanks to the equivalence between Hotelling's T^2 and Fisher F-distributions and the fact that the expected value of a Fisher $F_{(p,q)}$-distributed variate is $q/(q-2)$.

For deriving the coverage indicators to calculate accuracies and derived quantities, we can follow the same ideas as in the section about cokriging: one needs to select a univariate summary of the composition, compute that statistic for the true value and generate its probability distribution with the realizations. Again, in general terms, it makes sense to enforce that summary statistic to be affine-equivariant. The indicator variable defining the position of the true compositional vector within the local distribution can then be based on the square Aitchison–Mahalanobis distance: the distance $d_0(u_\alpha)$ between the true composition and the mean is compared with the empirical distribution of the distances $\{d_\ell(u_\alpha) : \ell = 1, 2, \ldots, L\}$ of the simulated results and the mean at u_α

$$d_0(u_\alpha) = \left(\boldsymbol{\zeta}(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha)\right)\widehat{\Sigma}(u_\alpha)^{-1}\left(\boldsymbol{\zeta}(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha)\right)^T,$$

$$d_\ell(u_\alpha) = \left(\boldsymbol{\epsilon}^\ell(u_\alpha)\right)\widehat{\Sigma}(u_\alpha)^{-1}\left(\boldsymbol{\epsilon}^\ell(u_\alpha)\right)^T,$$

where $\boldsymbol{\epsilon}^\ell(u_\alpha) = \boldsymbol{\zeta}^\ell(u_\alpha) - \overline{\boldsymbol{\zeta}}(u_\alpha)$ for each $\ell = 1, 2, \ldots, L$.

The values $\{d_\ell(u_\alpha) | \ell = 1, \ldots, L\}$ are arranged in ascending order $\left\{\widehat{d}_\ell(u_\alpha) | \ell = 1, \ldots, L\right\}$, and

$$i(u_\alpha, p) = \begin{cases} 1 \text{ if } d_0(u_\alpha) \leq \widehat{d}_{\lceil pL \rceil}(u_\alpha) \\ 0 \qquad \text{otherwise} \end{cases},$$

where $\lceil pL \rceil$ denotes the smallest integer greater than pL. The statistics A, P and G are then defined as in Sect. 4.1. This construction mimics that of Deutsch (1997) in that in the univariate case, the simulated values are used to derive a local distribution, and the location of the true value is determined relative to it. Note that, as in the case of cokriging, other uni-dimensional summaries may also be meaningful for specific cases: the simulation approach outlined here is particularly useful in these cases because the probability distribution of the target summary can always be derived from the set of realizations.

5 Illustration

In what follows, two applications are provided. The first demonstrates the cross-validation for cokriging of a regionalized subcomposition of the Tellus data (Young and Donald 2013), while the second concerns cross-validation and accuracy assessment in the case of direct simulation for the modeling of the structure of a tailing storage facility.

5.1 The Tellus Data Set

The composition considered in this example (Fig. 1) consists of the components MgO, Al_2O_3, CaO and Fe_2O_3 from a sample of the Tellus soil horizon A data (Young and Donald 2013; Tolosana-Delgado and Mueller 2021).

The subcomposition was closed through the inclusion of an additional component called Rest and transformed to logratios via the default ilr transform, as described in Tolosana-Delgado et al. (2019). The ilr variables exhibit geometric anisotropy with direction of greatest continuity N135 and a linear model of coregionalization comprised of a nugget, and an exponential structure of range 35 km was fitted, with an anisotropy ratio of 0.4. For the sake of comparison, an isotropic version of this model was also considered, with a range of 26.9 km. Tenfold cross-validation via ordinary cokriging with a moving neighborhood (search radius 60 km, minimum number of samples: 7, maximum number of samples: 20) was applied with both models.

A summary of the performance measures in Table 1 indicates that the anisotropic model is superior to the isotropic one. This is also supported by the graphs of coverage against confidence level in Fig. 2. Additionally, the actual coverage of the compositional model is greater than theoretically expected for the majority of the confidence levels, as evidenced by the value of A compared to those of the individual components. The accuracy plots in Fig. 2 provide further insight on the behavior of coverage versus confidence for the components and the entire composition. The coverage is generally closer to the chosen confidence level for the individual components compared to the model in its entirety (although the overall accuracy of the compositional model is superior to those of the constituent parts) for values of confidence (p) up to 0.7; for greater values of confidence, this behavior is no longer observed. This example thus provides support for our claim that evaluation of the performance of a geostatistical model for compositional data should not be based on the performance of the model on the individual components only. Figure 2 also shows that the model ignoring anisotropy appears to have higher accuracy, but at the price of constructing much wider p-intervals. In the accuracy plot, this is shown through the much greater deviation from the bisector. Accounting for anisotropy notably increases the model precision, as evidenced by the higher value of $P = 0.79$ and $G = 0.88$ in the anisotropic case compared to $P = 0.47$ and $G = 0.74$ for the isotropic model.

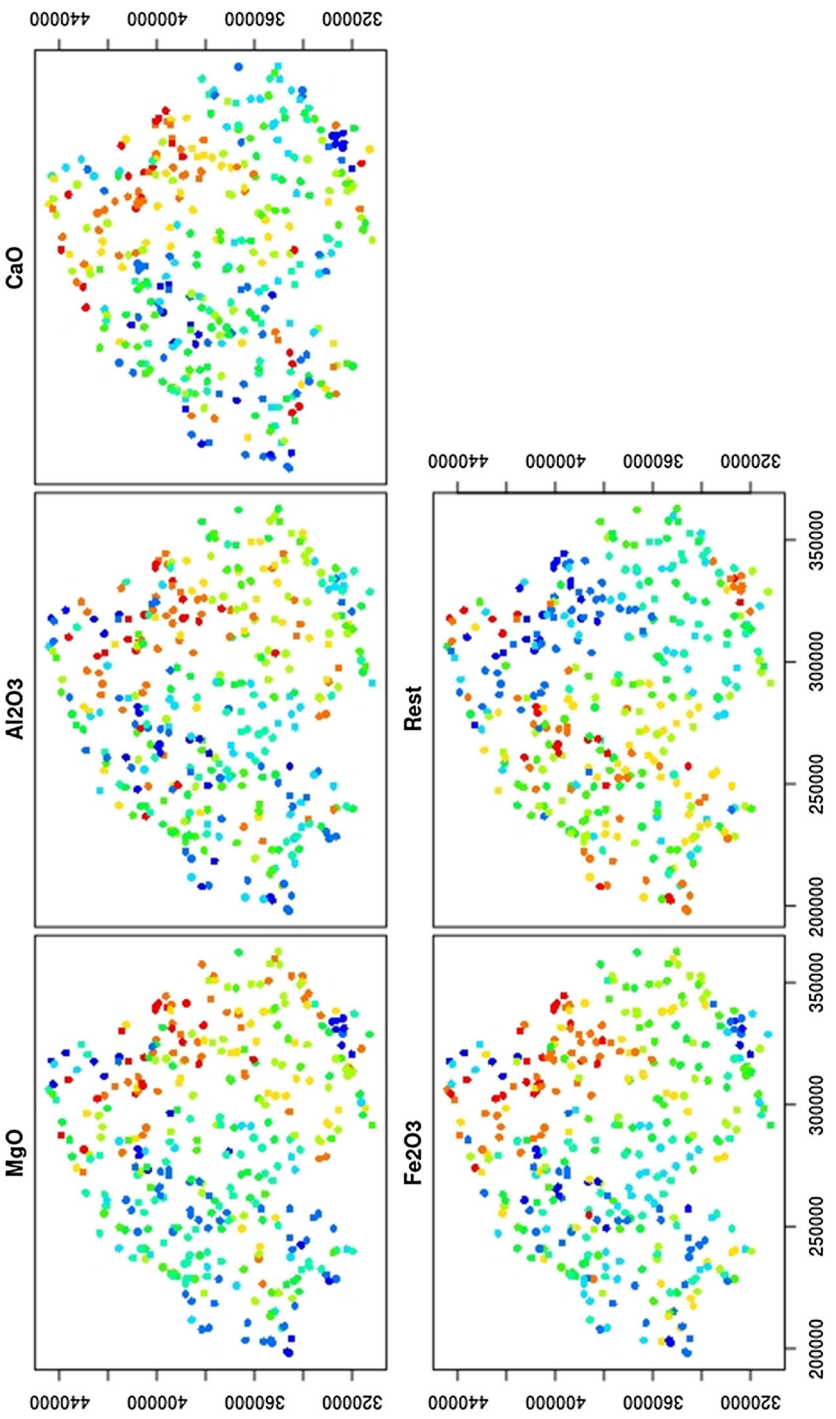

Fig. 1 Sample data for the Tellus subcomposition case study; blue indicates low values and red high

Table 1 Values of A, P and G from ordinary cokriging of the Tellus sample subcomposition obtained with the anisotropic model versus isotropic model

	ilr1	ilr2	ilr3	ilr4	Composition (anisotropic)	Composition (isotropic)
A	0.49	0.73	0.76	0.65	0.78	0.98
P	0.97	0.96	0.91	0.94	0.79	0.47
G	0.94	0.96	0.92	0.92	0.88	0.74

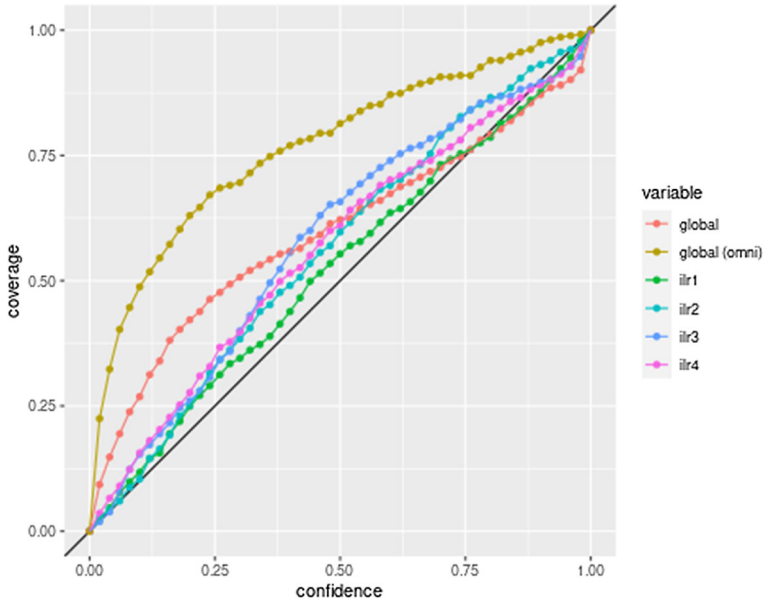

Fig. 2 Accuracy plot derived from tenfold ordinary cokriging cross-validation of the Tellus subcomposition sample, with an isotropic model (omni) and with an anistropic one (globally, and for each one of the ilr variables based on the anisotropic LMC)

Table 2 gives a summary of the compositional error measures for both the isotropic and anisotropic models. The vector of ilr mean errors (**ME**) is in both cases very close to zero, with mean square errors (MSE) of roughly 0.4 in both cases. The difference between the two models is apparent in the MSDR measures, where it is clear that the isotropic model overestimates the spread. MSDR_2 clearly shows that on average over all logratios, the anisotropic model meets the target of MSDR equal to 1.

This subcomposition was also studied using direct simulation (Mariethoz et al. 2010), with a validation subset of 100 samples, and the rest used as conditioning data and training image generation, reported in Tolosana-Delgado and Mueller (2021; Ch. 10–11). To complement the discussion in these chapters, the accuracy metrics in the original units (raw data) are shown in Table 2. The results show overly optimistic metrics in general terms, indicating a potential lack of variability in the training image,

Table 2 Compositional error measures obtained by cross-validation with isotropic and anisotropic models

	$MSDR_1$	$MSDR_2$	MSE	
Isotropic	2.03	0.52	0.39	
Anisotropic	4.09	1.00	0.40	
ME	ilr1	ilr2	ilr3	ilr4
Isotropic	−0.001	0.001	−0.001	−0.005
Anisotropic	−0.000	0.002	−0.001	−0.001

Table 3 Values of A, P and G from direct sampling of the subcomposition of interest for the validation sample in the Tellus data set

	MgO	Al_2O_3	CaO	Fe_2O_3	Rest	Compositional
A	0.95	0.55	0.45	0.20	0.85	0.95
P	0.83	0.97	0.97	0.99	0.84	0.53
G	0.91	0.95	0.95	0.94	0.92	0.77

Table 4 Parameter values for the stratigraphic forward simulation (left) and direct sampling (DS) simulation (right)

Delft3D parameter	Value	DS parameter	Value
Domain (m^3)	$25 \times 25 \times 6$	Grid size (voxel)	$18 \times 21 \times 21$
Pixel size (m^3)	$0.5 \times 0.5 \times 0.1$	Search area (voxel)	$10 \times 10 \times 3$
Silicate density	$2.7 kg/m^3$	TI fraction to scan	0.25
Sulfide density	$4.00 kg/m^3$	Data event size	15 points
Inflow	$0.5, 0.4, 0.3 m^3/s$	Distance threshold	0.01
Silicates	$0.5 kg/m^3$	No. of simulations	100
Sulfides	$0.02 kg/m^3$		

with generally high goodness values for the individual variables in contrast to a large range of accuracy values (from a low of 0.2 for Fe_2O_3 to a maximum of 0.95 for MgO). The corresponding compositional metrics have high accuracy but low precision and moderate goodness (Table 3).

5.2 The Tailings Storage Facility Case Study

The second example illustrates the usage of the simulation-based approach of Sect. 4.2, by means of a case study with direct sampling (Mariethoz et al. 2010) with a very complex setup, far from additive logistic normality. Here the aim is to model distributions of particle types in a multiple stream tailings storage facility (Selia et al.

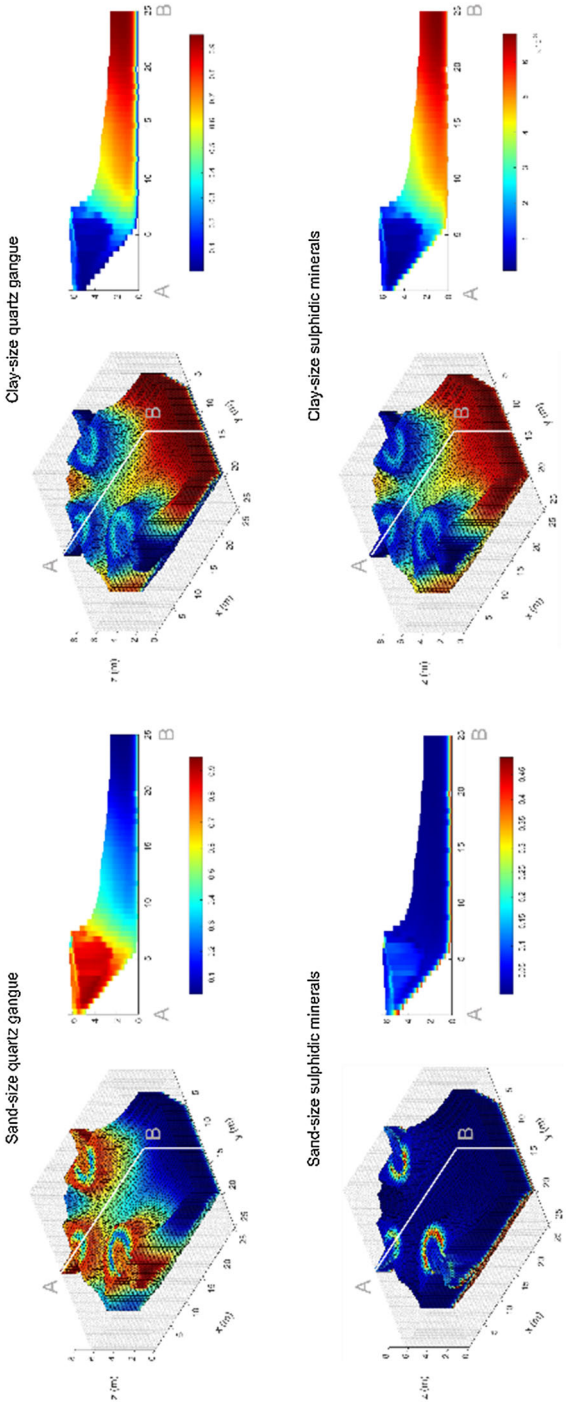

Fig. 3 Forward simulated four-component system as training image for the tailings storage case study

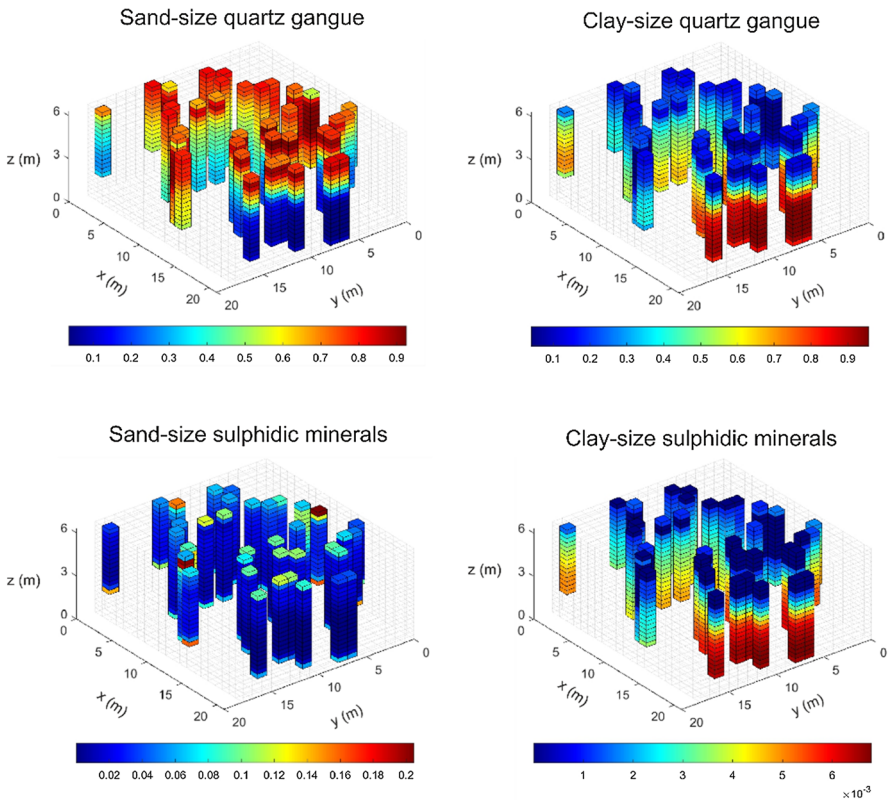

Fig. 4 Synthetic sample data for the tailing storage case study

Table 5 Values of A and G from simulation-based leave-one-out cross-validation of the direct sampling example (logratio coefficients; global: Aitchison–Mahalanobis distance summary)

	ilr1	ilr2	ilr3	Global
A	0.00	0.05	0.00	0.00
G	0.84	0.88	0.81	0.89

in prep). Stratigraphic forward modeling is used for training image generation and direct sampling for data fusion with the measured data. A multipoint method was considered as most appropriate because of the strong effects of non-linearity and non-stationarity of the patterns typical of these anthropogenic sedimentary systems. Briefly, the study considers four particle classes, according to size and dominant mineralogy, forming a four-part mass composition: sand-sized silicates (V1), clay-sized silicates (V2), sand-sized sulfides (V3) and clay-sized sulfides (V4). The forward model used is Delft3D-FLOW (Lesser et al. 2004), an open-source, process-based stratigraphic forward modeling software accounting for diffusion, advection in both bed load and

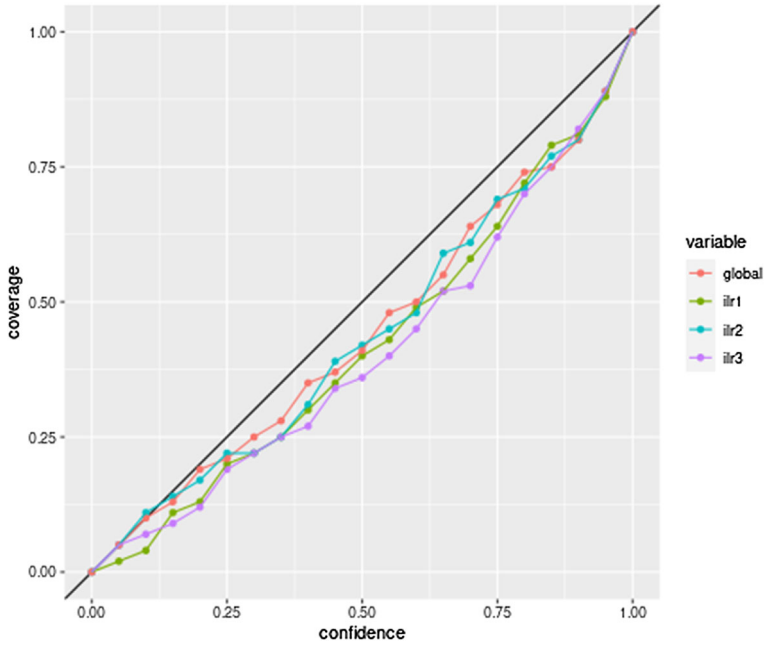

Fig. 5 Accuracy plot derived from simulation-based leave-one-out cross-validation of the direct sampling example (logratio coefficients; global: Aitchison–Mahalanobis distance summary)

Table 6 Values of A and G from simulation-based leave-one-out cross-validation of the direct sampling example (raw data)

	V1	V2	V3	V4
A	0.00	0.00	0.00	0.00
G	0.87	0.84	0.87	0.83

suspended load, erosion and compaction in both aerial and subaquatic environments. The parameters of the forward simulation are described in Table 4 (left). Boundary conditions are designed to mimic the behavior of tailings dams, allowing water to seep while retaining sediments. Results were cropped to the basin without the upstream channels and upscaled to $18 \times 21 \times 21$ voxels to form the ground truth, which for this study, is also taken to be the training image (Fig. 3). Synthetic boreholes were randomly taken at 36 locations (Fig. 4). Of the resulting 677 samples, 100 were randomly picked for leave-one-out cross-validation.

To predict each of these 100 samples, direct sampling was applied excluding that sample, based on the parameters specified in Table 4 (right), using the Aitchison distance (Eq. 3) as the measure of proximity between the composition at each training image pixel and the data set. The resulting simulations were used to calculate the accuracy and goodness as defined in Sect. 4.2.

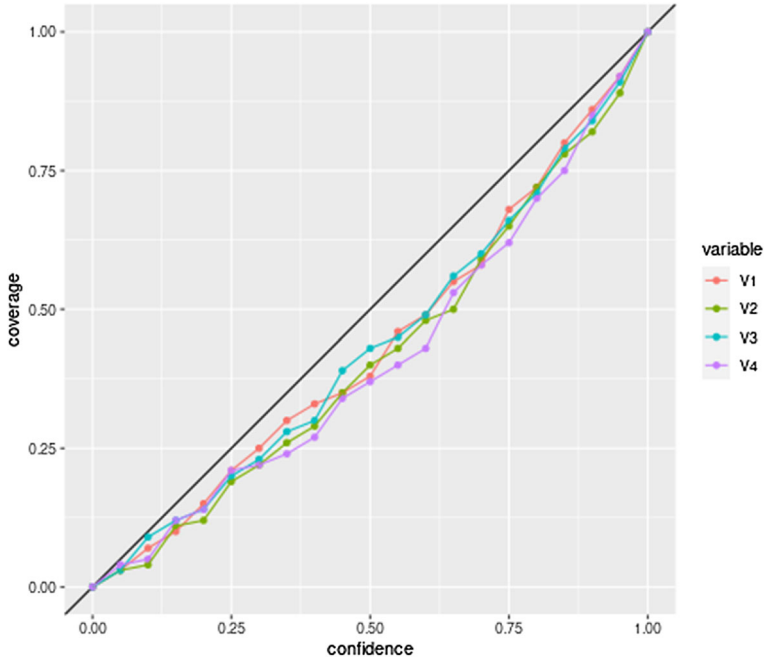

Fig. 6 Accuracy plot derived from simulation-based leave-one-out cross-validation of the direct sampling example (raw data)

As can be seen from the numerical (Table 5) and graphical results (Fig. 5), the simulations show inaccurate but moderately good results: the accuracy curve for both individual ilr coefficients and for the global Aitchison–Mahalanobis summary are systematically below but close to the reference line, particularly for theoretical coverage levels below 0.20. This indicates that confidence intervals tend to be too small to deliver the coverage promised by their nominal confidence. Correspondingly, the numerical values of accuracy A are close to zero, and precision is meaningless (hence not reported in the tables). In this situation, the goodness G becomes useful: G is above 0.81 for all three ilr variables, and the global goodness is the highest (0.89).

The availability of a simulation allows an evaluation of the accuracy in terms of the original four components. Results are reported in Table 6 and Fig. 6, and do not qualitatively diverge from the logratio-based results: confidence intervals are too large, so that accuracy values A are not useful. However, goodness values are high (above 0.83), suggesting that the model is acceptable in both representations, raw and logratio.

6 Conclusions

We have provided an extension of quantitative measures of goodness of fit to the multivariate context, particularly for compositional data, which allow ranking of models analogously to already existing univariate approaches. In addition to generalizations

of mean errors to vectorial quantities, the average Mahalanobis distance between estimates and observations ($MSDR_1$) gives an additional decision tool for choosing between competing models. This measure explicitly accounts for the multivariate structure, is affine-equivariant and shifts the focus from the individual components to the entire composition. Alternatively, an average mean square deviation ratio over all logratios ($MSDR_2$) can also be used, although this quantity is not affine-equivariant, that is, it depends on which specific logratios are being used to compute the statistic. In principle, both measures focus on different aspects and could result in different rankings of models, although both would result in the choice of the same model in the case studies provided here. The joint accuracy and precision measures introduced here can provide further insights into the system inasmuch as they evaluate the global structure, and may rank competing models differently as when evaluated in terms of just the marginals.

The Mahalanobis distance also provides a reasonable one-dimensional summary of the multivariate distribution, to derive a cumulative distribution and with them measures of accuracy, precision and goodness after Deutsch (1997). Accuracy cannot be evaluated isolated from goodness or precision measures, in either the univariate or multivariate case. This is particularly evident in the metrics for the isotropic versus anisotropic linear models of coregionalization of the Tellus subcomposition, where the accuracy is marginally higher for the isotropic model, but goodness and precision provide a higher contrast to choose the better model. In particular, the goodness metric is important here, as it accounts for both accurate and inaccurate cases. In the case of the tailings storage facility case study, the model overall ended up showing results being too far from the center and hence resulting in an accuracy value of 0. Nevertheless, examination of the actual versus theoretical coverage shows reasonable goodness, with the plot for the global metric quite close to the bisector. The tools could thus still be used for ranking competing models, such as training images or method parameter setups, had several of them been available.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall Ltd, London

Barnett RM, Manchuk JG, Deutsch CV (2014) Projection pursuit multivariate transform. Math Geosci 46(2):337–360

Chilès JP, Delfiner P (2012) Geostatistics – Modelling spatial uncertainty, 2nd edn. Wiley, New York

Deutsch CV (1997) Direct assessment of local accuracy and precision. In: Baafi E, Schofield NA (eds) Geostatistics Wollongong '96, vol 1. Kluwer, Dordrecht, pp 102–113
Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geosci 35(3):279–300
Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Math Geosci 40(3):233–248
Lesser G, Roelvink J, van Kester J, Stelling G (2004) Development and validation of a three-dimensional morphological model. Coast Eng 51(8–9):883–915
Mariethoz G, Renard P, Straubhaar J (2010) The direct sampling method to perform multiple-point geostatistical simulations. Water Resour Res 46(11):7621
Olea RA (2012) Building on cross-validation for increasing the quality of geostatistical modeling. Stoch Env Res Risk Assess 26:73–82
Pawlowsky-Glahn V, Egozcue JJ (2020) Compositional data in geostatistics: a log-ratio based framework to analyze regionalized compositions. Math Geosci 52:1067–1084
Pawlowsky-Glahn V, Olea RA (2004) Geostatistical analysis of compositional data. Oxford University Press, Oxford
Selia S, Tolosana-Delgado R, van den Boogaart KG (in prep) Compositional multi-point geostatistics for tailings deposits—a synthetic case study. Comput Geosci
Sepulveda E, Adeli A, Dowd PA, Ortiz JM, Abulkhair S, Xu C (under review) Evaluation of multivariate Gaussian transforms for geostatistical applications. Submitted to Stochastic Environmental Research and Risk Assessment. preprint https://doi.org/10.21203/rs.3.rs-2087808/v1
Tolosana-Delgado R, Mueller U (2021) Geostatistics for compositional data with R, UseR! series. Springer, Cham
Tolosana-Delgado R, Mueller U, van den Boogaart KG (2019) Geostatistics for compositional data: an overview. Math Geosci 51:485–526
van den Boogaart KG, Schaeben H (2002a) Kriging of regionalized directions, axes and orientations (I): directions and axes. Math Geosci 34(5):479–503
van den Boogaart KG, Schaeben H (2002b) Kriging of regionalized directions, axes and orientations (II): orientations. Math Geosci 34(6):671–677
van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing compositional data with R. Springer, Heidelberg
van den Boogaart KG, Mueller U, Tolosana-Delgado R (2017) An affine equivariant multivariate normal score transform for compositional data. Math Geosci 49(2):231–251
Webster R, Oliver M (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, Chichester
Young ME, Donald AE (2013) A Guide to the Tellus Data. Geological Survey of Northern Ireland, Belfast