HHS Public Access

Author manuscript

J Geochem Explor. Author manuscript; available in PMC 2019 March 01.

Published in final edited form as:

J Geochem Explor. 2018 March ; 186: 24–35. doi:10.1016/j.gexplo.2017.11.022.

Mapping of compositional properties of coal using isometric log-ratio transformation and sequential Gaussian simulation – A comparative study for spatial ultimate analyses data

C. Özgen Karacan[a,b,*] and **Ricardo A. Olea**[a]

[a]USGS, Reston, VA, USA

[2]NIOSH, PMRD, Pittsburgh, PA, USA

Abstract

Chemical properties of coal largely determine coal handling, processing, beneficiation methods, and design of coal-fired power plants. Furthermore, these properties impact coal strength, coal blending during mining, as well as coal's gas content, which is important for mining safety. In order for these processes and quantitative predictions to be successful, safer, and economically feasible, it is important to determine and map chemical properties of coals accurately in order to infer these properties prior to mining.

Ultimate analysis quantifies principal chemical elements in coal. These elements are C, H, N, S, O, and, depending on the basis, ash, and/or moisture. The basis for the data is determined by the condition of the sample at the time of analysis, with an “as-received” basis being the closest to sampling conditions and thus to the in-situ conditions of the coal. The parts determined or calculated as the result of ultimate analyses are compositions, reported in weight percent, and pose the challenges of statistical analyses of compositional data. The treatment of parts using proper compositional methods may be even more important in mapping them, as most mapping methods carry uncertainty due to partial sampling as well.

In this work, we map the ultimate analyses parts of the Springfield coal from an Indiana section of the Illinois basin, USA, using sequential Gaussian simulation of isometric log-ratio transformed compositions. We compare the results with those of direct simulations of compositional parts. We also compare the implications of these approaches in calculating other properties using correlations to identify the differences and consequences. Although the study here is for coal, the methods described in the paper are applicable to any situation involving compositional data and its mapping.

*Corresponding author at: USGS Eastern Energy Resources Science Center, Reston, VA, USA. ckaracan@usgs.gov (C.Ö. Karacan).

Disclaimer

For the National Institute for Occupational Safety and Health (NIOSH), the findings and conclusions in any paper are those of the authors and do not necessarily represent the views of NIOSH. Mention of any company name, product, or software does not constitute endorsement by NIOSH. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Keywords

Calorific value; Coal quality; Compositional modeling; Regression; Springfield coal

1. Introduction

Aside from being characterized as containing only organic and inorganic compounds, coal is a chemically, petrographically, and physically complex and heterogeneous natural material, which is not easy to fully characterize for all of its properties. Its organic part contains different macerals as the building blocks, whereas the inorganic part contains different clays, minerals, and various major and trace elements. These compositional properties of coal and how they may interact with air, for instance, can significantly influence how it should be mined, handled, processed, utilized, and even how coal-fired power plants should be designed to reduce emissions. Moreover, these properties affect coal strength, coal blending design, as well as coal's gas content, which is important for mining safety. In other words, coal's properties and composition impact all processes, from its safe mining to its utilization in different industries.

Two of the basic and most common analyses to describe properties of coal are proximate and ultimate analyses. Proximate analysis determines moisture, volatile matter, ash, and fixed carbon within the coal (ASTM D121, 2006). These properties are important for coal utilization, as moisture and ash affect heat absorption and thus calorific value of coal. Ultimate analysis, on the other hand, is more detailed and involves the determination of carbon, hydrogen, nitrogen, total sulfur, oxygen, and ash yield. Carbon and hydrogen contents are determined by analyzing the gaseous products of the complete combustion of the coal, oxygen by difference, and total sulfur, nitrogen, and ash yield are based on the coal material as a whole. After corrections for carbon, hydrogen, and sulfur derived from the inorganic material, and for ash to mineral matter, the ultimate analysis represents composition of the organic material in coal in terms of carbon, hydrogen, nitrogen, sulfur, and oxygen (Riley, 2007).

The elemental composition of coal determined by ultimate analysis can be important for various purposes. For instance, it gives an idea about maturity of coal. Carbon, oxygen, and hydrogen are, to some degree, rank-dependent elements. The highest rank coals have the highest carbon contents and the lowest oxygen contents. High-volatile B and C bituminous coals, on the other hand, have the highest hydrogen contents, with decreasing amounts as rank increases. Therefore, ratios of these elements can indicate the rank of the coal and its coalification degree.

The results of ultimate analysis can also be used in different correlations to predict various properties of coal. Chelgani et al. (2008) investigated the effects of proximate and ultimate analysis, maceral content, and coal rank for a wide range of Kentucky coal samples on Hardgrove grindability index (HGI) using multivariable regression and artificial neural network (ANN) methods. One of the most comprehensive papers that cover correlations, such as density, the Hardgrove grindability index, the free swelling index, pyrolysis yields, direct liquefaction yields, and carbon dioxide yields, by using ultimate, proximate, and

maceral analyses results, is published by Mathews et al. (2014). They reviewed 42 correlations found in the literature addressing multiple coal properties and compared against vitrinite-rich United States coals sampled from the Pennsylvania State University Coal Sample Bank and Database.

Mathews et al. (2014) concluded that while some correlations, such as calorific value predictions, are accurate over a wide range of coals, others are restricted in applicability to a select rank range. They interpret the limitation in applicability as a consequence of the creation intent of the correlations or to the complex nature of the coal. While these are true and valid arguments, all statistical analyses are done using compositional data (i.e. observations carrying relative information) such as ultimate, proximate, and maceral analyses. The outputs are used to correlate them to non-compositional data and, in some cases, to other compositional data. Therefore, the way that the data is handled and analyzed using classical statistics may make a difference as well in the predictive performance of correlations built on compositional data. Since ultimate analysis gives six partial components, as in the case of an “as received” basis to the whole—carbon (C), hydrogen (H), nitrogen (N), sulfur (S), oxygen (O) and ash—the results can be evaluated mathematically in the realm of compositional data analyses.

The ultimate analysis is a *D-part composition* of the whole, and its sampling space is the *D-part simplex* with a set of real vectors with positive components, which can be represented by a constant-sum constraint without loss of information (Otero et al., 2005). This constant value is 100% in the case of ultimate analysis on an “as received” basis. The compositional character of the vector, and thus its relative nature, imposes a geometry on the system, which is different than Euclidean geometry in real space; therefore a new mathematical system to work with compositional data has been developed (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001; Egozcue et al., 2003). In such a system, it is known that classical statistics based on the Euclidean distance are unable to honor the compositional properties of the data and may lead to spurious correlations. Such cases where improper treatment of compositional data lead to misleading or spurious results, and the causes are given in Aitchison (1986), Chayes (1960), Buccianti and Pawlowsky-Glahn (2005), Engle and Rowan (2014), Filzmoser et al. (2009), and Otero et al. (2005).

Compositional data analysis techniques have been applied in different problems related to coal and to different areas related to earth and environmental sciences (e.g. geochemistry, in particular). Geboy et al. (2013) used major oxide and trace metal data from the Pennsylvanian-age Pond Creek coal of eastern Kentucky, USA, and applied isometric log-ratio (ilr) to perform univariate statistics and analyze the importance of reporting basis. In classical statistics, it is preferable to work in orthonormal coordinates and in case of compositional data, they are obtained using the ilr transformation. The only peculiarity is the out of D-part compositional data, only D-1 ilr coordinates are obtained. As a consequence, a careful interpretation of the new variables is needed. Nevertheless, they concluded that the results of univariate statistics still differed between the ash basis and whole-coal basis, but in predictable and calculated manners. Further, they emphasized that the stability between two different components, a bivariate measure, is identical, regardless of the reporting basis. Otero et al. (2005) applied compositional principal components analyses and bi-plot

analyses to distinguish between different anthropogenic and geological pollution sources at 30 sampling stations along the Llobregat River and its tributaries in Barcelona, Spain. They showed that the differences in water quality and the pollution sources could be highlighted when compositional statistics was used. Reimann et al. (2012) applied compositional statistics and centered log-ratio (clr) to plot soil composition in Europe. They showed that, through application of compositional analyses, the regional distribution of some elements did not change substantially, while the patterns of some others changed considerably. In one of the most recent studies, Owen et al. (2016) explored the variability in major ion composition between gas bearing aquifers to delineate ground water associated with biogenic gas using compositional data analysis techniques. By using isometric log ratios, they described subtle differences in the behavior of Na, K, and Cl between coal seam gas water types and very similar Na-HCO_3 water types in adjacent aquifers.

As with any non-compositional data, the knowledge of the spatial distribution of compositional data is also often desirable as these properties of any commodity, e.g. coal, may have safety and health, as well as economic and environmental implications in exploiting and utilizing them. Geostatistics is a powerful technique for investigation of spatial relationships of regionalized variables and for modeling them (e.g. Deutsch and Journel, 1998; Pyrcz and Deutsch, 2014; Caers, 2011; Chilès and Delfiner, 2012). Common geostatistical modeling methods can still be used for compositional data too (Pawlowsky-Glahn and Olea, 2004; Pawlowsky-Glahn and Egozcue, 2016), as demonstrated by various applications (e.g. Tolosana-Delgado and van den Boogaart, 2014; Park and Jang, 2014). However, unlike with non-compositional data, spatial modeling of compositional data may require using compositional data analyses techniques combined with geostatistics to avoid spurious correlations and to solve problems related to the compositional character of spatially dependent data.

In this paper, we map the ultimate analyses parts of the Springfield coal from an Indiana section of the Illinois basin, USA. If determined on an “as-received” basis, weight percentages of the ultimate analyses parts, C, H, N, S, O, and ash-yield, sum to 100%—i.e. they are constrained data in a simplex, not data in real space that can take values from $-\infty$ to $+\infty$, and they also carry relative information, which is essential information of interest. Therefore, they need a special pre-processing for a mathematically adequate mapping using geostatistical methods. In this work, we use isometric log-ratio transformed compositions in sequential Gaussian simulation. We compare the results from this approach with those of direct simulations of compositional parts using sequential Gaussian simulation. We also compare the implications of these approaches in modeling heat value data for this coal to identify the differences and potential error incurred. In this study, the approach is demonstrated for coal. However, the methods described in the paper are applicable to any situation involving compositional data and its mapping.

2. Studied coal seam and its ultimate analysis data within the area of interest

The studied area of the Springfield coal seam is located in the Indiana section of the Illinois Basin, USA. The Springfield coal seam is in the Petersburg Formation of the Carbondale Group and is one of the most important coal seams of economic value along with the Danville, Hymera, Herrin, and Seelyville coals (Fig. 1). Springfield coal has high-volatile bituminous C and B rank and occurs generally at depths between 300 and 600 ft. The coal has average vitrinite reflectance (Ro) of 0.63%, and has 73.4% vitrinite, 6.4% liptinite, 15.4% inertinite, and 4.8% mineral matter.

The ultimate analysis data used in this work was compiled from the Indiana Geological Survey Coal Stratigraphic and Coal Quality Databases (Drobniak and Mastalerz, 2012a, 2012b). The studied area covers a range between UTM coordinates of 470,494 ft and 483,705 ft in the east-west direction, and UTM coordinates of 4,198,870 ft and 4,256,544 ft in the north-south direction. Figs. 2 and 3 show posting of the spatial ultimate analysis data of Springfield coal on an “as-received” basis in this area and the histograms of these data, respectively. The sample size per part shown in these figures is 54. The samples were collected from surface and underground mines of Springfield coal, as well as from its outcrops in different counties in Indiana.

The data given in Figs. 2 and 3 shows that the values of the individual parts are distributed within the area without a noticeable trend, and their frequency histograms are not following normal distributions. The basic results of the statistics conducted on these data are given in Table 1. However, it should be noted that the results given in this table have just explorative value, as the statistics is conducted using classical tools.

In most situations, for data that is concerned with Euclidian distance, classical statistics is applicable. However, a composition is a vector that carries information about the relative importance of the measured parts in the whole. Since the parts in a composition are relative to each other in the whole, this gives a composition an intrinsic multivariate property. This multivariate property can be attained by means of a comparison between parts as pairs for their importance in the whole (Otero et al., 2005). Due to the multivariate nature of compositions, simple cross-plots of compositional data, or Harker diagrams, are not informative in general (Filzmoser and Hron, 2009; Buccianti et al., 2014).

The compositional bi-plot that is based on the centered log-ratio (clr) transformation is one of the most popular ways to jointly represent the variables due to its connection to principal component analysis (PCA) (Aitchison, 1983). The bi-plot is a 2-D representation of the singular value decomposition (SVD), where the individual data points are represented as dots and the variables are represented as rays (Otero et al., 2005). The length of a ray is proportional to the standard deviation of the variable it represents. If two rays are near each other, the variables might be highly associated. If three or more vertices of rays are aligned, a compositional linear relationship between the parts is possible and should be checked. The links between the vertices of rays, on the other hand, are proportional to the standard deviation of the simple log-ratio of the variables corresponding to the rays. The cosine of the

angle between the two links closely approximates the correlation coefficient between the corresponding simple log-ratios (Egozcue and Pawlowsky-Glahn, 2011). If this angle is orthogonal, the two simple log-ratios are possibly uncorrelated.

Fig. 4 shows the compositional bi-plot of the ultimate analysis data given in Figs. 2 and 3. The plot was generated using Compositional Data Package (CoDaPack) (Comas and Thió-Henestrosa, 2011). The principal components shown in this plot represent 89% of the cumulative variance. The dots shown in the plot, which are partitioned based on the sampling location, are the individual data, whereas the rays are clr of the parts. The features of this plot show that clr(S) ray extending in the opposite direction with the highest standard deviation. Due to its direction compared to C, H, N, and O, which seem to be weakly associated with each other, its changes may be compensated by shared changes in these parts. Also, Fig. 4 shows that S is not associated, or proportional, with ash. The distribution of data, on the other hand, shows that all sampling locations have higher affinity to all parts, except ash. A higher proportion of ash seems to have higher affinity to surface mines. Also, outcrop samples seem to have higher affinity towards a higher proportion of O compared to other parts.

3. Methodology for spatial modeling of ultimate analysis data

Geostatistics applies to attributes defined in real number space with the assumption of Euclidian distance (Olea and Luppens, 2016). Compositional data, however, is constrained and represents relative information about parts of the whole represented usually in the closed space (e.g. 100%) called simplex. While direct application of geostatistical techniques will still produce maps for each of the modeled parts, the sum of the values of individual cells in these maps may exceed what the sum should be, which is clearly a violation to mathematics and to the core idea of the experimental analysis.

In this work, our objective was to map ultimate properties of Springfield coal shown in Fig. 2 by using two approaches and to compare the results for their adequacy in spatial modeling of compositional data. The first and the commonly applied approach, despite its aforementioned deficiencies, was direct geostatistical modeling of ultimate properties using sequential Gaussian simulation. The second approach was compositional geostatistical modeling of ultimate properties using ilr transformation and then sequential Gaussian simulation. Fig. 5 shows these approaches and the workflow for each of them. SGeMS (Remy et al., 2009) and CoDaPack (Comas and Thió-Henestrosa, 2011) for sequential Gaussian simulation and for ilr transformation were used, respectively.

3.1. Direct geostatistical modeling of ultimate properties

The details of sequential Gaussian simulation (e.g. Deutsch and Journel, 1998; Remy et al., 2009) and its applications for different problems were demonstrated in the literature (e.g. Srivastava, 2013)—readers are also referred to volume 112 of *International Journal of Coal Geology* for its different applications. Therefore, the exhaustive list of different publications and applications will not be repeated here. However, it should be mentioned that sequential Gaussian simulation was selected in this work because of its practicality, its efficiency in honoring the data and its histogram, and also due to its potential to enable uncertainty

analysis through as many equiprobable realizations as necessary that can represent a stabilized variance for the attribute of interest. In this work, 100 realizations—in both approaches shown in Fig. 5—were generated for each of the ultimate analysis parts.

In addition, central to its application, sequential Gaussian simulation requires that the data follow a univariate normal distribution. Since none of the parts met this requirement, as seen in Fig. 3, all parts were transformed to normal-score space before variogram analysis (Fig. 5). An example of this process for C% is shown in Fig. 6, and analytical variograms modeled for the normal-score data of all ultimate analysis parts are given in Table 2. Sequential Gaussian simulation was performed using these variograms to generate 100 realizations for each of the properties. The other parameters of the simulation algorithm, including the seed number and maximum conditioning data in the search ellipsoid, were kept constant between the two modeling approaches followed in this paper.

3.2. Compositional geostatistical modeling of ultimate properties through ilr-transformation

The workflow adopted for the compositional geostatistical analysis and mapping of the data is shown in Fig. 5. A similar approach was used before by Olea and Luppens (2016) and Olea et al. (2016) for proximate analysis of a Texas lignite deposit.

In this work, ilr transformation was used to produce a new set of variables in an unconstrained space with an orthogonal coordinate system, where standard geostatistical methods can be applied without violating the mathematics. The ilr transformation was selected because it avoids singular covariance matrices and sub-compositional incoherence (Egozcue et al., 2003; Filzmoser and Hron, 2009).

The first step of ilr transformation is to generate a class of interpretable ilr variables, called balances. As mentioned earlier, a D-part composition whose values sum up to a constant is in a simplex of D-1 dimension due to the relativeness of measurements and thus the multivariate nature of the data. Therefore, for a 6-part system, 5 ilr balances were generated through binary partition of parts of the composition by using a binary partition matrix, Θ (Egozcue and Pawlowsky-Glahn, 2005, 2006). The binary partition starts dividing parts into groups until all parts are in an individual group. The partition is conducted by +/− 1, indicating inclusion, or not, of a part in a particular group or by “0,” indicating a no action for that part at a particular partition level. It should be mentioned that the final results of the spatial modeling do not depend on the way that the binary partition is conducted, although it can be important for easy interpretation of the ilr coordinates. Eq. (1) shows the binary partition matrix used in this work and the values p_j and n_j, for +1 and −1 entries in each row, respectively. For each location $\mathbf{u_i}$, the D-1 transformations $ilr_j\ z(\mathbf{u_i})$ of $z_D(\mathbf{u_i})$ – of D-part compositions at each location - are given by Eq. (2). An example calculation for ilr_1 illustrating the use of binary partition matrix for this transformation is given in Eq. (3).

$$\theta = \begin{array}{c} \begin{matrix} \text{Ash} & \text{C} & \text{H} & \text{N} & \text{O} & \text{S} \end{matrix} \\ \begin{bmatrix} -1 & 1 & 1 & 1 & 1 & 1 \\ 0 & -1 & 1 & 1 & 1 & 1 \\ 0 & 0 & -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \end{array} \begin{array}{cc} \text{p}_\text{j} & \text{n}_\text{j} \\ 5 & 1 \\ 4 & 1 \\ 3 & 1 \\ 2 & 1 \\ 1 & 1 \end{array} \quad (1)$$

$$\text{ilr}_j(z(\boldsymbol{u}_i)) = \sqrt{\frac{p_j \times n_j}{p_j + n_j}}\ \ln \frac{[\prod_{p_j} \boldsymbol{z}_{p_{j-D}}(\boldsymbol{u}_i)]^{1/p_j}}{[\prod_{n_j} \boldsymbol{z}_{n_{j-D}}(\boldsymbol{u}_i)]^{1/p_j}}, i = 1, 2, \ldots, \text{N}; j = 1, 2, \ldots\ldots \text{D} - 1 \quad (2)$$

where N is the number of data locations with compositions. For example

$$\text{ilr}_1(u_1) = \sqrt{\frac{5 \times 1}{5+1}}\ \ln \frac{[C \times H \times N \times O \times S]^{1/5}}{(\text{Ash})} \quad (3)$$

Performing this transformation for all 54 spatial data locations with 6 parts gave 5 ilr balances, each of which had 54 spatial data points that could be used for analysis and modeling.

The next step was to do spatial modeling of ilr balances. This required two considerations; the first one was to check the normality of ilr balances, as before, to be able to proceed with variogram analysis and then with sequential Gaussian simulation. The histograms of ilr balances were not normally distributed, and thus variogram analyses were performed after normal-score transformation (Fig. 7). Table 3 gives the parameters of analytical variograms of normal-score transformed ilr balances.

The second consideration was to understand if there was spatial cross-correlation between ilr balances that would require sequential Gaussian co-simulation, instead of sequential Gaussian simulation. In order to explore cross-correlations, structural analysis was performed by generating cross-variograms between all normal-score transformed ilr balances (Fig. 8). This figure shows that spatial cross-correlations were absent and co-simulations were not needed, which in this case would not improve the results either as all data were collocated (Olea, 1999). Nevertheless, since compositions are based on relative measures in a D-1 dimensional simplex, and thus the data is multivariate in nature, conducting a structural analysis was needed to prove that spatial cross-correlations were absent. After these analyses and confirmations, analytical variograms given in Table 3 were used in sequential Gaussian simulation to model all 5 ilr balances (100 realizations for each) by using the rest of the algorithm parameters—the same as in the direct simulation of

compositional parts. Fig. 9 shows ilr-balance maps of the 25th realization (arbitrarily selected) generated using sequential Gaussian simulation.

The next and final step was to transform ilr balance realizations back to the simplex (percentages). The back transformation was performed using Eq. (4). In this equation, $\boldsymbol{z}_D^*(\boldsymbol{u}_i)$ are the back-transformed values (D of them) corresponding to the D-1 **ilr*(z(u$_i$)** transformed values in realizations generated by sequential Gaussian simulation, and c is the ratio of closure constant over the sum of the parts in the vector product $\boldsymbol{\Psi}^{\mathrm{T}} \cdot \mathbf{ilr}(\boldsymbol{u}_i)$. In this product, $\boldsymbol{\Psi}^{\mathrm{T}}$ is the transpose of the contrast matrix, whose elements are determined by the elements of partition matrix, $\theta_{j,k}$, and is determined by Eqs. (5) and (6).

$$\boldsymbol{z}_D^*(\boldsymbol{u}_i) = c^*(\boldsymbol{u}_i) \cdot \exp\left(\boldsymbol{\Psi}^T \cdot \boldsymbol{ilr}^*\right), i = 1, 2, \ldots, \mathrm{N} \quad (4)$$

$$c^*(\mathbf{u}_i) = \frac{\kappa}{\sum_{j=1}^{D} \exp([\boldsymbol{\Psi}^{\mathrm{T}} \cdot \mathbf{ilr}^*(\mathbf{u}_i)]_j)} \quad (5)$$

$$\boldsymbol{\Psi}_{i,k} = \begin{cases} 0, & \text{if } \theta_{i,k} \text{ is } 0 \\ \sqrt{\frac{n_j}{p_j \cdot (p_j + n_j)}}, & \text{if } \theta_{i,k} \text{ is } +1 \\ -\sqrt{\frac{p_j}{n_j \cdot (p_j + n_j)}}, & \text{if } \theta_{i,k} \text{ is } -1 \end{cases} \quad (6)$$

where the bracket $[\cdot]_j$ denotes the jth component of the vector and κ is the constant sum, 100 in our case. Applying back-transformation to each cell in all 100 realizations from ilr space to real space generated maps of composition values for all 6 parts of the ultimate analysis through a compositional approach. For this purpose, 500 ilr realizations (100 for each) were back-transformed as batches of 5 of each realization to generate ultimate analysis parts corresponding to the same realization numbers. Olea and Luppens (2016) give a numerical example calculation for transformation and inverse transformation, described there for a 4-part system for those who are interested in following the calculations.

4. Results and discussion

4.1. Results of mapping of parts

Application of direct simulation and compositional geostatistical simulation though ilr transformation generated 100 realizations for each of the parts of the ultimate analysis—i.e. 1200 total realizations for direct and compositional approaches combined. The obvious question is whether these are different from each other and what the consequences are of using direct simulation compared to the more complex compositional approach.

Fig. 10 shows maps of ultimate analysis parts (non-ranked realization 25) generated using direct simulation and compositional simulation for comparison. These maps show that the

values of the parts are visually within the similar range and have similar fluctuations. In fact, E-type maps generated using 100 realizations of C, H, and O (Fig. 11), as examples, are even more similar visually not only in terms of values, but also with their spatial distribution. The biggest difference, though, between direct simulation of analysis results and after a compositional approach using ilr transformation is the sum of values. The sum of parts in ultimate analysis should be 100%, when reporting on an as-received basis, which was the case in all 54 pointwise data used in this work for modeling. This means that the sum of each corresponding cell in all 100 sets of maps of ultimate analysis parts should be 100% too.

Fig. 12 shows the sum of cell values of parts of E-type maps as well as those of 11 realizations that were randomly selected (including realization 25, Fig. 10) for a comparison between direct and compositional simulation. In this figure, the distribution of data of E-type maps is based on 21,000 cell values, whereas the distribution of data of 11 realizations is based on 231,000 cell values. This figure shows that compositional simulation gives exactly 100% as the sum of values of parts, regardless of whether the map is E-type or individual realizations, or regardless of how many realizations are included in the statistics or which cell location is analyzed. However, direct modeling of parts causes the sum to exceed 100% by varying amounts depending on the cell location, violating the mathematics that the parts should sum up to a fixed number. For instance, in the case of E-type maps (21,000 cells), the sum of parts varies between 95% and 103%, with a mean of 99%. Due to the averaging effect of E-type maps, this is rather a good result. When we look at the sum of cell values of parts from 11 individual realizations separately (231,000 cells), the spread of the data is much wider and varies between 79% and 121%, indicating ~± 21% error and that some of the parts have larger values than they actually should have, while others have lesser values at some cell locations (Fig. 12). This is a significant error to exceed the maximum allowed limit for the sum of the parts and indicates that the mathematics is violated. This error in direct simulation is due to using variograms with raw data. As stated by Tolosana-Delgado et al. (2011), raw covariance functions and variograms are as spurious as the correlation matrix of raw data and may lead to spatial interpolations exceeding the maximum allowed limit or may create negative percentages.

Finally, in order to further explore the difference between direct geostatistical simulation using raw compositions and the ones with isometric log ratio transformation, compositional distances between the results were computed. Here, the interest is to exemplify the relative variation of the simulated parts between the two sample sets (raw versus compositional simulations) at each node location shown in the maps. For this purpose, two measures of distance were used. The first one was the Euclidian distance of the ilr balances (Eq. (7)) to calculate distances between orthogonal coordinates (Pawlowsky-Glahn and Egozcue, 2006), and the second one was to use the Aitchison distance (Eq. (8)) between the simulated compositions (van den Boogaart and Tolosana-Delgado, 2013).

$$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i,j}^{D-1}\left(\mathrm{ilr}\ (\boldsymbol{x})_i-\mathrm{ilr}\ (\boldsymbol{y})_j\right)^2} \quad (7)$$

$$d_A(\boldsymbol{x},\boldsymbol{y})=\sqrt{\frac{1}{D}\sum_{i=1}^{D}\sum_{j>i}^{D}\left(\ln\frac{x_i}{x_j}-\ln\frac{y_i}{y_j}\right)^2} \quad (8)$$

In these equations, ***x*** and ***y*** are the D-part compositional vectors; in this case, the ultimate analysis data simulated using raw and ilr-transformation approaches. The distances were calculated for each node of the computation domain.

The compositional distance maps are illustrated using realization 25 and E-type, and are given in Fig. 13. This figure shows that the two distance definitions give very similar results, although the node-based distances computed by using Eq. (8) are about 10% within those of Eq. (7). Further, distances based on E-type data of either compositions or isometric balances of coordinates are less compared to individual realizations of the same due to averaging effects. And finally, it is evident, especially from the realization 25 given in Fig. 13 that the distances are larger in areas where hard data is absent, but geostatistically simulated. This shows an additional benefit of using compositional data treatments prior to geostatistical simulations to eliminate dispersion of data and spurious spatial fluctuations of compositions.

4.2. Application of mapped parts for calorific value computation

Aside from a purely geostatistical modeling of compositional data point of view and a comparison of the two modeling approaches, the next question is what the implications of the differences may be from an engineering point of view. For instance, what would the consequences be if the parts modeled by direct simulation rather than compositional were to be used in correlations to predict other properties? In order to explore this aspect of geostatistical modeling of compositional data, a correlation was established to estimate an as-received calorific value (CV) of Springfield coal by using its ultimate properties. The correlation was established by using 265 data points available in the Indiana Coal Quality Database (Drobniak and Mastalerz, 2012a). This correlation (R^2 =0.85) is given in Eq. (7).

$$\mathrm{CV}\left(\frac{\mathrm{Btu}}{\mathrm{lb}}\right)=164.863\times(C\%)+46.307\times(O\%)+73.678\times(S\%) \quad (9)$$

In order to estimate calorific value distribution of Springfield coal, all parts realizations generated using direct and compositional simulation were used in the correlation. To continue with the same realization as an example, Fig. 14 shows the calorific value distribution of realization 25 using parts maps generated through direct and compositional simulation. These maps show that the calorific value is distributed very similarly within the coal seam. However, the map generated using the parts from compositional simulation exhibits data in a narrow range compared to the one from direct simulation. Fig. 15 shows calorific value distribution of 231,000 cells of 11 realizations studied previously in this section for the sum of their cell values. This figure shows that the calorific value estimated using the results of compositional simulation ranges between 10,800 and 12,500 Btu/lb, whereas the one estimated using the results of direct simulation ranges between 10,000 and

13,300 Btu/lb. The larger range of calorific value data calculated for direct simulation may be due to the range of simulated cell values of the parts.

In order to compare which one of these mapping techniques—direct or compositional simulation of parts—and the resultant data distributions give closer estimates to the maps of calorific value that is modeled directly from the pointwise data, measured calorific values of initial 54 data points collocated with ultimate analysis parts were modeled by sequential Gaussian simulation. In this process, variograms of 54 pointwise calorific values were modeled and 100 realizations were generated by using exactly the same simulation parameters, including the initial seed number, as in the simulation of ultimate analysis parts. For comparison, the realizations corresponding to the order of the 11 realizations presented before were selected as the benchmark data.

Fig. 16-A shows the distribution of calorific values in 11 realizations that are used as a benchmark. The data ranges from 10,600 Btu/lb to 12,600 Btu/lb, with a mean of 11,500 Btu/lb. This data spread is closer to the data range from calorific value maps predicted using parts from compositional modeling through correlation. In order to calculate the prediction errors, relative errors with respect to the benchmark data were calculated (Eq. (10)). For this purpose, corresponding realizations of compositional and direct simulation predictions were subtracted from the corresponding realization of the benchmark data, and relative errors were calculated for each cell. Fig. 16-B shows the distribution errors based on 231,000 cells and that the errors incurred by using compositional modeling in correlation are fewer compared to using direct modeling results in correlations. Basic statistics of benchmark data and the calculated errors are given in Table 4.

$$\text{Error } (\%) = 100 \times \frac{(\text{BTU}_{\text{bm}_i})_{x,y} - (\text{BTU}_{\text{simi}})_{x,y}}{(\text{BTU}_{\text{bmi}})_{x,y}} \quad (10)$$

In this equation, $(BTU_{bmi})_{x,y}$ is the ith realization of benchmark data and $(BTU_{simi})_{x,y}$ is the corresponding realizations computed through Eq. (9) using parts mapped by direct or compositional simulation. The subscripts x and y are the grid cells.

Although the methodology explained in this paper is for coal and the data is ultimate analysis data, the method is general and is applicable to any setting and any compositional data. Analysis and modeling of ultimate data showed that the compositional modeling approach is coherent and does not violate mathematical constraints that are imposed on the measurements or the basis that the data is reported by. As important as mathematical coherence, using results of compositional modeling in correlations, rather than direct modeling, to predict non-compositional parameters produces more accurate results. This may have engineering and safety consequences. In this paper, one such application was demonstrated for predicting in-situ calorific value of coal due to data availability. This has economic and resource utilization benefits. However, for instance, methane content of coal seams can be predicted using proximate analysis or maceral compositions in correlations by using the same approach. Similarly, self-heating temperature can be predicted using correlations involving composition of coal. Accurate prediction of such parameters has the

potential to improve safety and productivity of mines and also has the potential to guide engineers better in the design of mines and of methane and self-heating control systems. Therefore, going one extra step in modeling through compositional modeling, when the data requires such, has multiple benefits to improve the economics and safety of operations.

5. Summary and conclusions

In this paper, a compositional modeling approach through ilr transformation was demonstrated to map ultimate properties of Springfield coal in the Indiana section of the Illinois basin. The results were compared with the results of direct spatial modeling of raw measurements. Further, the results were used in a correlation to predict calorific value distribution within the coal seam for comparison purposes.

This study demonstrated that by using compositional simulation using ilr and sequential Gaussian simulation, the benefits of both techniques are taken advantage of. Sequential Gaussian simulation offers the advantages that the errors are conditionally unbiased, the histogram of each realization is the same as the realization of the data, and multiple realizations generated stochastically can be used for uncertainty assessment. On the other hand, compositional modeling ensures that mathematics and the limits of the data are not violated. In our study, compositional simulation gave exactly 100% as the sum of values of parts, regardless of the type of the map and how many realizations are included in assessment. Direct modeling of parts using raw data, on the hand, created as much as ~ ± 21% error in sum of the part values in individual realizations due to direct variogram modeling of compositional data.

In order to exemplify the relative variation of the simulated parts using raw compositions and the ones with isometric log ratio transformation, compositional distances between the results were computed. The results showed that the distances were larger in areas where hard data was absent, but geostatistically simulated. This indicates that using compositional data treatments prior to geostatistical simulations helps eliminating dispersion of data and spurious spatial fluctuations of compositions.

Application of part values modeled using both approaches were used in a correlation to predict calorific value of coal. The results were compared with modeling of collocated pointwise calorific value data. Results showed that the data spread was closer to the data range from calorific value maps predicted using parts from compositional modeling, which showed less prediction error compared to direct geostatistical modeling of parts.

Using, or not using, compositional modeling when the data calls for such may have important implications for safety and engineering consequences. For instance, besides other non-compositional parameters, methane content of coal seams can be predicted using proximate analysis or maceral compositions in correlations. Likewise, self-heating temperature can be predicted using correlations involving composition of coal. Better prediction of such parameters and their distributions within the seam has the potential to improve safety and productivity of mines and to guide engineers towards more efficient design of methane and self-heating control systems.

Acknowledgments

Profs. Vera Pawlowsky-Glahn, Juan Jose Egozcue and Karel Hron are gratefully acknowledged for reviewing earlier version of this paper and for making useful comments.

References

Aitchison J. Principal component analysis of compositional data. Biometrika. 1983; 70(1):57–65.

Aitchison, J. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd.; London: 1986. The Statistical Analysis of Compositional Data; p. 416

ASTM D121. Annual Book of ASTM Standards. Vol. 05.06. ASTM International; West Conshohocken, PA: 2006.

van den Boogaart, KG., Tolosana-Delgado, R. Analyzing Compositional Data With R. Springer-Verlag; Berlin Heidelberg: 2013. p. 258

Buccianti A, Pawlowsky-Glahn V. New perspectives on water chemistry and compositional data analysis. Math. Geol. 2005; 37:271–275.

Buccianti A, Egozcue JJ, Pawlowsky-Glahn V. Variation diagrams to statistically model the behavior of geochemical variables: theory and applications. J. Hydrol. 2014; 519:988–998.

Caers, J. Modeling Uncertainty in the Earth Sciences. Wiley-Blackwell; Chichester, UK: 2011. p. 229

Chayes F. On correlation between variables of constant sum. J. Geophys. Res. 1960; 65(12):4185–4193.

Chelgani SC, Hower JC, Jorjani E, Mesroghli Sh, Bagherieh AH. Prediction of coal grindability based on petrography, proximate and ultimate analysis using multiple regression and artificial neural network models. Fuel Process. Technol. 2008; 89(1):13–20.

Chilès, JP., Delfiner, P. Geostatistics: Modeling Spatial Uncertainty. second. John Wiley & Sons, Inc.; Hoboken, NJ: 2012. p. 734

Comas, M., Thió-Henestrosa, S. CoDaPack 2.0: a stand-alone, multiplatform compositional software. In: Egozcue, JJ.Tolosana-Delgado, R., Ortego, MI., editors. Proceedings of the 4th International Workshop on Compositional Data Analysis, Sant Feliu de Guixols. International Center for Numerical Methods in Engineering; 2011. p. 10http://ima.udg.edu/codapack/

Deutsch, CV., Journel, AG. GSLIB—Geostatistical Software Library and User's Guide. 2. Oxford University Press; New York: 1998. p. 369

Drobniak A, Mastalerz M. The Indiana Geological Survey Coal Stratigraphic Database—The Database and Interactive Map: Indiana Geological Survey Report of Progress 39. CD-ROM. 2012a

Drobniak A, Mastalerz M. The Indiana Coal Quality Database—The Database and Interactive Map: Indiana Geological Survey Report of Progress 40. CD-ROM. 2012b

Egozcue JJ, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. Math. Geol. 2005; 37(7):795–828.

Egozcue, JJ., Pawlowsky-Glahn, V. Simplicial geometry for compositional data. In: Buccianti, A.Mateu-Figueras, G., Pawlowsky-Glahn, V., editors. Compositional Data Analysis in the Geosciences: From Theory to Practice. Vol. 264. Geological Society, London: Special Publications; 2006. p. 67-77.

Egozcue, JJ., Pawlowsky-Glahn, V. Basic concepts and procedures. In: Pawlowsky-Glahn, V., Buccianti, A., editors. Compositional Data Analysis, Theory and Practice. Wiley; 2011. p. 378

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. Math. Geol. 2003; 35(3):279–300.

Engle MA, Rowan EL. Geochemical evolution of produced waters from hydraulic fracturing of the Marcellus Shale, northern Appalachian Basin: a multivariate compositional data analysis approach. Int. J. Coal Geol. 2014; 126:45–56.

Filzmoser M, Hron K. Correlation analysis for compositional data. Math. Geosci. 2009; 41(8):905–919.

Filzmoser P, Hron K, Reimann C. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci. Total Environ. 2009; 407:6100–6108. [PubMed: 19740525]

Geboy NJ, Engle MA, Hower JC. Whole-coal versus ash basis in coal geochemistry: a mathematical approach to consistent interpretations. Int. J. Coal Geol. 2013; 113:41–49.

Mathews JP, Krishnamoorthy V, Louw E, Tchapda AHN, Castro-Marcano F, Karri V, Alexis DA, Mitchell GD. A review of the correlations of coal properties with elemental composition. Fuel Process. Technol. 2014; 121:104–113.

Olea, RA. Geostatistics for Engineers and Earth Scientists. Kluwer Academic Publishers; Norwell, MA: 1999. p. 303

Olea RA, Luppens JA. Mapping of coal quality using stochastic simulation and isometric logratio transformation with an application to a Texas lignite. Int. J. Coal Geol. 2016; 152:80–93.

Olea RA, Luppens JA, Egozcue JJ, Pawlowsky-Glahn V. Calorific value and compositional ultimate analysis with a case study of a Texas lignite. Int. J. Coal Geol. 2016:27–33.

Otero N, Tolosana-Delgado, Soler A, Pawlowsky-Glahn V, Canals A. Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a Mediterranean river. Water Res. 2005; 39(7):1404–1414. [PubMed: 15862341]

Owen DDR, Pawlowsky-Glahn V, Egozcue JJ, Buccianti A, Bradd JM. Compositional data analysis as a robust tool to delineate hydrochemical facies within and between gas-bearing aquifers. Water Resour. Res. 2016; 52:5771–5793.

Park N-W, Jang DH. Comparison of geostatistical kriging algorithms for intertidal surface sediment facies mapping with grain size data. Sci. World J. 2014; 2014:1–11. issue.. http://www.hindawi.com/journals/tswj/2014/145824/.

Pawlowsky-Glahn V, Egozcue JJ. Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. 2001; 15(5):384–398.

Pawlowsky-Glahn V, Egozcue JJ. Spatial analysis of compositional data: a historical review. J. Geochem. Explor. 2016; 164:28–32.

Pawlowsky-Glahn, V., Olea, RA. Monograph 7, Studies in Mathematical Geology. Oxford University Press; New York: 2004. p. 181

Pawlowsky-Glahn, V., Egozcue, JJ. Compositional data and their analysis: an introduction. In: Buccianti, A.Mateu-Fugueras, G., Pawlowsky-Glahn, V., editors. Compositional data analysis in the geosciences: from theory to practice. Vol. 264. Geological Society, London: Special Publications; 2006. p. 1-10.

Pyrcz, MJ., Deutsch, CV. Geostatistical Reservoir Modeling. second. Oxford University Press; New York: 2014. p. 433

Reimann C, Filzmoser P, Fabian K, Hron K, Birke M, Demetriades A, Dinelli A, Ladenberger A. The GEMAS Project Team. The concept of compositional data analysis in practice— Total major element concentrations in agricultural and grazing land soils of Europe. Sci. Total Environ. 2012; 426:196–210. [PubMed: 22503552]

Remy, N., Boucher, A., Wu, J. Applied Geostatistics With SGeMS, A User's Guide. Cambridge University Press; Cambridge, United Kingdom: 2009. p. 264

Riley, JT. Manual. Vol. 57. ASTM International; West Conshohocken, PA: 2007. Routine Coal and Coke Analysis: Collection, Interpretation, and Use of Analytical Data.

Srivastava RM. Geostatistics: a toolkit for data analysis, spatial prediction and risk management in the coal industry. Int. J. Coal Geol. 2013; 112:2–13.

Thompson TA, Sowder K, Johnson M. Generalized Stratigraphic Column of Indiana Bedrock, Indiana Geological Survey Poster 6. 2013

Tolosana-Delgado R, van den Boogaart KG. Towards compositional geochemical potential mapping. J. Geochem. Explor. 2014; 141:42–51.

Tolosana-Delgado, R., van den Boogaart, KG., Pawlowsky-Glahn, V. Geostatistics for compositions. In: Pawlowsky-Glahn, V., Buccianti, A., editors. Compositional Data Analysis, Theory and Practice. Wiley; 2011. p. 378

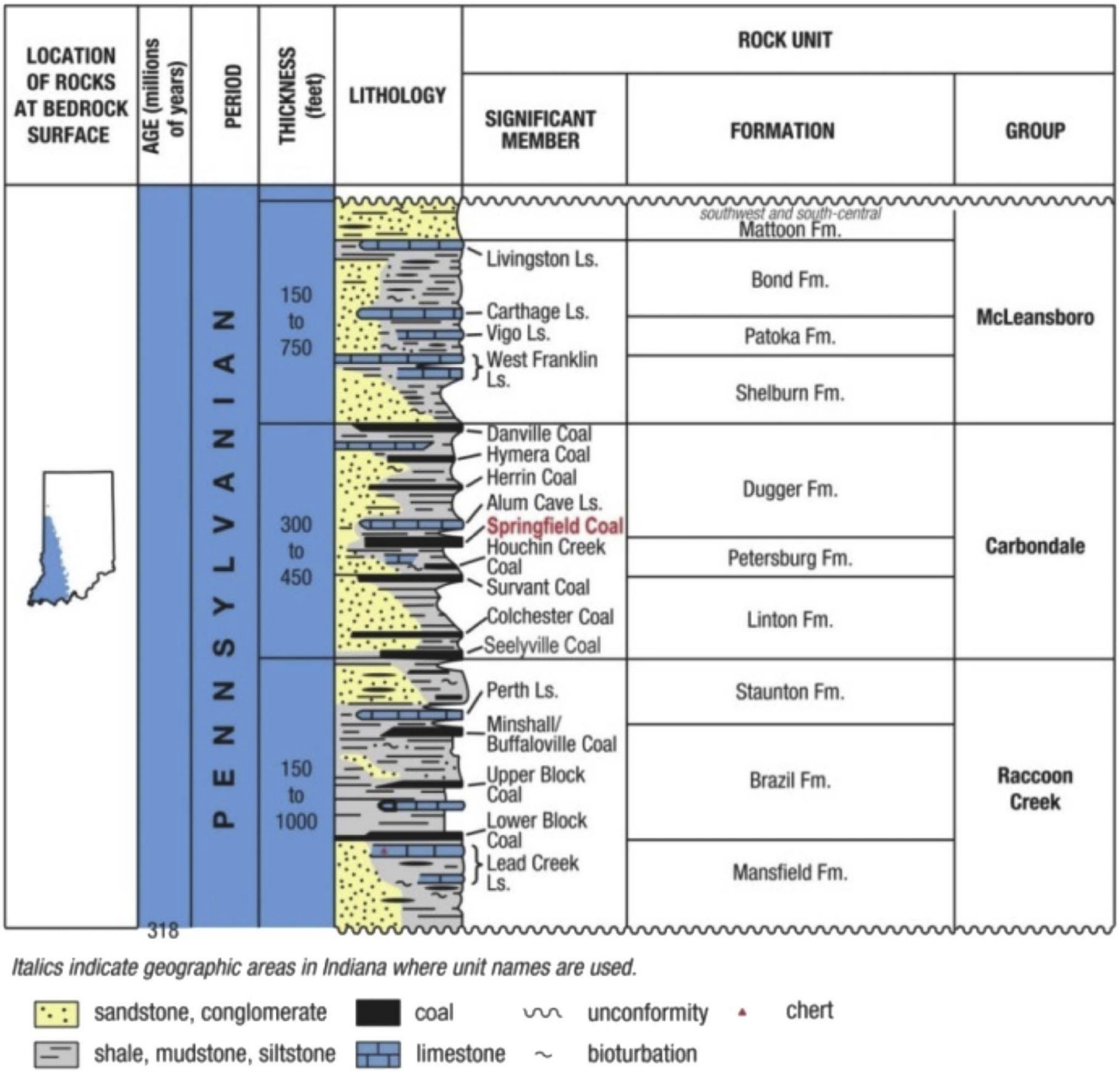

Fig. 1.

General stratigraphy of Indiana's Pennsylvanian system of formations and coal members (after Thompson et al., 2013).

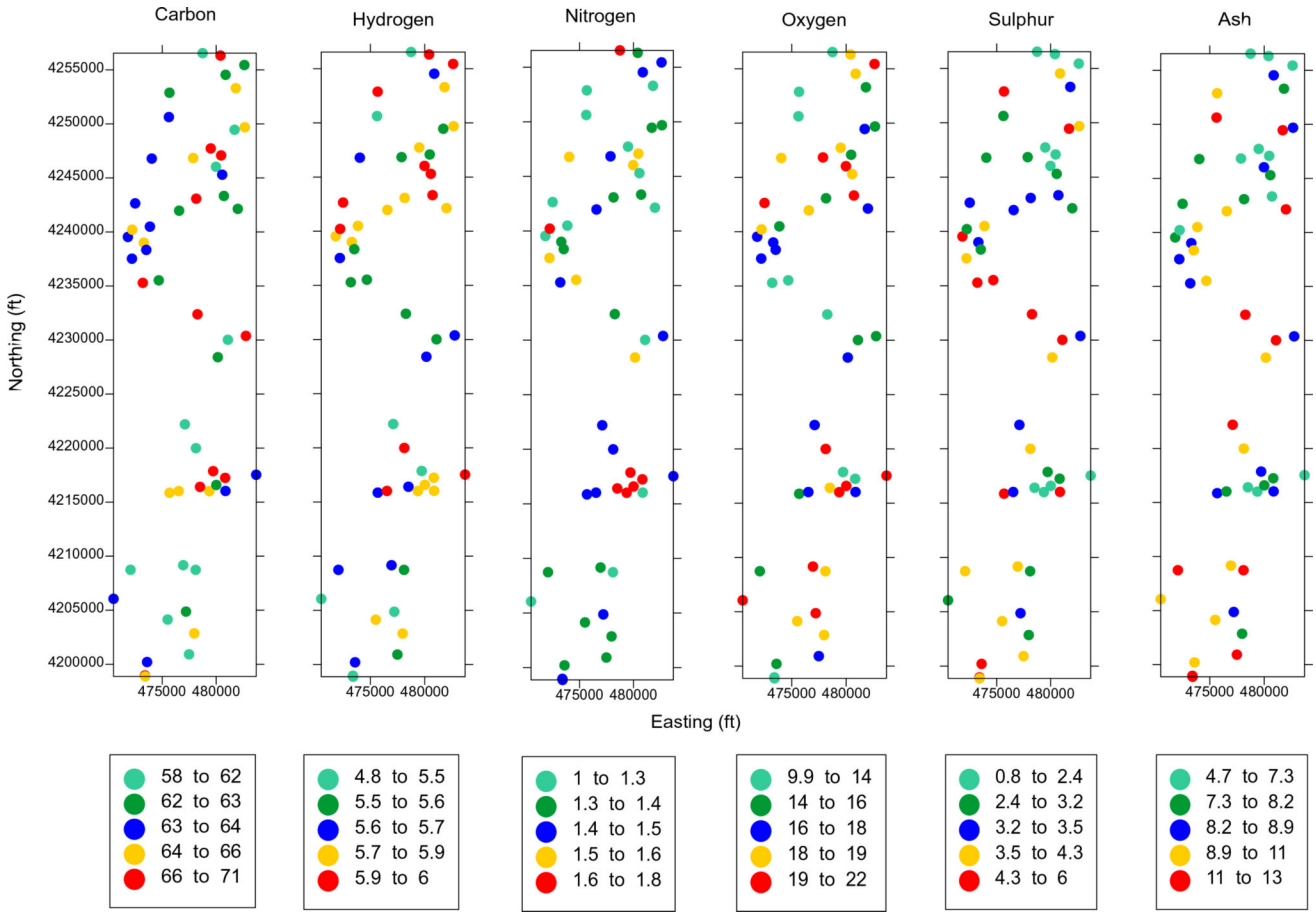

Fig. 2.

Posting of the ultimate properties of Springfield coal within the studied area. Values are in %.

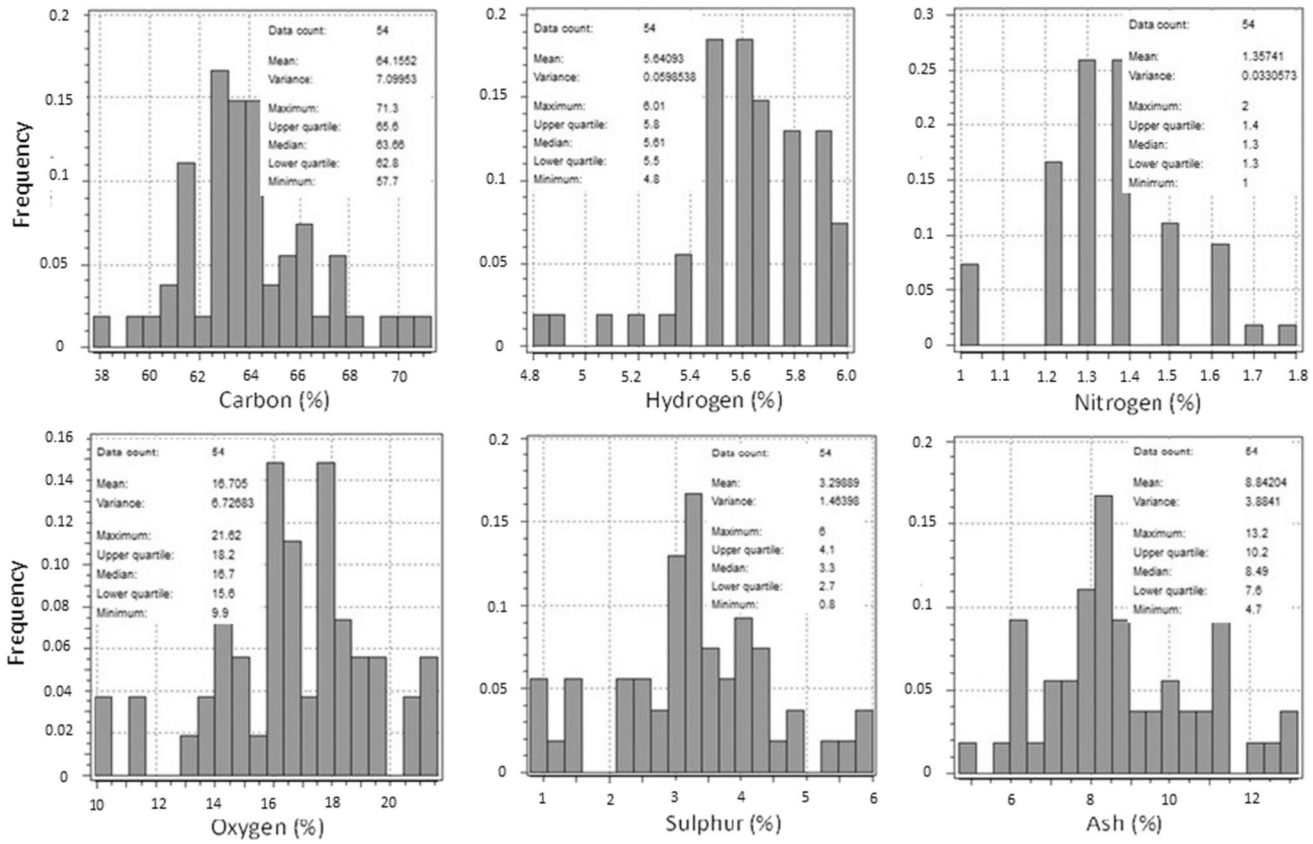

Fig. 3.

Frequency histograms of the raw ultimate analysis of Springfield coal within the studied area.

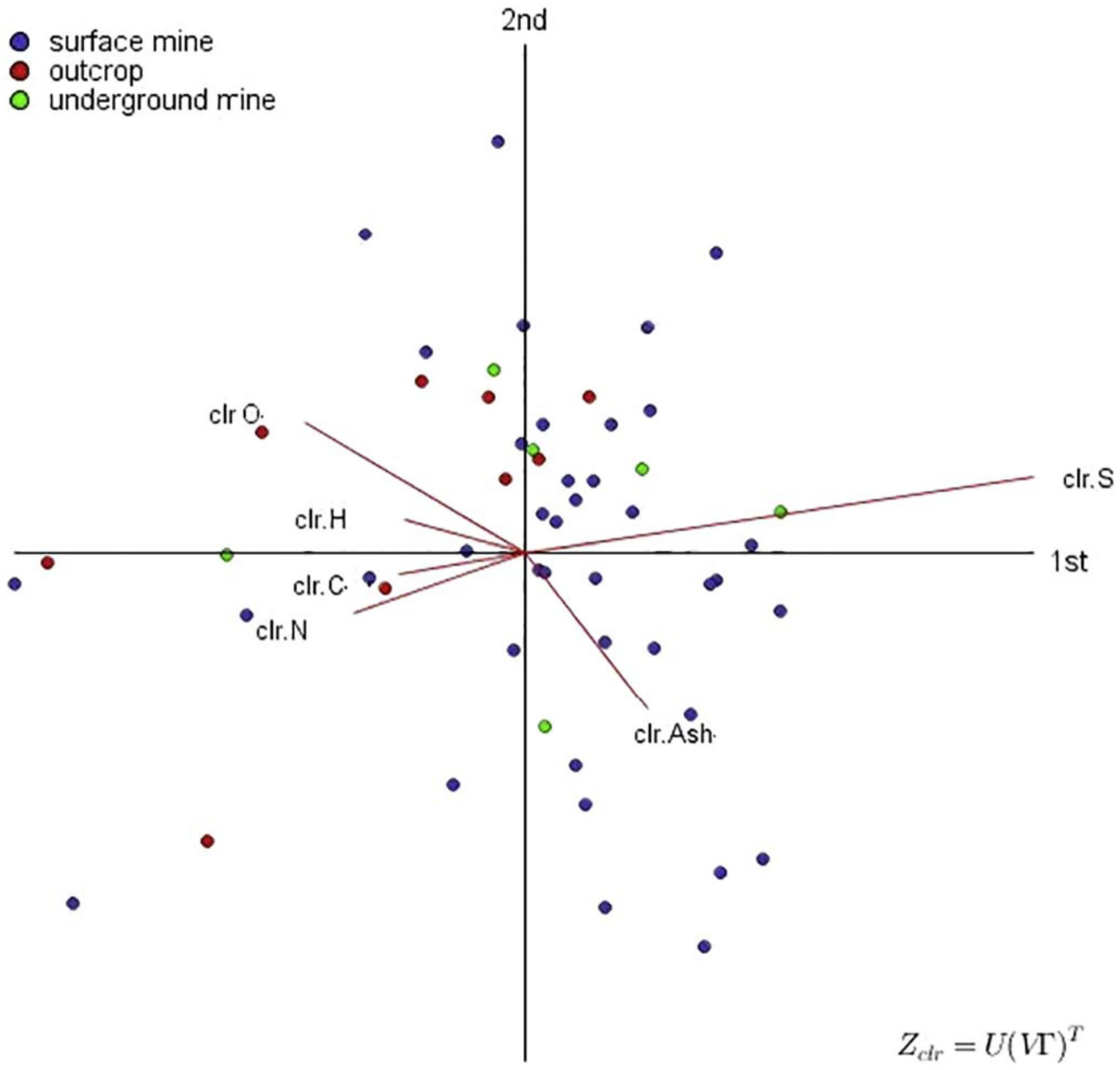

Fig. 4.
Compositional clr bi-plot of the ultimate analysis parts from 54 samples presented in Figs. 2 and 3.

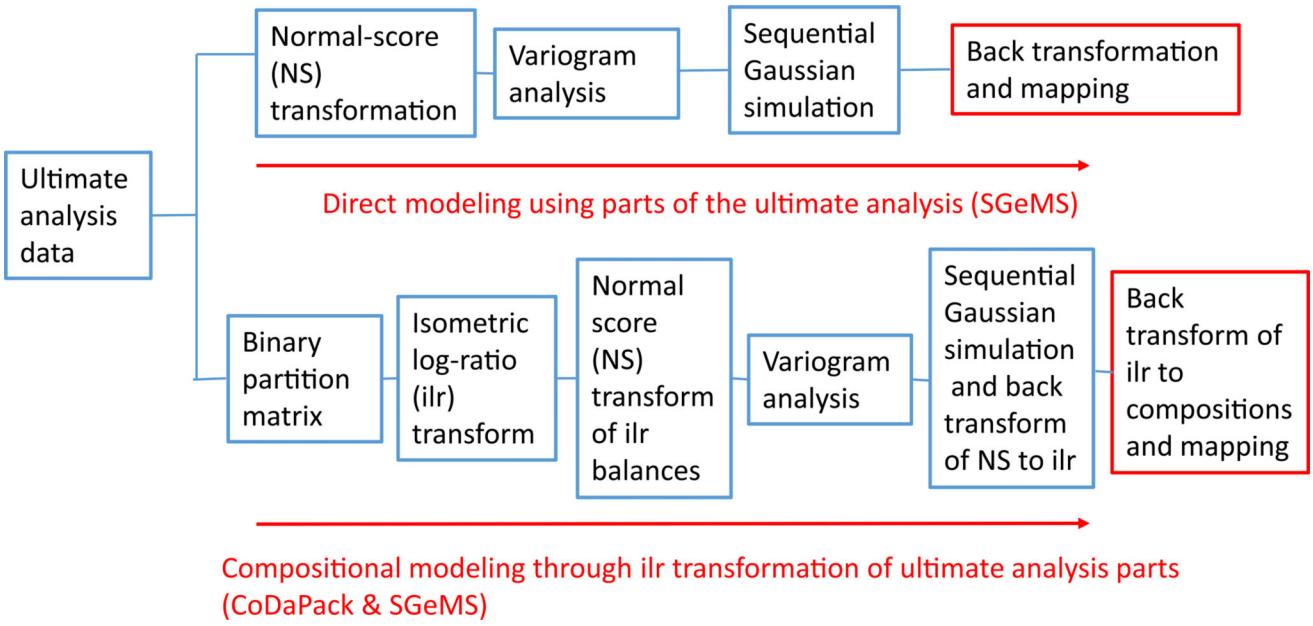

Fig. 5.
Methodology applied in this work and the workflow for each approach.

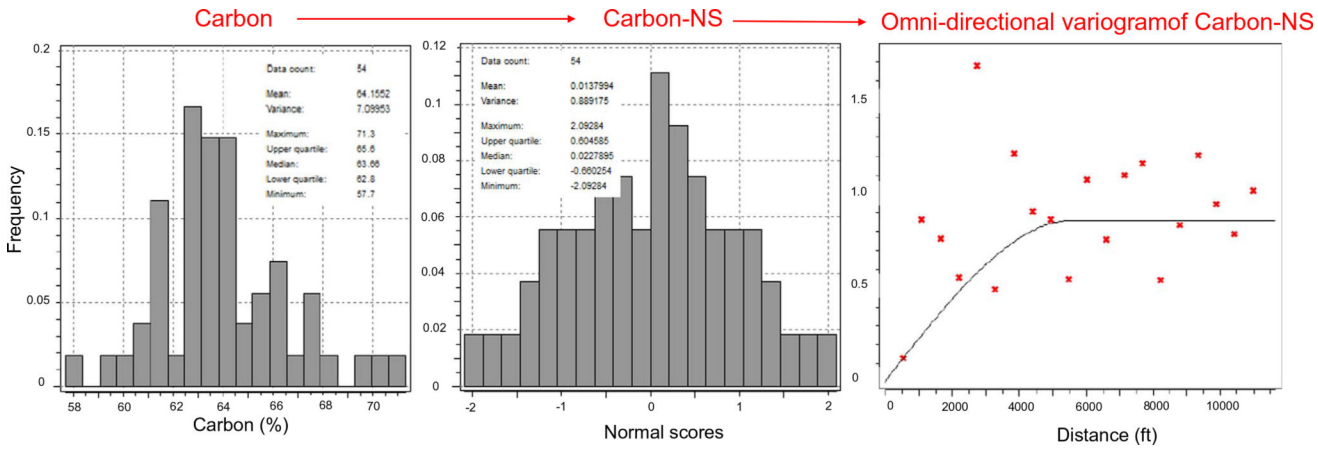

Fig. 6.

Progression of variogram building (for carbon %) for direct geostatistical modeling.

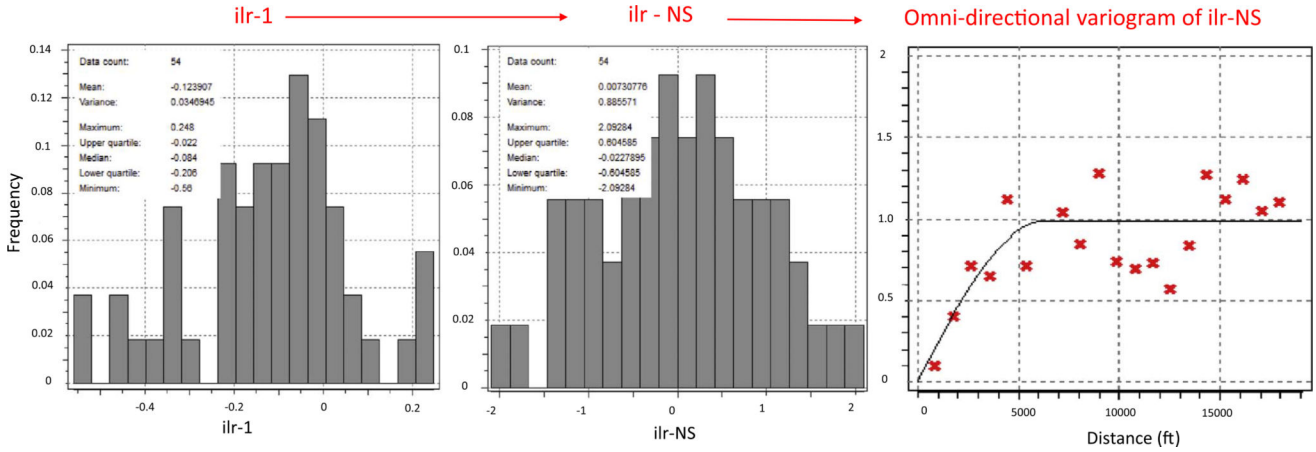

Fig. 7.

Progression of variogram building for normal scores of ilr balances to be used in sequential Gaussian simulation in compositional modeling.

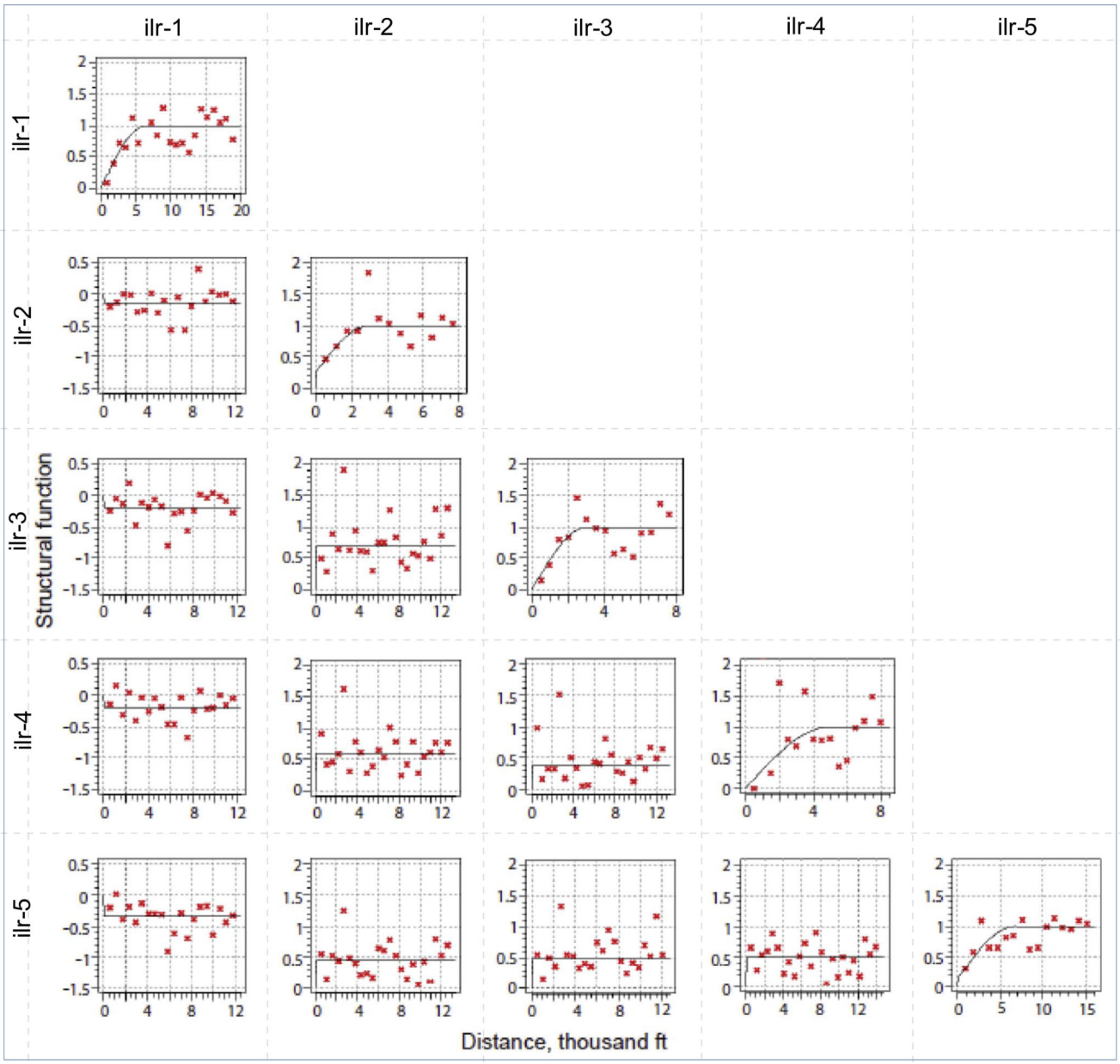

Fig. 8.
Structural functions, auto- and cross-variograms, between normal scores of ilr balances.

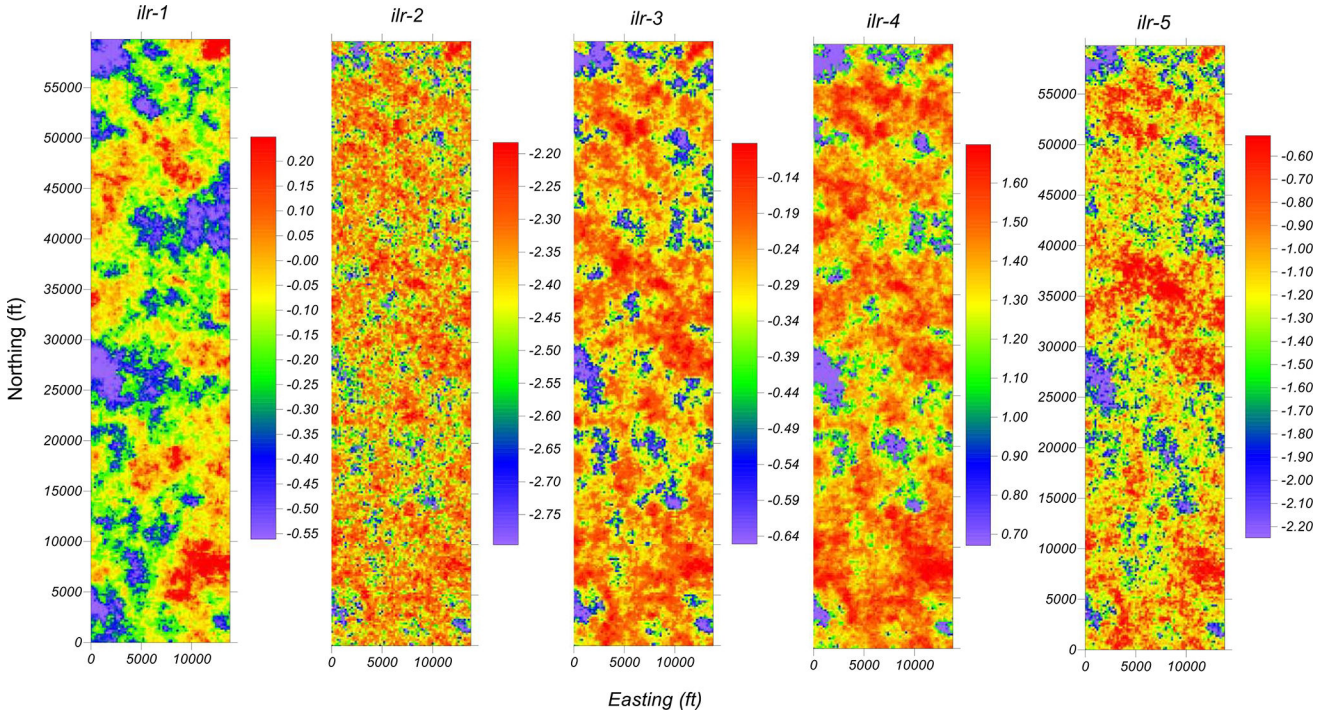

Fig. 9.

Maps of ilr balances (realization 25) generated using sequential Gaussian simulation.

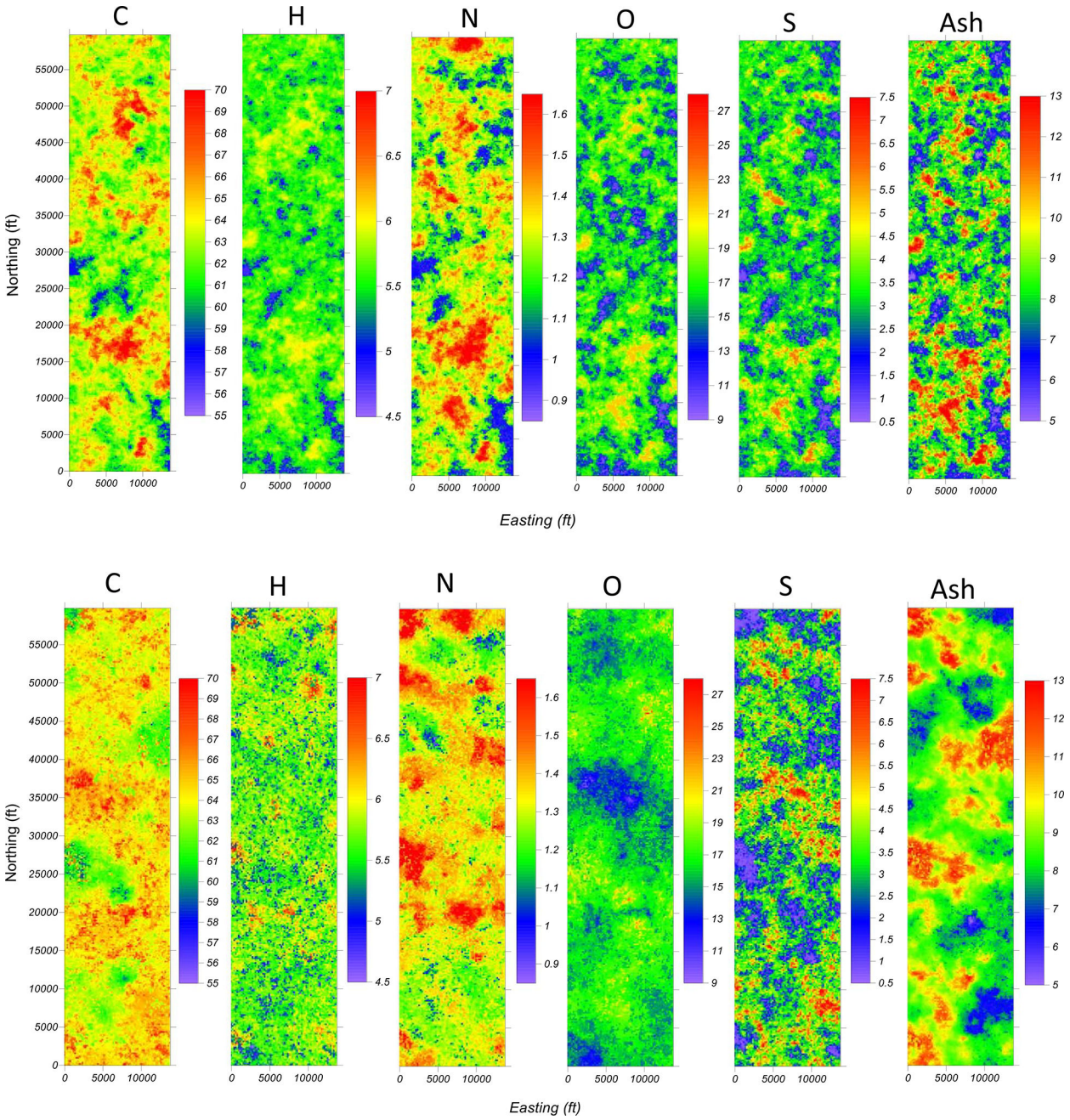

Fig. 10.

Maps of ultimate analysis parts (realization 25) using direct simulation (top row) and compositional simulation (bottom row). Values are in %.

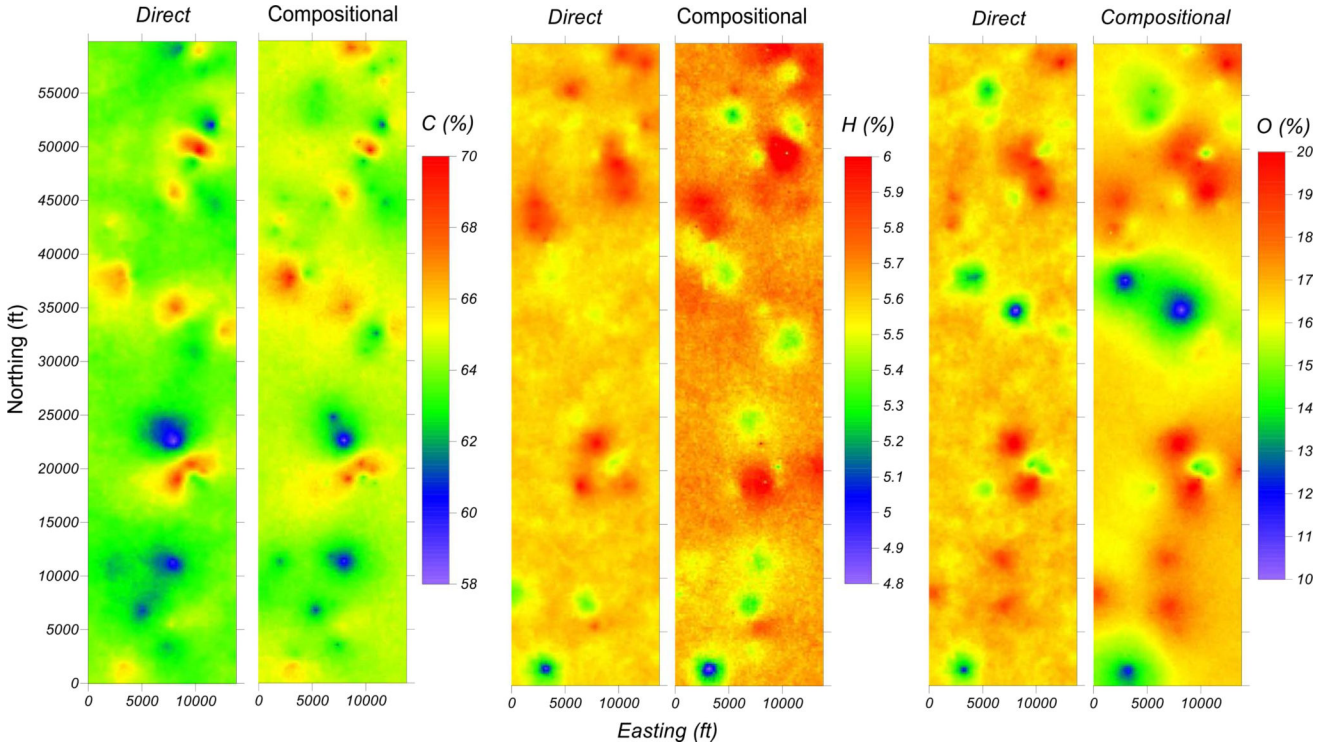

Fig. 11.

E-type maps of C, H, and O generated based on 100 realizations.

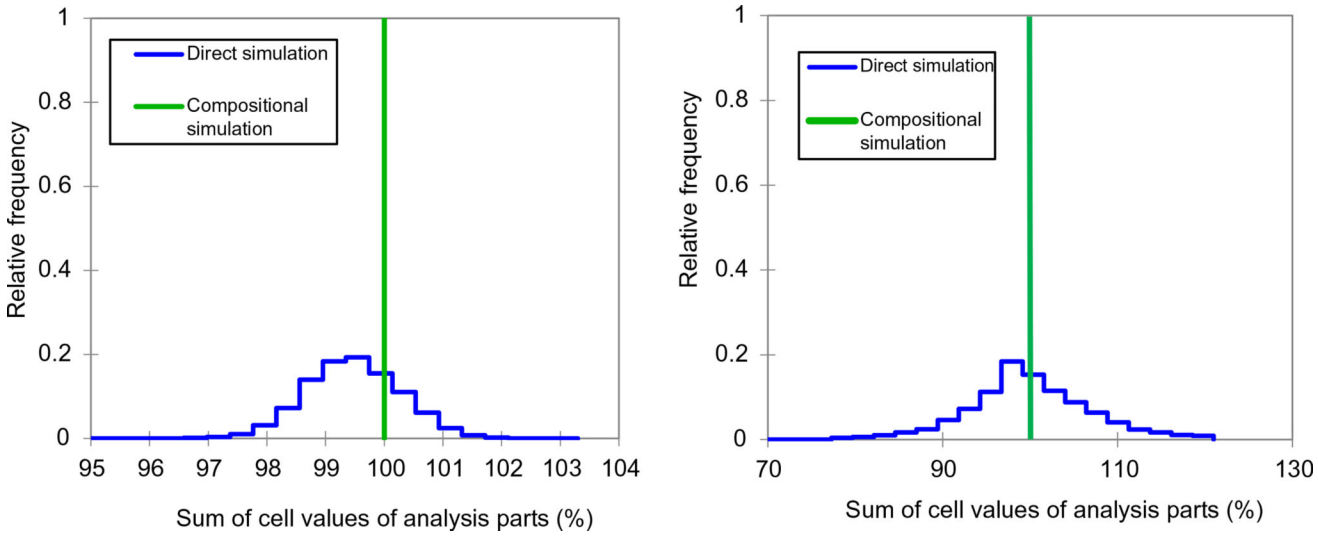

Fig. 12.

Sum of cell values of parts of E-type maps (A) as well as those of 11 randomly selected realizations (B), as a comparison between direct and compositional simulation.

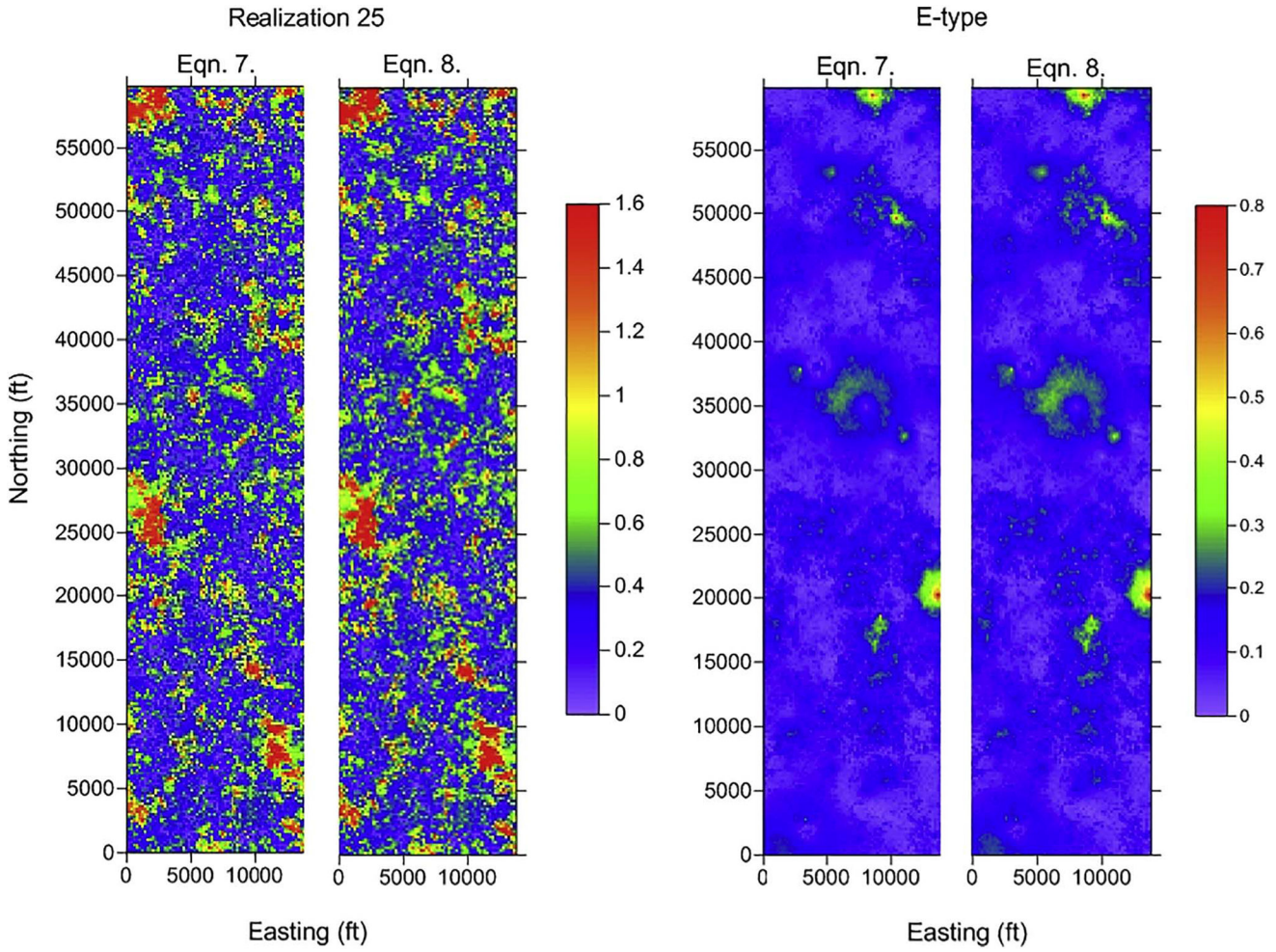

Fig. 13.
Compositional distance maps computed using Eqs. (7) and (8).

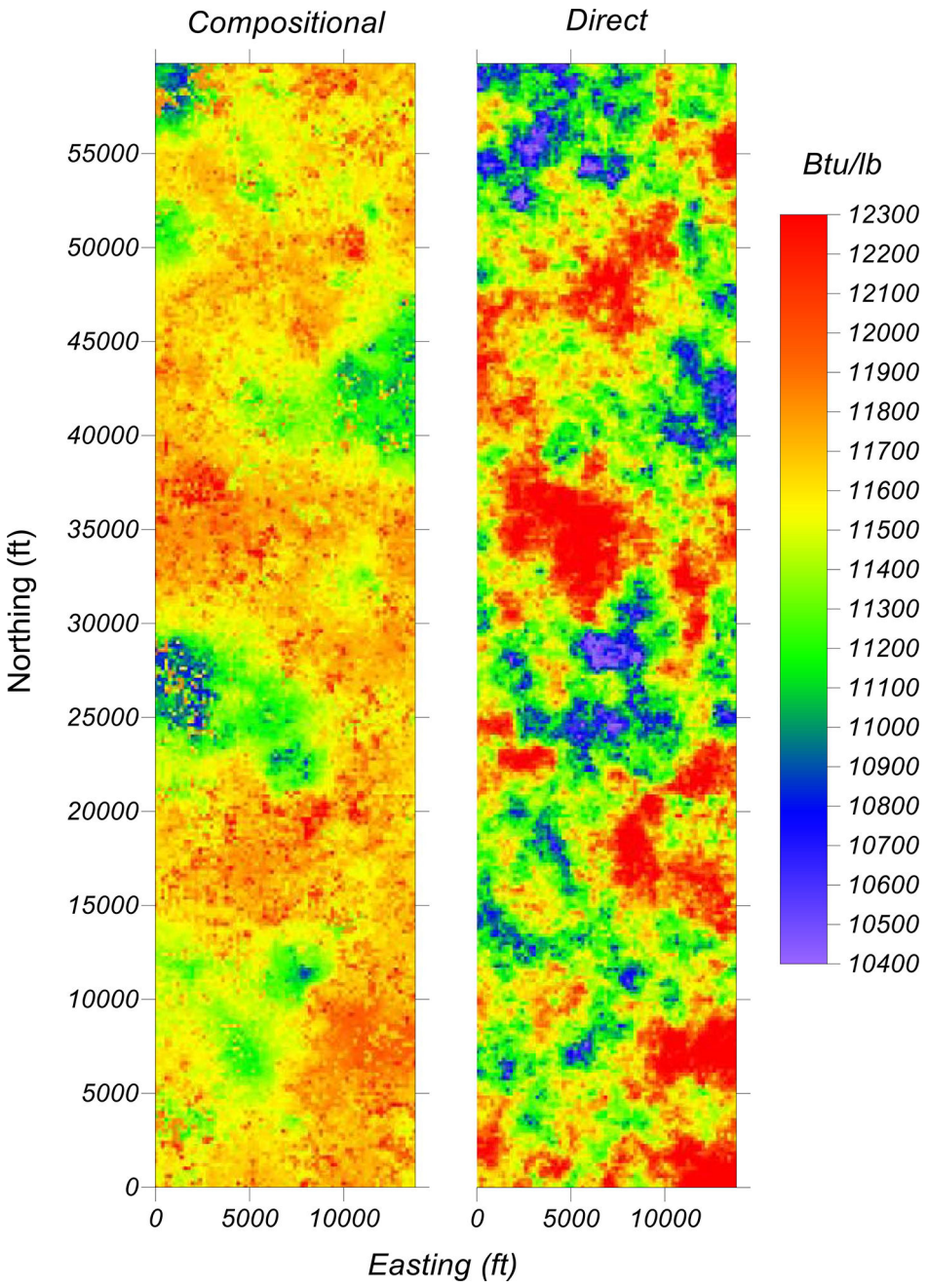

Fig. 14.

Calorific value distribution of realization 25 using parts maps generated through direct and compositional simulation.

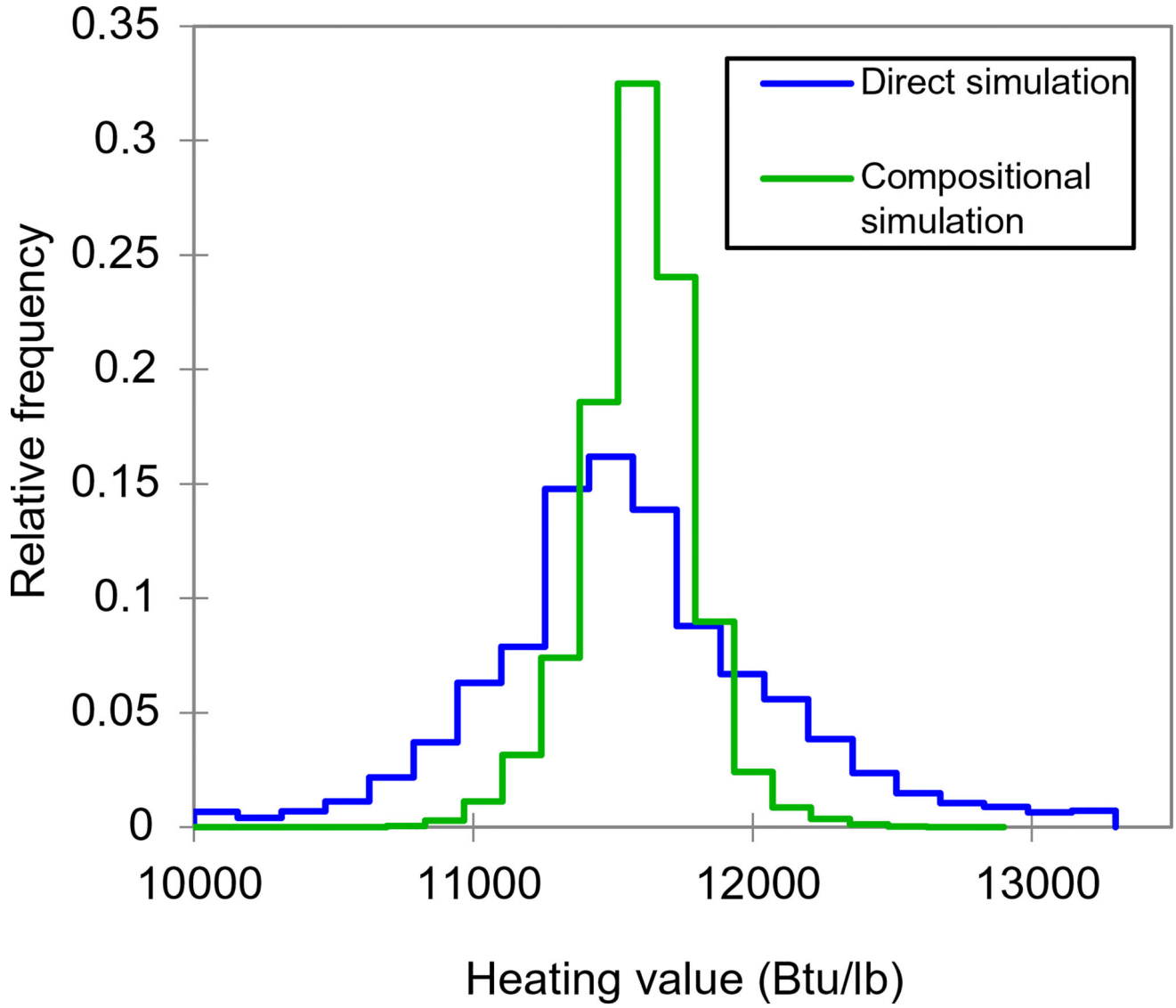

Fig. 15.
Calorific value distribution of cells of 11 realizations calculated using the results of direct and compositional simulation of ultimate analysis parts.

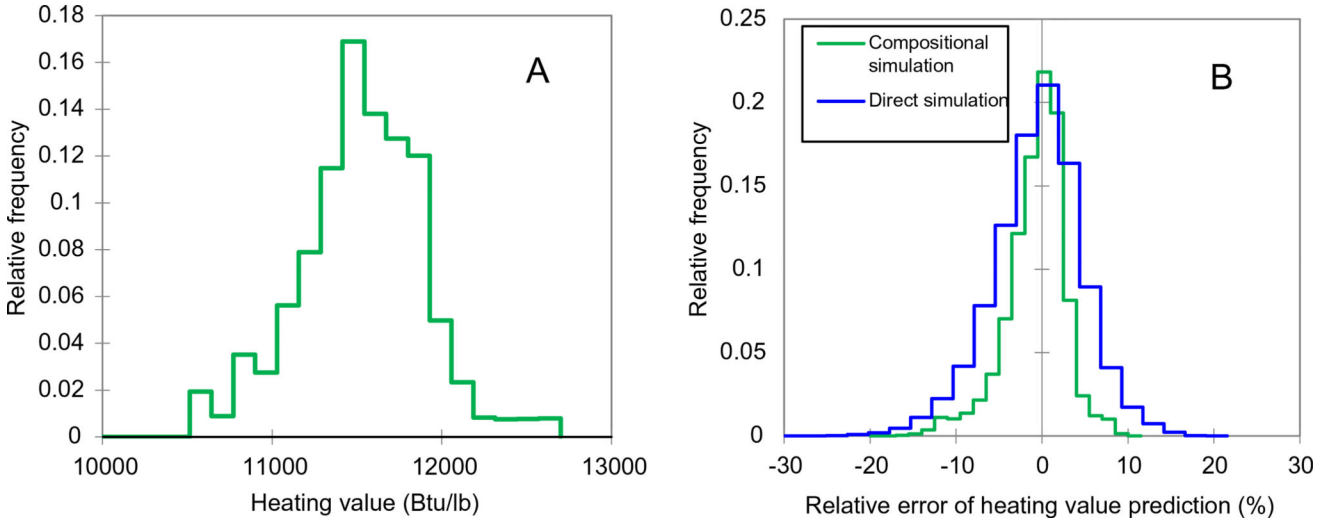

Fig. 16.

Distribution of data from cells of 11 realizations modeled using 54 collocated calorific value points with the ultimate analysis data (A), and the errors incurred by using the results of compositional and direct modeling of parts in correlation (B).

Table 1

Basic standard descriptive statistics of the parts of ultimate analysis based on raw data just for exploration purposes.

	Mean	Std. dev.	Upper quartile	Median	Lower quartile	Min	Max
Ash (%)	8.8	2.0	10.2	8.5	7.6	4.7	13.2
C (%)	64.0	2.6	65.2	63.6	62.7	57.7	71.3
H (%)	5.6	0.3	5.8	5.6	5.5	4.8	6.0
N (%)	1.4	0.2	1.4	1.4	1.3	1.0	1.8
O (%)	16.7	2.6	18.2	16.8	15.1	9.9	21.6
S (%)	3.3	1.2	4.1	3.3	2.7	0.8	6.0

Table 2

Analytical variograms and their parameters of normal scores of each of the raw parts (in %).

Component	Model	Nugget	Sill-nugget	Range (ft)
C	Spherical	0.01	0.84	5618
H	Spherical	0.01	0.96	3040
N	Spherical	0.01	0.81	5408
O	Spherical	0.00	0.85	2500
S	Spherical	0.01	0.64	2700
Ash	Spherical	0.01	0.74	2665

Table 3

Analytical variograms and their parameters of normal scores of ilr balances.

ilr	Model	Nugget	Sill-nugget	Range (ft)
1	Spherical	0.01	0.98	6300
2	Spherical	0.25	0.75	3050
3	Spherical	0.01	0.96	3100
4	Spherical	0.02	0.98	5000
5	Spherical	0.15	0.85	6700

Table 4

Basic statistics of benchmark calorific value data and of the errors (Eq. (10)) incurred using compositional and direct simulations in correlation (Eq. (9)).

Variable	Cells	Min.	Max.	Mean	Std. dev.
Calorific value (Btu/lb) – benchmark data	231,000	10,589	12,617	11,526	365.2
Rel. error (%) – compositional sim.	231,000	− 18.1	10.5	− 0.703	3.477
Rel. error (%) – direct sim.	231,000	− 24.8	20.6	− 0.505	5.269

J Geochem Explor. Author manuscript; available in PMC 2019 March 01.