ACS Combinatorial Science

Research Article

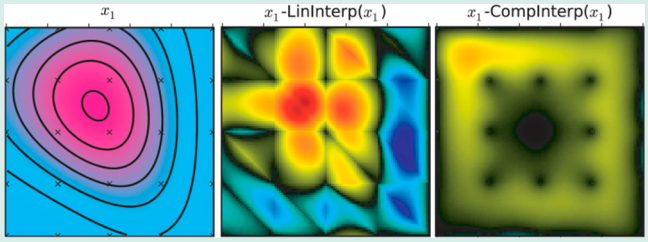
pubs.acs.org/acscombsci

Statistical Analysis and Interpolation of Compositional Data in Materials Science

Misha Z. Pesenson,* Santosh K. Suram, and John M. Gregoire*

Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, California 91125, United States

ABSTRACT: Compositional data are ubiquitous in chemistry and materials science: analysis of elements in multicomponent systems, combinatorial problems, etc., lead to data that are non-negative and sum to a constant (for example, atomic concentrations). The constant sum constraint restricts the sampling space to a simplex instead of the usual Euclidean space. Since statistical measures such as mean and standard deviation are defined for the Euclidean space, traditional correlation studies, multivariate analysis, and hypothesis testing may lead to erroneous dependencies and incorrect inferences when applied to compositional data. Furthermore, composition measurements that are used for data analytics may not include all of the elements contained in the material; that is, the measurements may be subcompositions of a higher-dimensional parent composition. Physically meaningful statistical analysis must yield results that are invariant under the number of composition elements, requiring the application of specialized statistical tools. We present specifics and subtleties of compositional data processing through discussion of illustrative examples. We introduce basic concepts, terminology, and methods required for the analysis of compositional data and utilize them for the spatial interpolation of composition in a sputtered thin film. The results demonstrate the importance of this mathematical framework for compositional data analysis (CDA) in the fields of materials science and chemistry.



KEYWORDS: *high-throughput screening, electrocatalyst, inkjet printing, sputtering, thin-films, interpolation, compositional data, big data, complex data, statistical data analysis*

I. INTRODUCTION

Recent and forthcoming advances in instrumentation are creating materials science data sets that are not amenable to the application of the existing, standard methods of analysis.[1−12] Present-day data do not "speak for themselves" as suggested by the familiar slogan, which was coined *before* the modern era of complex high-throughput data. Rather, one has to learn from data, and statistical learning and modeling provide a means for extracting knowledge from data.[13,14] Traditional techniques are often inadequate not merely because of the size in bytes of the data sets but also because of the complexity of modern data sets.[1−6] At the same time, it is precisely the richness and complexity of new data sets that provide materials science with a wealth of information. Indeed, most of the research progress expected from such sets inherently rests in their enormity and complexity which enable data-driven decisions. Developing and applying statistically sound methods that allow one to accurately represent and interpret data is very important for data-driven materials optimization.

Traditionally, research in materials science has focused on synthesis, characterization and property measurement for select compositions of interest. With the quest for rapid discovery of functional materials, large scale combinatorial experiments that produce composition-property maps of higher order composition spaces have become increasingly important, in particular due to the ability to tailor material properties using multiple elements.[6,15,16] Because individual components of a composition are not free to vary independently, each component cannot be interpreted without relating it to any of the other components. This makes compositional data intrinsically multivariate. Statistical methods for compositional data analyses have been and continue to be developed in the field of statistics since the 1980s,[17,18] but have not been readily applied to the relevant topics in materials science.

The extensive amount of relevant work on statistical analysis of compositional data already accomplished in disciplines outside of materials science does not allow us to offer a complete review of all aspects of these complex topics and problems, and an interested reader is referred to works in bibliography.[18−22] The structure of the paper is as follows: In section II, we discuss challenges related to compositional data processing together with some principled methods of CDA. Section III describes an application of such methods for spatial interpolation of compositional data.

II. COMPOSITIONAL DATA

It is common to express elemental concentrations as percentages, so that the sum of the concentrations is 100%. Data expressed as part of a whole are called compositional data, and the study of such data is a relatively new part of statistics.[22]

Received: September 17, 2014
Revised: November 26, 2014
Published: December 29, 2014

DOI: 10.1021/co5001458
ACS Comb. Sci. 2015, 17, 130−136

The core concepts are presented here through illustrative examples from solid state chemistry.

II.A. Closure Effects. *Induced Correlations.* The traditional way to describe the pattern of variability of data is through the estimates of the raw mean, covariance, and correlation matrices. Individual components of compositional data are not free to vary independently: if the proportion of one component decreases, the proportion of one or more other components must increase, thus leading an artificial correlation that is, in fact, caused by the constant sum constraint. Indeed, the closure, or in other words the constant sum constraint, affects correlation between variables.[24] Consider for example a set of N-part compositions that can be treated as a $M \times N$ matrix $\boldsymbol{W}$ where N is the number of elements in the composition, and M is the number of measurements, or samples, with the component sum $\Sigma_{k=1}^{N} w_{ik} = 1$, where $i = 1, ..., M$. Let us denote $Y_k = \{w_{ik}\}$, $i = 1, ..., M$, to be the k-th column of the matrix $\boldsymbol{W}$. Since

$$\mathrm{cov}(Y_k, \sum_{j=1}^{N} Y_j) = 0$$

we have

$$\sum \mathrm{cov}(Y_k, Y_j) = -\mathrm{var}(Y_k),\ j \neq k \qquad (1)$$

so the sum of the covariances of any variable is negative. Thus, each variable must be negatively correlated with at least one other variable and, in general, there is a strong bias toward negative correlation between variables of (relatively) large variance. One of the important consequences of closure for materials science is that usual correlation analysis can produce misleading associations between elemental concentrations.

Illustrative Example. As an example, consider a set of M materials each containing N elements for which we would like to ascertain if there is correlation in the concentration of element 1 with respect to the other elements. Figure 1a shows correlation coefficients for a synthetic data set created by generating random quantities of four elements ($N = 4$) from normal distributions. Because of the randomness, the element-pairwise correlation over the M = 400 materials is negligible when considering the quantities of the elements, which is non-normalized data. Measurements of the (normalized) composition of each material produce the $M \times N$ closed data set $\{w_{ik}\}$. Using this simulated data, the Pearson correlation coefficient ${}^{4}C_{k,l}$ of the concentration vectors $\boldsymbol{Y}_k$ and $\boldsymbol{Y}_l$ (elements k and l) can be calculated, where the superscript 4 indicates the dimension of the composition space ($N = 4$). Consider an extension of this example in which the concentration of the fourth element cannot be measured so instead composition measurements are made in the $N = 3$ space and correlations ${}^{3}C_{k,l}$ are calculated, and a similar exercise can by performed for $N = 2$.

The values of ${}^{N}C_{k,l}$ plotted in Figure 1b demonstrate some limitations of the usual statistics. Indeed, the correlation coefficients are skewed toward negative values because of the normalization-induced correlation, as indicated by eq 1. In fact, for the $N = 2$ case, the correlation coefficient is -1 because due to the normalization $x_{i,2} = 1 - x_{i,1}$. In other words, correlation structure of a composition cannot be used to interpret correlations among the measured elemental concentrations and vice versa. It should be mentioned that other distance-based statistics like means, variances and standard deviations, as well as tasks such as clustering and multidimensional scaling exhibit similar limitations when applied to compositional data.

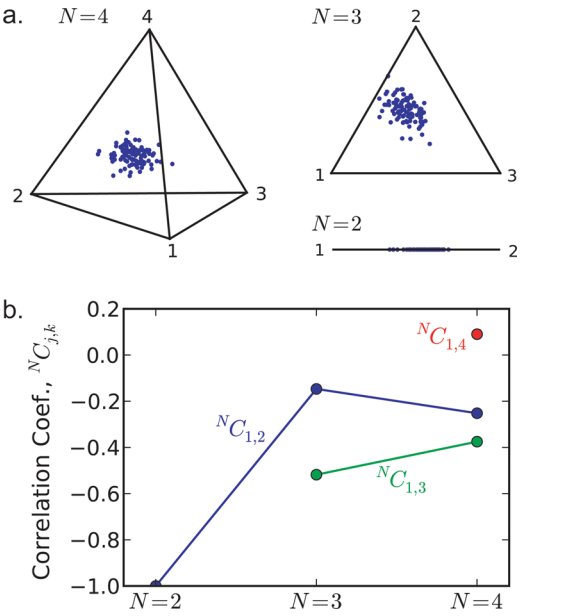


Figure 1. (a) Set of 100 compositions generated from normal distributions of element quantities with normalizations corresponding to the quaternary ($N = 4$), ternary ($N = 3$), and binary ($N = 2$) composition spaces. (b.) The correlation of the concentration of element 1 with each other element; the magnitude change demonstrates induced correlation; the variation with respect to N shows subcompositional incoherence.

Subcompositional Coherence. An n-part composition $(x_1, x_2, ..., x_n)$ with $\Sigma_{i=1}^{n} x_i = 1$ is called a subcomposition of an m-part composition $(x_1, x_2, ..., x_m)$ with $\Sigma_{i=1}^{m} x_i = 1$, if $m > n$ and $(x_1, ..., x_n)$ is a subset of the elements $(x_1, ..., x_m)$. A consequence of the constant sum constraint for compositional data is that subcompositions may not reflect the variations present in the parent data, and as a result covariance of elements may change substantially between different subsets of the parent data set. Every composition is a sub- or a parent- composition depending on the objective of an experiment or the goal of data analysis. An experimentalist or a data analyst may not be able to take into account all elements (some elements may not be accessible), or may disregard some of the available elements if they are not pertinent to the objective. The following principle of subcompositional coherence is an important concept of compositional analysis: any compositional data analysis should be done in a way that produces the same results in a subcomposition, regardless of whether we analyzed only that subcomposition or a parent composition.[18] Subcompositional incoherence of Pearson correlation coefficient is demonstrated in Figure 1, where for a given pair of elements, ${}^{N}C_{k,l}$ varies with the order N of the composition space.

These effects of closure on statistical analysis of compositional data, induced correlations and subcompositional incoherence, make traditional statistical methods invalid, and artificial correlation obtained by applying such techniques may lead to false scientific discoveries and incorrect predictions. Moreover, methods that are based on a correlation matrix of observations, such as factor analysis, principal component analysis (PCA), cluster analysis,[13,14] kriging interpolation[25] to name just a few, would lead to inaccurate, warped results. Thus, correlation analysis, and multivariate statistical analysis in

DOI: 10.1021/co5001458
ACS Comb. Sci. 2015, 17, 130−136

general, of compositional data require special techniques in order to avoid producing false results.

II.B. Principled Analysis of Compositional Data. *Sample Space.* The fundamental building block of statistical analysis is the probabilistic model. A well-defined sample space is one of the basic elements in a probabilistic model. The constraint of constant sum does not allow the components of a composition to vary from $-\infty$ to ∞. Because of the constraint, an N-element composition is confined to a restricted part of the Euclidean space called the simplex $S^N = \{\boldsymbol{x};\ x_k \geq 0,\ \Sigma^N_{k=1}\ x_k = 1\}$.[18] All standard statistical methods assume that the sample space is the entire Euclidean space, while compositional data clearly do not satisfy this assumption. In order to deal with the closure effects described in the previous section, an approach based on a family of transformations, the so-called logratio transformations, has been introduced.[18] These transformations based on logarithm of ratios of compositions map the components of the compositions onto a Euclidean space, thus enabling one to apply classical statistical methods. In what follows, we briefly describe a few key concepts of such analysis.[18,21,22] The so-called *alr* transform is defined for a given N-element composition $\boldsymbol{x}$ as an $(N-1)$-element vector $\boldsymbol{z}$ with the following components:

$$\mathbf{z} = alr(\mathbf{x}) = (\ln(x_1/x_N), \ldots, \ln(x_{N-1}/x_N)) \tag{2}$$

where one of the composition components is chosen as common divisor. This logratio transform is invertible since there is a one-to-one correspondence between any N-part composition $\boldsymbol{x}$ and its logratio vector $\boldsymbol{z}$. This means that any statement about the components of a composition can be expressed in terms of logratios and vice versa. By defining the sum

$$s_i = \sum_{j\neq i} \exp(z_j - z_i) = \sum_{j\neq i} x_j/x_i$$

the transformation from log ratio to composition coordinates is given by

$$x_i = 1/(s_i + 1) \tag{3}$$

Because *alr* depends on the choice of x_N, this transform is not employed in our calculations and a more suitable transform is discussed below. In section III we utilize eqs 2 and 3 only to illustrate the results of spatial compositional interpolation.

To build a vector space structure on the simplex the following operations were introduced by Aitchison. The *closure* operation C is defined as

$$\mathbf{x} = C[u_1, \ldots, u_N] = (u_1/(u_1 + \ldots + u_N), \ldots, u_N/(u_1 + \ldots + u_N));\ u_i \geq 0,\ \mathbf{x} \in S^N \subset R^{N-1}$$

where u_i represent the raw data such as element quantities. *Perturbation* $\oplus$ is an equivalent of addition in the Euclidean space and defined as

$$\mathbf{w} = \mathbf{x} \oplus \mathbf{y} = C[x_1y_1, x_2y_2, \ldots, x_Ny_N];\ \mathbf{w}, \mathbf{x}, \mathbf{y} \in S^N \subset R^{N-1} \tag{4}$$

Powering $\odot$ is an equivalent of multiplication a vector by a scalar and defined as

$$\mathbf{w} = a \odot \mathbf{x} = C[x_1^a, x_2^a, \ldots, x_N^a];\ \mathbf{x} \in S^N,\ a \in R$$

Aitchison inner product replaces the Euclidean inner product and defined as

$$\langle \mathbf{x}, \mathbf{y}\rangle_A = 1/N \sum_{i=1}^{N} \sum_{j>i}^{N} \ln(x_i/x_j)\ln(y_i/y_j);\quad \mathbf{x}, \mathbf{y} \in S^N \subset R^{N-1} \tag{5}$$

Thus, the norm of a vector, or its simplicial length, is $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x},\boldsymbol{y}\rangle_A}$. This enables one to compute distances between compositional vectors, projections of compositional vectors, etc.

The Aitchison distance is defined as

$$d_A(\mathbf{x}, \mathbf{y}) = \{1/N \sum_{i=1}^{N} \sum_{j>i}^{N} [\ln(x_i/x_j) - \ln(y_i/y_j)]^2\}^{1/2};\quad \mathbf{x}, \mathbf{y} \in S^N \subset R^{N-1} \tag{6}$$

Establishing a metric vector space structure in the simplex and utilizing orthonormal bases facilitates application of complex statistical methods to analysis of compositional data. The so-called isometric logratio (ilr) transform has important conceptual advantages and enables one to use balances, a particular form of ilr coordinates in an orthonormal basis. A balance is defined as

$$b_{pq} = ([pq/(p + q)]^{1/2})\ln(g(x_p)/g(x_q)) \tag{7}$$

where $g(\cdot)$ is the geometric mean of the argument, x_p is the group with p parts and x_q is the group of q parts which are obtained by sequential binary partition (see works[21,23] and references there). However, there is no obvious "optimal" basis, and the compositional biplot approach should be used to find one.[18,23] For an analysis to be subcompositionally coherent, it suffices to define variables using the ratios of the composition values. The quantities x_1/x_2 and $ln(x_1/x_2)$ are invariant under changes of the composition order as they quantify the relative magnitudes of elemental concentration rather than their absolute values, though the interpretation of the results in terms of the original variables is not always trivial. To study correlation structure of compositions Aitchison introduced a variation matrix $\boldsymbol{T} = \{\ \tau_{ij}\ \}$ of dimensions $N \times N$, with the elements

$$\tau_{ij} = \mathrm{var}[\ln(Y_i/Y_j)] \tag{8}$$

When τ_{ij} are large, there is no proportionality between the corresponding elements. If, however, the elements i and j are exactly proportional then $\tau_{ij} = 0$. The scale of these variations can be determined by introducing total variance as a normalized sum of the variances of all logratios

$$V_{\mathrm{tot}} = 1/(2N) \sum_{i=1}^{N} \sum_{j=1}^{N} \tau_{ij} \tag{8a}$$

The variation matrix $\boldsymbol{T}$ (eqs 8, 8a) is instrumental in the analysis of associations between elemental concentrations in compositions. Such analysis will be discussed in greater detail in our forthcoming paper dedicated to covariance structures of screening libraries. In what follows we apply balances (eq 7) to spatial interpolation of compositional data.

III. INTERPOLATION OF COMPOSITIONAL DATA

III.A. Interpolation of Compositional Data and Materials Science: Sputtering. Section II highlighted the importance of using logratio variables for statistical analysis and in this section we use spatial interpolation of composition measurements, a standard operation in combinatorial research, to demonstrate how behavior of logratio variables differs from

DOI: 10.1021/co5001458
ACS Comb. Sci. 2015, 17, 130−136

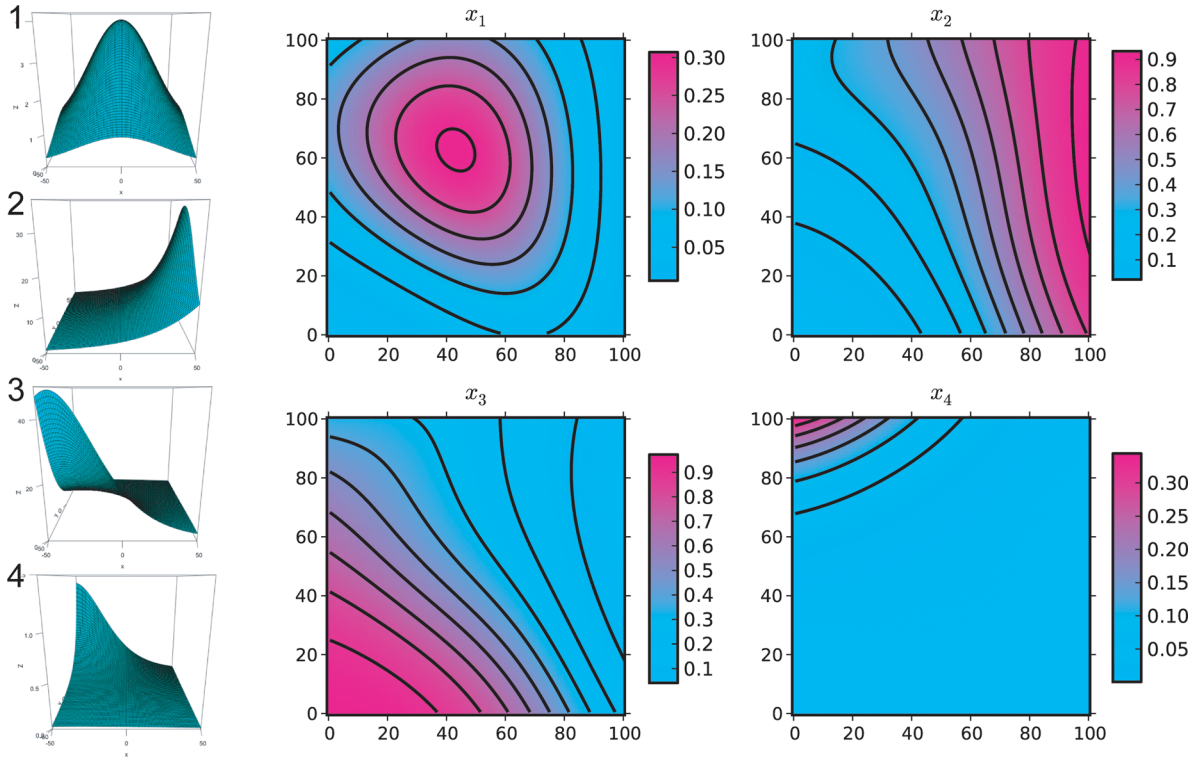


Figure 2. (left) Modeled profiles of deposition rate as a function of position on a 100 mm square substrate are shown for 4 elemental deposition sources. (right) The corresponding elemental compositions are shown for each element as a colored contour plot.

that of raw compositional variables. To create a synthetic data set, we employ a common combinatorial synthesis technique, multisource cosputtering of a composition spread thin film. Combinatorial sputtering is commonly used for synthesis of binary, ternary, and quaternary thin film libraries.[26] In general, spatial variation of the composition of each of the components is dependent on several parameters, such as sputtering target, sputtering gas, reactive gases, deposition geometry, and resputtering phenomenon.[27] The continuous nature of the thin films and ability to measure film properties with high spatial resolution require accurate modeling of spatio-compositional structure to observe meaningful composition-property studies. The compositions of a continuous film can be measured using appropriate elemental analysis techniques, and to enable high throughput and measurement efficiency, it is desirable to measure compositions on a sparse grid and rely on spatial interpolation to obtain the full composition map. Addressing the complex nature of compositional variation as a function of the spatial coordinates requires compositional statistical methods.

Composition libraries containing 4 components can be deposited onto a substrate using 1 central face-to-face sputtering source surrounded by 3 tilted sources. The deposition rate from each of the 4 elemental sources in this configuration was calculated using a standard sputtering model, yielding the deposition rate profiles in Figure 2a. To provide some variability in the deposition rate profiles, different tilt angles of the sources and overall deposition rates were used in the model, but we note that these details are inconsequential for the purposes of generating a demonstration data set. Using the deposition rate profiles, the elemental compositions were calculated over a 100 mm × 100 mm square substrate area, yielding the composition coverage shown in Figure 2.

Figure 2 provides the spatial map of the 4 compositional variables x_i, and in this paper we will assume that accurate, noise-free compositions measurements are made on a set of 25 substrate positions chosen as a 5 × 5 square grid with 25 mm spacing. An appropriate spatial interpolation method should at least guarantee that the non-negativity and constant-sum constraints are satisfied. In fact, among the conventional unconstrained interpolation techniques, linear interpolation satisfies these requirements. However, usual straightforward approaches, even if they satisfy the constraints, interpolate each component x_i independently, thus ignoring the inner relationships between the compositional elements. Since our end goal is to enable the analysis of covariance structure of compositions without the artifacts of induced correlation and subcompositional incoherence, an approach that leads to accurate logratio values employed by the simplicial distance (eq 6) and compositional covariance matrix $\boldsymbol{T}$ (eq 8) is required. To achieve this, we utilize a broadly applicable and highly versatile technique based on kriging.[25] The kriging-based interpolation was computed with **R** language and environment for statistical computing[28] by applying the R-package "compositions" developed by van der Boogaart and Tolosana-Delgado.[29,30] The method exploits codependences in the composition and takes into account the spatial covariance structure by modeling the set of variograms for all possible pairwise balances (eq 7). It takes into account various effects and parameters including the nugget effect, the choice of exponential and spherical variograms which parameters we chose to be 62.5 and 162.0, respectively. Since this interpolation technique is specialized for compositional data, we refer to it as "compositional interpolation" and represent the result of compositional interpolation of the 25 z_i measurements as CompInterp(z_i). To attain an analogous result using traditional linear

DOI: 10.1021/co5001458
ACS Comb. Sci. 2015, 17, 130−136

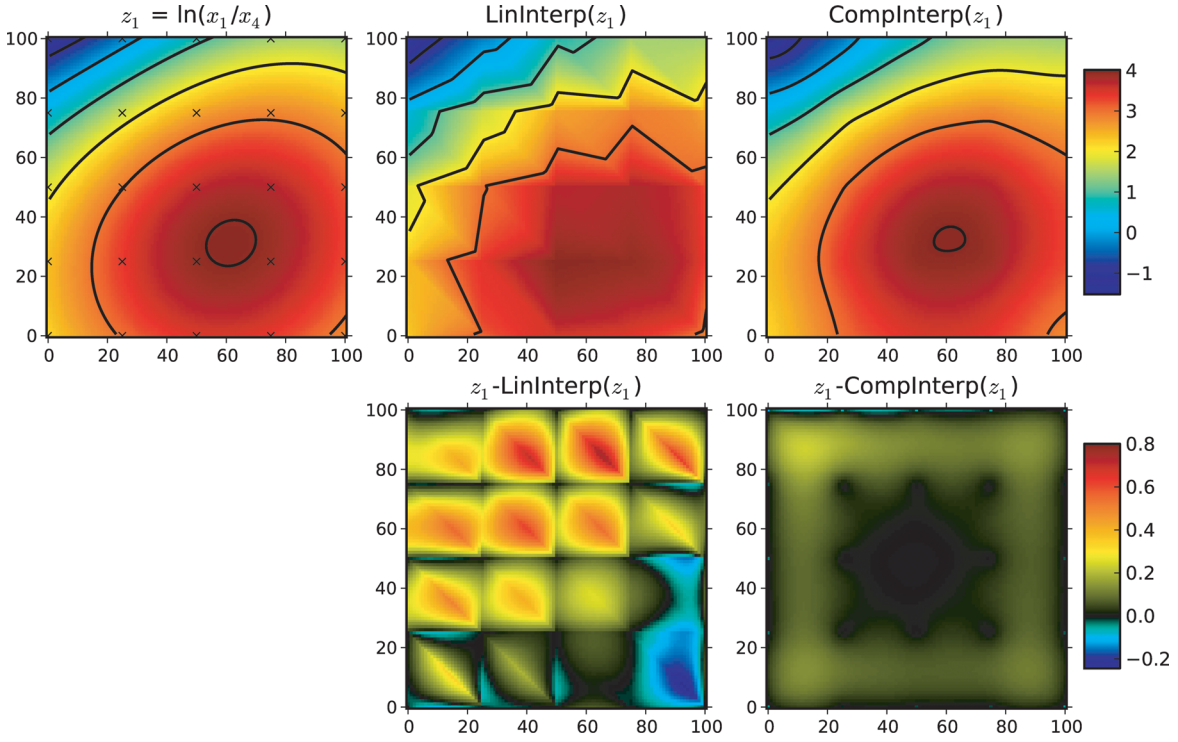


Figure 3. Interpolation of a 4-element composition. z_1: logratio of compositions x_1 and x_4 calculated from Figure 2 with the 25 sampling points marked by "×". LinInterp(z_1): log ratio of the linearly interpolation of x_1 and x_4. CompInterp(z_1): log ratio of the compositionally interpolated z_1; the difference between the model data and its linear interpolation; the difference between the model data and its compositional interpolation.

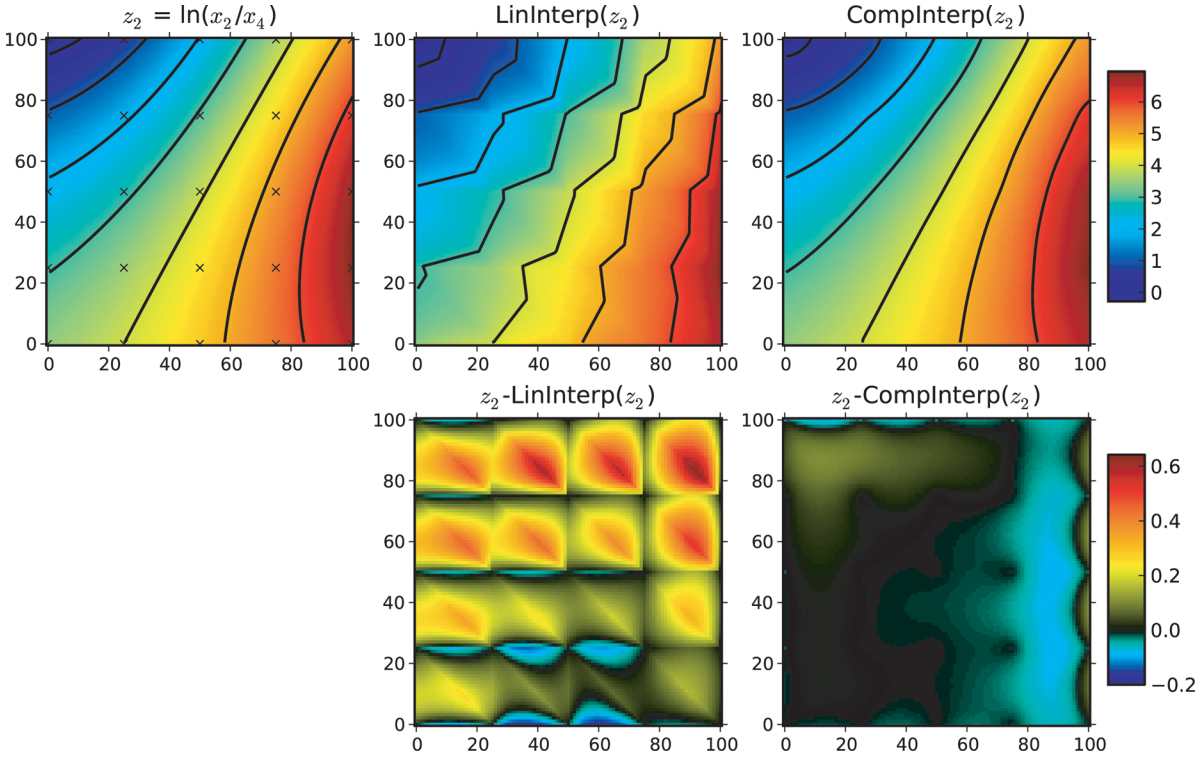


Figure 4. Results of interpolation for z_2. See Figure 3 for detailed description.

interpolation, x_i and x_N can be independently interpolated followed by calculating z_i (eq 2), resulting in a spatial map of z_i referred to as LinInterp(z_i). The results of compositional and linear interpolations and their comparisons with the "perfect" data calculated from the model compositions of Figure 2 are shown in Figures 3 and 4. By definition, both interpolation methods produce exact values at each of the 25 locations in the sampling grid. The assessment of the performance of a given interpolation is thus performed by evaluating the absolute magnitude and pattern of interpolation error in the regions

DOI: 10.1021/co5001458
ACS Comb. Sci. 2015, 17, 130−136

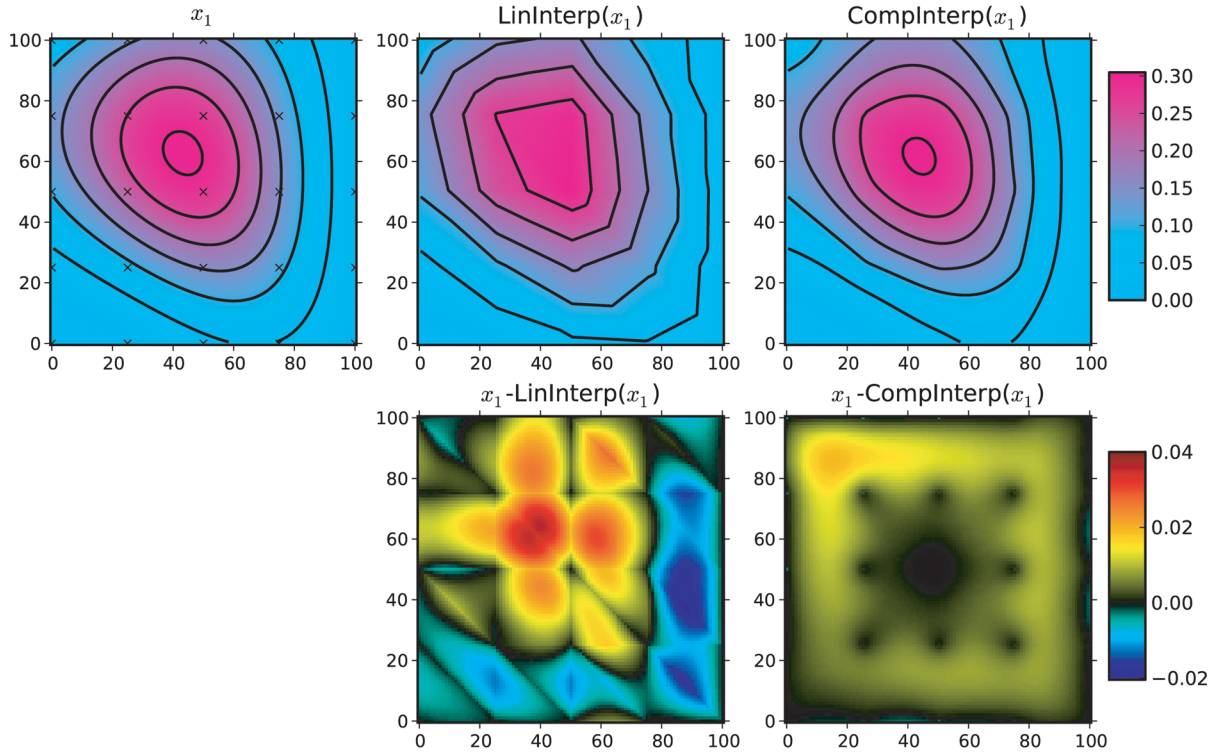


Figure 5. Interpolation of a 4-element composition. x_1: Model data taken from Figure 2 with the 25 sampling points marked by "×". LinInterp (x_1): Linear interpolation of these 25 samples of x_1. CompInterp(x_1): Result of applying eq 3 to the set of compositionally interpolated logratios; the difference between the model data and its linear interpolation; the difference between the model data and its compositional interpolation.

between the sampling points (the grid). Compared to linear interpolation, the compositional interpolation provides more accurate results and the discrepancy varies smoothly over the entire interpolation region. It is important to note that artificial "patchiness" of the linear interpolation would distort simplicial distances (eq 6) between different compositions. Such distortions would lead to artificial associations in the analysis of correlational structure of compositions and, more generally, to erroneous results in all calculations that involve distances, e.g. mean and standard deviation.

Provided the compositional interpolation of each z_i (eq 2), a spatial map of x_i can be calculated using the inverse transformation (eq 3). We refer to this spatial map as CompInterp(x_i). This result is compared to the linear interpolation of x_i LinInterp(x_i) in Figure 5. It can be seen, that linear interpolation and compositional interpolation both perform well for the interpolation of x_1, with maximum deviation of linear interpolation being 0.04 and maximum deviation of compositional interpolation being 0.02. The compositional interpolation produces a smoother result due to the inherent consideration of the spatial covariance structure of all pairwise balances. While it is worth noting that many experimental measurements of composition have associated uncertainties on par with these deviations, this demonstration data set was chosen to illustrate the difference in behavior of compositional and linear interpolation. This data set lacked experimental noise and contained smooth, very slowly varying functions (Figure 2.). Kriging assumes that the observed values are a realization of a stochastic process, so the quantitative advantages of compositional interpolation based on kriging should become more pronounced as variation of the composition variables increases. It is worth noting, that there are other interpolation methods that preserve the non-negativity and constant-sum constraints such as local sample mean, inverse distance interpolation, and triangulation (since the weights they use range from 0 to 1, and sum to unity). However, unlike the approach developed in refs 29 and 30 and utilized here, those methods do not take into account the spatial covariance structure which may be critical for statistical analysis.

From subcomposition coherence and closure effects of Figure 1 to the shortcomings of linear interpolation in Figures 2−5, we demonstrate fundamental issues of applying Euclidean-based methods to compositional data. In the example of spatial interpolation of compositions, the data vary smoothly as a function of position, allowing linear interpolation to provide reasonable results, which are still improved by the use of simplex-based methods. The shortcomings of Euclidean-based techniques strongly depend on the sparseness of the measurements and nonlinearity of the measured signals, but on a fundamental level the use of simplex-based methods generally provides a more accurate treatment of compositional data. As combinatorial materials science continues to expand into high order compositions spaces, the prudent application of statistical methods developed specifically for CDA will be required to enable accurate data mining.

IV. CONCLUSIONS

Probably no other field has so much of its data intrinsically expressed as percentages as do chemistry and materials science. In this Research Article, we brought the attention of materials scientists to the importance of CDA. We first demonstrated that Euclidean-based correlation structure should not be used to interpret associations among measured elemental concentrations. By using simulated data we presented and illustrated induced correlations and subcompositional incoherence of the

DOI: 10.1021/co5001458
ACS Comb. Sci. 2015, 17, 130−136

Pearson correlation coefficient. These effects are caused by the constant sum constraint that restricts the sampling space to a simplex instead of the usual Euclidean space. Since statistical measures such as mean, standard deviation, etc., are defined for the Euclidean space, traditional correlation studies, multivariate analysis, and hypothesis testing may lead to erroneous dependencies and incorrect inferences when applied to compositional data. These issues demonstrate that prior to applying usual statistical methods data should be transformed to remove the constant sum constraint. Logratio transforms remove the data-sum constraint by mapping the components of the compositions into a Euclidean space, thus enabling one to apply classical statistical methods. Moreover, a metric vector space structure can be introduced in the simplex (via the simplicial metric based on log ratios), thus enabling meaningful statistical analysis of compositional data. We applied logratio analysis to interpolation of simulated composition data. Comparison of a consistent compositional interpolation based on balances with traditional linear approach revealed discrepancies between their results that are crucial for correct statistical analysis of composition-property relationships. Altogether these results demonstrate the importance of using adequate, mathematically consistent approaches to compositional data, particularly in high-order composition spaces.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: mzp@caltech.edu.

*E-mail: gregoire@caltech.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon work performed by the Joint Center for Artificial Photosynthesis, a DOE Energy Innovation Hub, supported through the Office of Science of the U.S. Department of Energy under Award Number DE-SC000499.

REFERENCES

(1) Fan, J.; Han, F.; Liu, H. Challenges in Big Data. *Natl. Sci. Rev.* **2014**, 1−22.

(2) Committee on the Analysis of Massive Data, *Frontiers in Massive Data Analysis*; The National Academies Press: Washington, DC, 2013.

(3) Pesenson, M. Multiscale Analysis—Modeling, Data, Networks, and Nonlinear Dynamics. In *Multiscale Analysis and Nonlinear Dynamics*, Wiley Reviews of Nonlinear Dynamics and Complexity; Pesenson, M., Ed.; Wiley-VCH: Weinheim, Germany, 2013; pp 1−19.

(4) Data-Enabled Science in the Mathematical and Physical Sciences, A workshop funded by the National Science Foundation, 2010. https://www.nsf.gov/mps/dms/documents/Data-EnabledScience.pdf.

(5) Leek, J.; Scharpf, R.; Bravo, H. Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. *Nat. Rev.* **2010**, *1*, 733−739.

(6) Rajan, K. Combinatorial Materials Sciences: Experimental Strategies for Accelerated Knowledge Discovery. *Annu. Rev. Mater. Res.* **2008**, *38*, 299−322.

(7) Gregoire, J. M.; Van Campen, D. G.; Miller, C.; Jones, R.; Suram, S.; Mehta, A. High Throughput Synchrotron X-ray Diffraction for Combinatorial Phase Mapping. *J. Synchrotron Radiat.* **2014**, DOI: 10.1107/S1600577514016488.

(8) Gregoire, J. M.; Xiang, C.; Liu, X.; Marcin, M.; Jin, J. Scanning Droplet Cell for High Throughput Electrochemical and Photoelectrochemical Measurements. *Rev. Sci. Instrum.* **2013**, *84*, No. 024102.

(9) Gregoire, J. M.; Xiang, C.; Mitrovic, S.; Liu, X.; Marcin, M.; Cornell, E. W.; Fan, J.; Jin, J. Combined Catalysis and Optical Screening for High Throughput Discovery of Solar Fuels Catalysts. *J. Electrochem. Soc.* **2013**, *160* (4), F337−F342.

(10) Maier, W. F.; Stowe, K.; Sieg, S. Combinatorial and High-Throughput Materials Science. *Angew. Chem., Int. Ed* **2007**, *46*, 6016−6067.

(11) Jiang, C.; Wang, R.; Parkinson, B. A. Combinatorial Approach to Improve Photoelectrodes Based on BiVO4. *ACS Comb. Sci.* **2013**, *15* (12), 639−645.

(12) Park, S. H.; Choi, C. H.; Koh, J. K.; Pak, C.; Jin, S.; Woo, S. I. Combinatorial High-Throughput Screening for Highly Active Pd−Ir−Ce Based Ternary Catalysts in Electrochemical Oxygen Reduction Reaction. *ACS Comb. Sci.* **2013**, *15* (11), 572−579.

(13) Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2009.

(14) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, 2013.

(15) Green, M. L.; Takeuchi, I.; Hattrick-Simpers, J. R. Applications of High Throughput (Combinatorial) Methodologies to Electronic, Magnetic, Optical, and Energy-Related Materials. *J. Appl. Phys.* **2013**, *113*, No. 231101.

(16) Haber, J. A.; Cai, Y.; Jung, S.; Xiang, C.; Mitrovic, S.; Jin, J.; Bell, A. T.; Gregoire, J. M. Discovering Ce-Rich Oxygen Evolution Catalysts, from High Throughput Screening to Water Electrolysis. *Energy Environ. Sci.* **2014**, *7* (2), 682−688.

(17) Aitchison, J. The statistical analysis of compositional data (with discussion). *J. R. Stat. Soc., Ser. B* **1982**, *44*, 139−177.

(18) Aitchison, J. *The Statistical Analysis of Compositional Data*, Monographs on Statistics and Applied Probability; Chapman & Hall Ltd., The Blackburn Press: London, 1986 (2d ed. with additional materials, 2003).

(19) Pawlowsky-Glahn, V.; Egozcue, J. Compositional Data and Their Analysis: an Introduction. *Geol. Soc. London Spec. Publ.* **2006**, *264*, 1−10.

(20) Pawlowsky-Glahn, V., Buccianti, A., Eds. *Compositional Data Analysis: Theory and Applications*; Wiley: Chichester, U.K., 2011.

(21) Egozcue, J. J; Pawlowsky-Glahn, V. Basic Concepts and Procedures. In *Compositional Data Analysis: Theory and Applications*; Pawlowsky-Glahn, V., Buccianti, A., Eds.; Wiley: Chichester, U.K., 2011; Chapter 2, pp 12−28.

(22) Bacon-Shone, J. A Short History of Compositional Data Analysis. In *Compositional Data Analysis: Theory and Applications*; Pawlowsky-Glahn, V., Buccianti, A., Eds.; Wiley: Chichester, U.K., 2011; Chapter 1, pp 3−11.

(23) Egozcue, J. J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barcelo-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **2003**, *35* (3), 279−300.

(24) Chayes, F. *Ratio Correlation*; Chicago University Press: Chicago, 1971.

(25) Chilès, J. P.; Delfiner, P. *Geostatistics — Modeling Spatial Uncertainty*, 2nd ed., Wiley Series in Probability and Statistics; Wiley: New York, 2012.

(26) Gregoire, J. M.; van Dover, R. B.; Jin, J.; DiSalvo, F. J.; Abruña, H. D. Getter sputtering system for high-throughput fabrication of composition spreads. *Rev. Sci. Instrum.* **2007**, *78* (7), No. 072212.

(27) Gregoire, J.; Lobovsky, M.; Heinz, M.; DiSalvo, F.; van Dover, R. Resputtering phenomena and determination of composition in codeposited films. *Phys. Rev. B* **2007**, *76* (19), No. 195437.

(28) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2004.

(29) Tolosana-Delgado, R.; van den Boogaart, K.; Pawlowsky-Glahn, V. Geostatistics for Compositions. In *Compositional Data Analysis: Theory and Applications*; Pawlowsky-Glahn, V., Buccianti, A., Eds.; Wiley: Chichester, U.K., 2011; Chapter 6, pp 73−86.

(30) van den Boogaart, K.; Tolosana-Delgado, R. *Analyzing Compositional Data with R*, Use R! Series; Springer: Berlin, 2013.

DOI: 10.1021/co5001458
ACS Comb. Sci. 2015, 17, 130−136