Interval Compositional Data: Problems and Possibilities

Ondřej Pavlačka and Karel Hron

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University Olomouc,
17. listopadu 12, 771 46 Olomouc, Czech Republic
ondrej.pavlacka@upol.cz, hronk@seznam.cz

Abstract. In statistics, compositional data are defined as multivariate observations that quantitatively describe contributions of parts on a whole, carrying exclusively relative information. As a consequence, compositions can be represented as proportions or percentages without loss of information (contained in ratios between parts). Nevertheless, in the practice parts of compositional data are frequently formed by intervals; for example, concentrations of chemical elements are provided not as exact numbers, but rather in an interval range. Intuitively, a natural question arises, whether the relative information is preserved, when the original compositional data with interval-valued parts are represented in proportions. Namely, from the arithmetic properties of interval data, normalizing of intervals does not simply follow the case of real values, but a special procedure according to constrained interval arithmetic is needed. The aim of the contribution is to discuss possibilities of representing the interval compositional data in proportions.

Keywords: compositional data, Aitchison geometry, interval arithmetic, descriptive statistics.

1 Introduction

The concept of compositional data frequently occurs in many applications, covering such situations, where not the absolute values of variables, but rather relative information they contain is of primary interest [1,8]. Typical examples are formed by concentrations of chemical elements in a rock (in mg/kg), proportional representation of political parties resulting from elections, but also household expenditures on various costs (like foodstuff, housing, clothing, culture, etc.), when the relative structure of costs is to be analyzed. Consequently, compositional data are popularly represented as multivariate observations with a constant sum constraint (like in proportions or percentages). Nevertheless, the above examples clearly imply that compositions themselves are not necessarily induced by any such constraint (household expenditures can be represented both in the original units, like EUR or USD, and in proportions, the ratios between parts remain the same).

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing,*
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_4

In the practice also such situations occur that compositional parts are represented by intervals rather than by precise values. One natural source of such data comes from aggregation of information over individuals in symbolic data analysis [2] in order to obtain representants of specified sets of individuals (like household expenditures in different parts of a certain region) that capture variability of the aggregation process. Another source of interval compositional data is formed by measurement process itself that leads naturally to unprecise values. Such situations arise usually in analytical chemistry or geochemistry, e.g. due rounding effects of values near to detection limit. In contrast to symbolic data analysis case, here the interval values of variables are often combined with precise ones what makes the use of procedures based on logarithmic transformation [3] conceptually not possible.

To illustrate the methodology, presented further, we use a small real-world data set, obtained from The National Institute of Public Health of Czech Republic (2011), where chemical composition of seven popular mineral waters was analyzed. For our purposes, just four chemical elements were chosen (calcium, sodium, magnesium and potassium), and the resulting values (measured in mg/l and already collected in form of interval values with lower and upper boundary) are presented in Table 1 (the mineral waters are listed with their original Czech names). We can observe that the interval values in the data set occur in all variables, except for the first mineral water (called Hanácká kyselka) and potassium in case of Magnesia.

Table 1. Interval concentrations of chemical elements (in mg/l) in Czech mineral waters

Mineral waters	Calcium		Sodium		Magnesium		Potasium	
Hanácká kyselka	275.0	275.0	68.0	68.0	251.0	251.0	17.7	17.7
Korunní	78.3	86.5	29.5	30.9	98.0	111.1	23.0	25.5
Magnesia	35.3	41.3	179.0	200.0	4.3	6.8	1.4	1.4
Mattoni	87.6	88.6	24.8	24.9	71.9	79.8	18.0	19.0
Ondrášovka	192.0	199.4	19.4	19.8	29.2	30.9	1.4	1.6
Poděbradka	142.2	145.5	45.4	49.3	344.0	360.0	47.0	49.8
Dobrá voda	8.7	10.7	9.5	9.7	9.5	10.0	9.3	9.4

Obviously, also for interval compositional data, analogously as for compositions with precise values, not absolute values of single element concentrations, but rather their relative contributions to the overall chemical composition of mineral waters is of primary interest. In other words, also here the ratios between (interval) compositional parts form the source of relevant information. Nevertheless, due to limitations of interval arithmetics, treatment of interval compositional data is more complex than in the standard (precise) case. The aim of this contribution is to draw up possible problems and challenges, related to geometrical properties and subsequent statistical analysis of interval compositional data, that might lead to a concise methodology in the future.

The paper is organized as follows. In the next section, basics of interval arithmetics are refreshed, with a focus on positive interval data (forming the compositional parts). Section 3 is devoted to the problem of forming the ratios of compositional data. In Section 4, problems related to proportional representation of interval compositional data are analyzed. Consequently, implications for descriptive statistics of interval compositions are briefly discussed. Finally, possibilities of further development are collected in the last section.

2 Computing with Intervals

Before we proceed to introduce interval compositional data and their geometrical concepts, let us briefly refresh basic possibilities of computing with intervals. Due to definition of compositional data, it is sufficient to restrict the general case for positive intervals only. We will distinguish two cases: First, we will assume that the input intervals are independent, i.e. all combinations of values belonging to the intervals are admissible. Second, we will consider interactive input intervals, where the set of all admissible combinations of values is given.

Let $I_1, \ldots, I_n$ be independent intervals and $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. Then

$$f(I_1, \ldots, I_n) = \left[\underline{y}, \overline{y}\right],$$

where

$$\begin{aligned}\underline{y} &= \min\{f(x_1, \ldots, x_n) \mid x_i \in I_i,\ i = 1, \ldots, n\},\\ \overline{y} &= \max\{f(x_1, \ldots, x_n) \mid x_i \in I_i,\ i = 1, \ldots, n\}.\end{aligned}$$

If f stands for one of the basic arithmetic operations, we get the well-known concept of *standard interval arithmetic*. Let $[a, b]$, $0 < a \leq b$, and $[c, d]$, $0 < c \leq d$, be independent intervals. Then arithmetic operations are extended as follows:

$$\begin{aligned}[a, b] + [c, d] &= [a + c, b + d],\\ [a, b] - [c, d] &= [a - d, b - c],\\ [a, b] \cdot [c, d] &= [a \cdot c, b \cdot d],\\ \frac{[a, b]}{[c, d]} &= \left[\frac{a}{d}, \frac{b}{c}\right].\end{aligned}$$

However, the above concept cannot be applied in the case when it is given a constraint set $Q \subset \mathbb{R}^n$ that represents all admissible combinations of the values of $x_1, \ldots, x_n$ (see e.g. [5]). If $Q \cap (I_1 \times \ldots \times I_n)$ is a nonempty convex set, then

$$f(I_1, \ldots, I_n; Q) = \left[\underline{y}, \overline{y}\right],$$

where

$$\begin{aligned}\underline{y} &= \min\{f(x_1, \ldots, x_n) \mid x_i \in I_i,\ i = 1, \ldots, n,\ (x_1, \ldots, x_n) \in Q\}\\ \overline{y} &= \max\{f(x_1, \ldots, x_n) \mid x_i \in I_i,\ i = 1, \ldots, n,\ (x_1, \ldots, x_n) \in Q\}.\end{aligned}$$

In our case, the role of a constraint set Q will play, for instance, the set representing proportional representation of interval compositional data.

3 Ratios of Compositional Parts

Following the Introduction section, a sample of D-part compositional data are positive vectors $\mathbf{x}_i := (x_{i1}, \ldots, x_{iD})$, $i = 1, \ldots, n$, that describe quantitatively contributions of parts on a whole, carrying exclusively relative information. This means that the relevant information is expressed by the ratios $r^i_{jk} := x_{ij}/x_{ik}$, $j, k = 1, \ldots, D$, $j \neq k$.

Now, let us consider the case of interval compositional data

$$\mathbf{X}_i := \left([\underline{x}_{i1}, \overline{x}_{i1}], \ldots, [\underline{x}_{iD}, \overline{x}_{iD}]\right), \quad i = 1, \ldots, n,$$

where

$$0 < \underline{x}_{ij} \leq \overline{x}_{ij}, \ j = 1, \ldots, D.$$

For the sake of simplicity, let us assume further that $[\underline{x}_{ij}, \overline{x}_{ij}]$ and $[\underline{x}_{lk}, \overline{x}_{lk}]$ are independent intervals for any $i, l = 1, \ldots, n$, and $j, k = 1, \ldots, D$, $j \neq k$.

According to the assumption, we can apply the concept of standard interval arithmetics for computing the ratios between the compositional parts:

$$R^i_{jk} := \frac{[\underline{x}_{ij}, \overline{x}_{ij}]}{[\underline{x}_{ik}, \overline{x}_{ik}]} = \left[\frac{\underline{x}_{ij}}{\overline{x}_{ik}}, \frac{\overline{x}_{ij}}{\underline{x}_{ik}}\right], \quad i = 1, \ldots, n, \ j, k = 1, \ldots, D, \ j \neq k. \tag{1}$$

For illustration, the ratios between concentrations of chemical elements presented in Table 1 are shown in Table 2.

Table 2. Ratios between concentrations of chemical elements in Czech mineral waters

Mineral waters	Calcium/Sodium		Calcium/Magnesium		Calcium/Potasium	
Hanácká kyselka	4.044	4.044	1.096	1.096	15.537	15.537
Korunní	2.534	2.932	0.705	0.883	3.071	3.761
Magnesia	0.177	0.231	5.191	9.605	25.214	29.500
Mattoni	3.518	3.573	1.098	1.232	4.611	4.922
Ondrášovka	9.697	10.278	6.214	6.829	120.0	142.429
Poděbradka	2.884	3.205	0.395	0.4236	2.855	3.096
Dobrá voda	0.897	1.126	0.870	1.126	0.926	1.151

For possible further dealing with interval ratios R^i_{jk} obtained by (1), it is worth to note that R^i_{jk} and R^i_{jl}, $k \neq l$, are not independent intervals since the same interval $[\underline{x}_{ij}, \overline{x}_{ij}]$ is used for their calculation.

4 Proportional Representation

The original compositions $\mathbf{x}_i$, $i = 1, \ldots, n$, are often represented so that the sums of the components for each composition are equal to an arbitrary (but

fixed) constant $\kappa > 0$. Such a representation is formally expressed by the *closure operation*

$$\mathcal{C}(\mathbf{x}_i) := \left(\frac{\kappa \cdot x_{i1}}{\sum_{j=1}^{D} x_{ij}}, \ldots, \frac{\kappa \cdot x_{iD}}{\sum_{j=1}^{D} x_{ij}} \right), \quad i = 1, \ldots, n.$$

The constant κ is popularly taken as 1 (or 100) in case of proportional (percentage) representation. It is essential that the proportional representation keeps the ratios between the compositional parts as

$$\frac{\frac{\kappa \cdot x_{ik}}{\sum_{j=1}^{D} x_{ij}}}{\frac{\kappa \cdot x_{il}}{\sum_{j=1}^{D} x_{ij}}} = \frac{x_{ik}}{x_{il}}, \quad k, l = 1, \ldots, D.$$

Note that the resulting *scale invariance* is one of the basic properties of compositional data, reflected also by the Aitchison geometry [8] that forms a natural algebraic-geometrical structure of compositions. Without the loss of generality, we will assume $\kappa = 1$ further in the paper.

For the interval compositional data $\mathbf{X}_i$, $i = 1, \ldots, n$, the situation becomes more complex. Observe that in the proportions

$$\mathcal{C}(\mathbf{x}_i)_k := \frac{x_{ik}}{\sum_{j=1}^{D} x_{ij}}, \ k = 1, \ldots, D,$$

the variable x_{ik} appears both in the numerator and the denominator. Hence, we cannot apply the concept of standard interval arithmetic and compute the k-th interval proportion in the following way:

$$\frac{[\underline{x}_{ik}, \overline{x}_{ik}]}{\sum_{j=1}^{D} [\underline{x}_{ij}, \overline{x}_{ij}]} = \frac{[\underline{x}_{ik}, \overline{x}_{ik}]}{\left[\sum_{j=1}^{D} \underline{x}_{ij}, \sum_{j=1}^{D} \overline{x}_{ij}\right]} = \left[\frac{\underline{x}_{ik}}{\sum_{j=1}^{D} \overline{x}_{ij}}, \frac{\overline{x}_{ik}}{\sum_{j=1}^{D} \underline{x}_{ij}} \right],$$

as the numerator and the denominator are not independent intervals. The correct procedure for computing the intervals $[\underline{c}_{ik}, \overline{c}_{ik}]$, $k = 1, \ldots, D$, that express the ranges of particular proportions from $\mathcal{C}(\mathbf{X}_i)$ is given in the following way (the formulas were developed for the first time in [4] for normalizing interval weights):

$$\underline{c}_{ik} = \min \left\{ \frac{x_{ik}}{\sum_{j=1}^{D} x_{ij}} \mid x_{ij} \in \left[\underline{x}_{ij}, \overline{x}_{ij}\right], \ j = 1, \ldots, D \right\} = \frac{\underline{x}_{ik}}{\underline{x}_{ik} + \sum_{j=1,\, j \neq k}^{D} \overline{x}_{ij}},$$

$$\overline{c}_{ik} = \max \left\{ \frac{x_{ik}}{\sum_{j=1}^{D} x_{ij}} \mid x_{ij} \in \left[\underline{x}_{ij}, \overline{x}_{ij}\right], \ j = 1, \ldots, D \right\} = \frac{\overline{x}_{ik}}{\overline{x}_{ik} + \sum_{j=1,\, j \neq k}^{D} \underline{x}_{ij}}.$$

The interval proportions of concentrations of chemical elements presented in Table 1 are given in Table 3.

However, the obtained intervals $[\underline{c}_{ik}, \overline{c}_{ik}]$, $k = 1, \ldots, D$, are not independent, so it is not correct to compute their ratios by means of standard interval arithmetic. Applying the results concerning normalization of interval weights that were proved in [7], we find out that the following general relations hold:

Table 3. Interval proportions of concentrations of chemical elements from Table 1

Mineral waters	Calcium		Sodium		Magnesium		Potasium	
Hanácká kyselka	0.45	0.45	0.111	0.111	0.41	0.41	0.029	0.029
Korunní	0.319	0.365	0.117	0.134	0.407	0.459	0.092	0.011
Magnesia	0.145	0.183	0.783	0.823	0.017	0.031	0.006	0.006
Mattoni	0.415	0.436	0.117	0.123	0.352	0.38	0.085	0.094
Ondrášovka	0.786	0.8	0.077	0.082	0.117	0.127	0.006	0.007
Poděbradka	0.237	0.25	0.076	0.085	0.584	0.605	0.078	0.086
Dobrá voda	0.23	0.274	0.24	0.261	0.242	0.267	0.234	0.253

$$R^i_{jk} \subseteq \frac{[\underline{c}_{ij}, \overline{c}_{ij}]}{[\underline{c}_{ik}, \overline{c}_{ik}]} \quad j,k = 1,\ldots,D,\ j \neq k.$$

Example 1. Let us consider the interval ratio between concentrations of calcium and sodium in mineral water Korunní $[2.534, 2.932]$ (see Table 2). If we compute, by means of the standard interval arithmetic operations, the ratio between interval proportions of calcium and sodium on the whole presented in Table 3, we obtain the following result:

$$\frac{[0.319, 0.365]}{[0.117, 0.134]} = [2.373, 3.125]\,.$$

We can see that the interval ratio $[2.534, 2.932]$ is indeed a strict subset of $[2.373, 3.125]$.

The interactions among the proportions $[\underline{c}_{ik}, \overline{c}_{ik}]$, $k = 1,\ldots,D$, mean that the proper proportional representation of interval compositional data $\mathbf{X}_i$, $i = 1,\ldots,n$, has to be given in the following way:

$$\mathcal{C}(\mathbf{X}_i) := \left\{\mathcal{C}(\mathbf{x}_i) \in [0,1]^D \mid \mathbf{x}_i \in [\underline{x}_{i1}, \overline{x}_{i1}] \times \ldots \times [\underline{x}_{iD}, \overline{x}_{iD}]\right\}. \tag{2}$$

Employing the results proved in [6] concerning normalization of interval weights, we can see that, unless $D = 2$, the interval proportions $[\underline{c}_{ik}, \overline{c}_{ik}]$, $k = 1,\ldots,D$, alone do not carry the whole information about the proportional representation of interval compositional data. From (2), we can see that we still have to know the initial interval compositional data $\mathbf{X}_i$, $i = 1,\ldots,n$. For $D = 2$, it is on the contrary sufficient to know only one interval proportion, e.g. $[\underline{c}_{i1}, \overline{c}_{i1}]$, $\mathcal{C}(\mathbf{X}_i)$ can be then given as follows:

$$\mathcal{C}(\mathbf{X}_i) = \left\{(c_{i1}, c_{i2}) \in [0,1]^2 \mid c_{i1} \in [\underline{c}_{i1}, \overline{c}_{i1}]\,,\ c_{i2} = 1 - c_{i1}\right\}.$$

Remark 1. Note that if the ratios between two proportions are calculated properly, they are equal to the ratios between the corresponding original compositional parts (the following procedure is inspired by the procedure introduced

in [7] for computing the ratios between normalized fuzzy weights). For $j, k = 1, \ldots, D$, $j \neq k$, let us denote the proper ratio between the j-th and k-th proportions by $[\underline{r}^i_{jk}, \overline{r}^i_{jk}]$. Then

$$\underline{r}^i_{jk} = \min\left\{ \tfrac{c_{ij}}{c_{ik}} \mid c_{ij} \text{ and } c_{ik} \text{ express the } j\text{-th and } k\text{-th components of at least one } \mathcal{C}(\mathbf{x}_i) \in \mathcal{C}(\mathbf{X}_i) \right\},$$

$$\overline{r}^i_{jk} = \max\left\{ \tfrac{c_{ij}}{c_{ik}} \mid c_{ij} \text{ and } c_{ik} \text{ express the } j\text{-th and } k\text{-th components of at least one } \mathcal{C}(\mathbf{x}_i) \in \mathcal{C}(\mathbf{X}_i) \right\}.$$

It can be shown (see [7, Theorem 8]) that $[\underline{r}^i_{jk}, \overline{r}^i_{jk}] = R^i_{jk}$ for any $j, k = 1, \ldots, D$, $j \neq k$.

5 Descriptive Statistics

Specific properties of (precise) compositional data, captured by the Aitchison geometry, should be reflected also by their descriptive statistics [1,9]. For instance, the arithmetic mean as a measure of location needs to be replaced by the geometric mean (centre) of compositional data, $\mathbf{g}(\mathbf{x}) := \left(g\left(\mathbf{x}^1\right), \ldots, g\left(\mathbf{x}^D\right)\right)$, where $\mathbf{x}^j := (x_{1j}, \ldots, x_{nj})$ and $g\left(\mathbf{x}^j\right) := \sqrt[n]{\prod_{i=1}^n x_{ij}}$, $j = 1, \ldots, D$. Note that the centre can be computed from an arbitrary representation of the input compositions $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the ratios between parts of $\mathbf{g}(\mathcal{C}(\mathbf{x}))$ remain always the same, i.e.

$$\frac{g\left(\mathbf{x}^j\right)}{g\left(\mathbf{x}^k\right)} = \frac{g\left(\mathcal{C}\left(\mathbf{x}^j\right)\right)}{g\left(\mathcal{C}\left(\mathbf{x}^k\right)\right)}, \quad j, k = 1, \ldots, D. \tag{3}$$

Now, let us consider the case of interval compositional data $\mathbf{X}_i$, $i = 1, \ldots, n$, introduced in Section 3. Since the particular intervals are assumed to be independent, the centre of these data is given as a vector $\mathbf{g}(\mathbf{X}) = \left(g\left(\mathbf{X}^1\right), \ldots, g\left(\mathbf{X}^D\right)\right)$, where

$$\mathbf{X}^j := \left(\left[\underline{x}_{1j}, \overline{x}_{1j}\right], \ldots, \left[\underline{x}_{nj}, \overline{x}_{nj}\right]\right), \quad j = 1, \ldots, D,$$

and

$$g\left(\mathbf{X}^j\right) := \left[\sqrt[n]{\prod_{i=1}^n \underline{x}_{ij}}, \sqrt[n]{\prod_{i=1}^n \overline{x}_{ij}}\right], \quad j = 1, \ldots, D.$$

Note that the particular intervals $g\left(\mathbf{X}^j\right)$, $j = 1, \ldots, D$, are independent. Hence, their ratios have to be computed applying the concept of standard interval arithmetic.

At the end of this section, let us verify the validity of equality (3) in the case of interval compositional data. Let

$$\mathcal{C}\left(\mathbf{X}^j\right) := \left(\left[\underline{c}_{1j}, \overline{c}_{1j}\right], \ldots, \left[\underline{c}_{nj}, \overline{c}_{nj}\right]\right), \quad j = 1, \ldots, D,$$

and $\mathbf{g}(\mathcal{C}(\mathbf{X})) = \left(g\left(\mathcal{C}\left(\mathbf{X}^1\right)\right), \ldots, g\left(\mathcal{C}\left(\mathbf{X}^D\right)\right)\right)$, where, for $i = 1, \ldots, n$, $j = 1, \ldots, D$, $\left[\underline{c}_{ij}, \overline{c}_{ij}\right]$ expresses the range of proportion of the j-th part in $\mathbf{X}_i$. Since the intervals $\left[\underline{c}_{ij}, \overline{c}_{ij}\right]$, $i = 1, \ldots, n$, $j = 1, \ldots, D$, are not independent, also the obtained intervals $g\left(\mathcal{C}\left(\mathbf{X}^1\right)\right), \ldots, g\left(\mathcal{C}\left(\mathbf{X}^D\right)\right)$ are not independent and their ratios cannot be computed applying the concept of standard interval arithmetic. If we do so, we obtain the following relation instead of equality (3):

$$\frac{g\left(\mathbf{X}^j\right)}{g\left(\mathbf{X}^k\right)} \subseteq \frac{g\left(\mathcal{C}\left(\mathbf{X}^j\right)\right)}{g\left(\mathcal{C}\left(\mathbf{X}^k\right)\right)}, \quad j, k = 1, \ldots, D.$$

Therefore, for retaining the validity of equality (3), we have to respect the interactions among $g\left(\mathcal{C}\left(\mathbf{X}^1\right)\right), \ldots, g\left(\mathcal{C}\left(\mathbf{X}^D\right)\right)$ when computing their ratios.

6 Future Work

Interval compositional data form a natural extension of the precise case. We have studied possible ways of dealing with such data. Future work will be aimed at extension of other procedures developed for dealing with compositional data. Another problem worth to study will be utilization of the information about precise sum of compositional parts, that is frequently available in the practice.

Acknowledgments. The paper is supported by the Operational Program Education for Competitiveness-European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic), and the grant IGA_PrF_2014028 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London (1986)
2. Billard, L., Diday, E.: From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. J. Amer. Statist. Assoc. 98, 470–487 (2003)
3. Brito, P., Silva, A.P.D.: Modelling Interval Data with Normal and Skew-Normal Distributions. J. Appl. Stat. 39, 3–20 (2012)
4. Dubois, D., Prade, H.: The Use of Fuzzy Numbers in Decision Analysis. In: Gupta, M.M., Sanchez, E. (eds.) Fuzzy Information and Decision Processes, pp. 309–321. North-Holland, Amsterdam (1982)
5. Klir, G.J., Pan, Y.: Constrained Fuzzy Arithmetic: Basic Questions and Some Answers. Soft Comput. 2, 100–108 (1998)
6. Pavlačka, O.: Note on the lack of equality between fuzzy weighted average and fuzzy convex sum. Fuzzy Set. Syst. 213, 102–105 (2013)
7. Pavlačka, O.: On various approaches to normalization of interval and fuzzy weights. Fuzzy Set. Syst. 243, 110–130 (2014)
8. Pawlowsky-Glahn, V., Buccanti, A. (eds.): Compositional Data Analysis: Theory and Applications. Wiley, Chichester (2011)
9. Pawlowsky-Glahn, V., Egozcue, J.J.: BLU Estimators and Compositional Data. Math. Geol. 34, 259–274 (2002)