Food Science's Chemical Element Compositions

Subjects: Chemistry, Applied
Contributor: Matthias Templ

In recent years, many analyses have been carried out to investigate the chemical components of food data. However, studies rarely consider the compositional pitfalls of such analyses. This is problematic as it may lead to arbitrary results when non-compositional statistical analysis is applied to compositional datasets. In this study, compositional data analysis (CoDa), which is widely used in other research fields, is compared with classical statistical analysis to demonstrate how the results vary depending on the approach and to show the best possible statistical analysis. For example, honey and saffron are highly susceptible to adulteration and imitation, so the determination of their chemical elements requires the best possible statistical analysis. Our study demonstrated how principle component analysis (PCA) and classification results are influenced by the pre-processing steps conducted on the raw data, and the replacement strategies for missing values and non-detects. Furthermore, it demonstrated the differences in results when compositional and non-compositional methods were applied. Our results suggested that the outcome of the log-ratio analysis provided better separation between the pure and adulterated data and allowed for easier interpretability of the results and a higher accuracy of classification. Similarly, it showed that classification with artificial neural networks (ANNs) works poorly if the CoDa pre-processing steps are left out. From these results, we advise the application of CoDa methods for analyses of the chemical elements of food and for the characterization and authentication of food products.

Keywords: composition of food ; log-ratio analysis ; PCA ; classification ; artificial neural networks ; adulteration ; honey ; saffron ; chemical profiling

1. Introduction

The importance of food composition data to nutrition and public health has been long acknowledged [1]. Currently, hundreds of articles have been published on the chemical composition of various kinds of food. The statistical techniques most often used are cluster analysis, principal component analysis (PCA), numerous classification methods, regression [2][3][4] and partial least-squares regression methods [5][6].

An inspection of the literature on the analytical and statistical methods frequently used in food science [2][3][4] as well as in chemometrics of honey [7] do not mention compositional data analysis (CoDa) [8]. A composition is the quantified decomposition of a whole into its component parts. Historically, a composition was described as random vectors with strictly positive components that added up to a whole, e.g., 100. Currently, it stands for all vectors that represent parts of a whole and carry relative information. The whole may only exist theoretically and be different for each composition [9]. CoDa, including the log-ratio methodology described later, is a method for describing the parts/connections of a whole that conveying relative information. Compositional methods are well established in many fields dealing with compositional data, such as material science [10], water chemistry [11], geochemistry [12], and air pollution chemistry [13]. Recently, the successful application of CoDa was demonstrated in food chemistry [14] by analyzing the chemical compounds in beer samples. It is well-known from the literature [9][15][16][17], that if traditional statistical analysis is applied to compositional datasets, correlations will be arbitrary [9][14] and even the arithmetic mean is not an adequate measure for the center of the distribution [18], which may lead to wrong conclusions [9][14][15]. Therefore, CoDa can be a way to gain additional insight and see beyond a constrained space (the simplex). While in most articles non-compositional methods for the statistical analysis of food are applied, there are a few exceptions. Cayuela-Sanchez (2020) used CoDa to investigate the composition of various pastries, biscuits [19] and olive oil [20]. Furthermore, E. Parent and their lab use CoDa theory for the diagnosis of various nutrients [21], for instance, fruit crops [22][23], bananas [24] and citrus [25]. Compositional data analysis focuses on log ratios between the parts (see Equations (3) and (4) for isometric and centered log-ratios), so that their relative scale and inherent interplay are accounted for. To demonstrate problems that may arise during the analysis of chemical elements in food science, datasets on the chemical compositions of honey and saffron were selected. Chemical profiling of honey [26][27] and saffron [28][29] is an important issue when determining their botanical and geographical origins. Honey is mainly composed of sugars and water with minor amounts of minerals, vitamins, amino acids, organic acids, flavonoids and other phenolic compounds, and aromatic substances [27][30]. The determinants of its composition,

color, aroma and flavor are the flowers, geographical regions, climate and species of honeybee [30][31]. As mislabeling and adulteration of honey has become a worldwide problem, it is crucial not only to detect the adulterants in honey but also to classify honey samples correctly. The technical challenge of detecting adulterants in honey is widely discussed [7][32][33], the challenge of finding a theoretically sound statistical analysis is little understood. Similarly, saffron, which has numerous health benefits and is the world's most expensive spice, is the object of fraudulent production and unethical trade practices [34]. The three major secondary metabolites which are important for the high quality of saffron are: crocins, which account for the yellow pigmentation from the stigma; picrocrocin, which gives it its rusty, bittersweet flavor; and safranal, which lends an earthy fragrance to the spice. It was hypothesized that additional insights into the chemical composition of honey and saffron samples might be obtained from a better interpretable results using CoDa. It was also assumed that a higher misclassification rate, lower predictive power, and a lower explained variance were inherent in a non-compositional analysis of compositional datasets.

The aim of this research was to compare compositional data analysis with classical statistical analyses to demonstrate how data pre-processing can influence a multivariate analysis, how a proper analysis can improve interpretation, and how a compositional method improves the accuracy of classification.

2. Development and Findings

Compositional data analysis using log-ratios is a theoretically sound concept that is well known in many sciences but rarely applied in food science. It is problematic because, if traditional statistical analysis is applied to a compositional dataset, correlations can be arbitrary and even the arithmetic mean is not an adequate measure for the center of the distribution [9]. Both of our null hypotheses for interpretability and misclassification rates were supported: higher explained variance and smaller misclassification rates were obtained when the compositional nature of the datasets was considered in the analysis.

Whenever a method takes the nature of compositional data into account, it leads to better interpretability of the results. Biplots obtained from various PCAs demonstrated how the pre-processing of data may influence the analysis. When CoDa was not considered, biplots were clearly distorted, which was best seen from the direction of the loading vectors. This is because the concept of linear correlation was not working and was theoretical unsound since the correlation between the parts of a composition is always biased toward a negative one. The variance of the first principal components was the highest when clr and ilr were applied, which confirmed that it was not advisable to apply PCAs to compositional data without using an appropriate log-ratio presentation of the data. The highest misclassification instances were gathered when no transformation was performed before data analysis or when the data were closed and standarized. Thus, the accuracy of the classification methods improved when CoDa was used.

To sum up some advantages, the theoretical correctness of compositional data analysis methods is undoubted and has been proven by many authors starting with the main works of [8]. In addition, the size effect—when a true measurement (e.g., an instrumental signal) $x=[x1,x2,\ldots,xn]$ cannot be observed directly but $cx=[cx1,cx2,\ldots,cxn]$ is observed—can be ignored when using compositional data analysis. The measurement is from the same equivalence class and the ratios between parts are also the same. Higher predictive power and better results are generally obtained.

Note that class modelling approaches can also be used for classification with respect to a one-class classification problem [35], such as when investigating adulterated versus non-adulterated honey or genuine honey versus all other non-real honey samples. One way to do this is classical soft independent modelling by class analogy (SIMCA) [36] or robust SIMCA [37]. The results were not satisfactory and the three other methods (LDA, KNN, and ANN) outperformed SIMCA, so the results were excluded so as not to go beyond the scope of the paper.

However, there are also several drawbacks to be discussed. Outliers are produced after presenting data in centered or isometric log-ratio coordinates whenever an observation lays on the boarder of the simplex. One solution is to use robust statistical methods to analyse such data [9]. True zeros and rounded zeros are not in the simplex by definition and a log-ratio with a zero is not possible. True zeros are still an unsolved problem in CoDa even though some solutions have already been presented [38]. Rounded zeros often come from too-small concentrations with too few precise measurement units. Rounded zeros cause extra work when using compositional methods, and they must be imputed first by using a censored method. Solution by imputation of rounded zeros are outlined in this contribution [39][40] and were applied to the honey samples. In addition, the centered log-ratio transformation is often used because of its simplicity, but using well-selected balances [8] for specific isometric log-ratio transformations often leads to better interpretable results. In addition, for instrumental signals (e.g., NMR, LC-MS, or GC-MS), all possible log-ratios may be used instead of centered or isometric log-ratios [41]; that is, each variable is divided by one of the other variables before the logarithm is taken. For a

dataset with 10 variables, there are already 45 possible log-ratios between the variables. The authors of [42] suggested using all possible log-ratios, but since this would lead to a large number, they suggest using feature selection to reduce the number of log-ratios. They argue that a centered log-ratio transformation may average too much leading to a higher false discovery rates of biomarkers [42].

This study had limitations. The usage of CoDa was demonstrated only on two datasets (honey and saffron), which originated from different fields of food science (food substance and spice). Therefore, other food datasets need to be analyzed with CoDa to establish its broad usage in food science. Furthermore, for the application of ANNs we would have needed larger datasets as these methods work better with big data, but in the field of food science, large datasets are seldom available. However, our results indicated that by incorporating the theory of CoDa, the predictive classification methods will lead to better performance, which may be used to improve the characterization of food products.

Our aim was to create awareness of the choice of compositional methods when compositional data to be analyzed. The CoDa of mineral elements of honey samples as well as trace element concentrations of saffron samples allowed us to demonstrate the correct assessment of compositions and to recommend that this application be extended to an analysis of any food composition. It would also allow for the establishment of CoDa in food science. It is expected that similar results can be gathered from the analysis of other datasets of other food substance and spices.

3. Conclusions

Principal component analysis revealed the pitfalls of classical analysis conducted on compositional data: distorted biplots and less-explained variance. Classification resulted in a less predictive power when a non-CoDa method was used. Replacement strategies of non-detects should be also based on log-ratio methods. Generally, using CoDa for chemical elements not only resulted in higher explained variance and lower misclassification rates but also enabled better interpretability of the results. However, depending on the type of data, one can expect some difficulties, which are mentioned in the discussion section (outliers, the zero problem, and the choice of log-ratio transformation). It is therefore advisable to apply compositional analysis (CoDa) methods in the analysis of chemical elements in food.

References

1. Elmadfa, I.; Meyer, A.L. Importance of food composition data to nutrition and public health. Eur. J. Clin. Nutr. 2010, 64, S4–S7.
2. Granato, D.; de Araújo Calado, V.M.; Jarvis, B. Observations on the use of statistical methods in Food Science and Tec hnology. Food Res. Int. 2014, 55, 137–149.
3. Nunes, C.A.; Alvarenga, V.O.; de Souza Sant'Ana, A.; Santos, J.S.; Granato, D. The use of statistical software in food s cience and technology: Advantages, limitations and misuses. Food Res. Int. 2015, 75, 270–280.
4. Granato, D.; Santos, J.S.; Escher, G.B.; Ferreira, B.L.; Maggio, R.M. Use of principal component analysis (PCA) and hi erarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. Trends Food Sci. Technol. 2018, 72, 83–90.
5. Gottardo, P.; Penasa, M.; Lopez-Villalobos, N.; De Marchi, M. Variable selection procedures before partial least square s regression enhance the accuracy of milk fatty acid composition predicted by mid-infrared spectroscopy. J. Dairy Sci. 2 016, 99, 7782–7790.
6. Kamruzzaman, M.; ElMasry, G.; Sun, D.W.; Allen, P. Non-destructive prediction and visualization of chemical compositi on in lamb meat using NIR hyperspectral imaging and multivariate regression. Innov. Food Sci. Emerg. Technol. 2012, 16, 218–226.
7. Fakhlaei, R.; Selamat, J.; Khatib, A.; Razis, A.F.A.; Sukor, R.; Ahmad, S.; Babadi, A.A. The Toxic Impact of Honey Adult eration: A Review. Foods 2020, 9, 1538.
8. Aitchison, J. The Statistical Analysis of Compositional Data; Chapman & Hall: London, UK, 1986.
9. Filzmoser, P.; Hron, K.; Templ, M. Applied Compositional Data Analysis. With Worked Examples in R; Springer Series in Statistics; Springer: Cham, Switzerland, 2018.
10. Pesenson, M.Z.; Suram, S.K.; Gregoire, J.M. Statistical Analysis and Interpolation of Compositional Data in Materials S cience. ACS Comb. Sci. 2015, 17, 130–136.
11. Buccianti, A.; Pawlowsky-Glahn, V. New Perspectives on Water Chemistry and Compositional Data Analysis. Math. Ge ol. 2005, 37, 703–727.

12. Buccianti, A.; Grunsky, E. Compositional data analysis in geochemistry: Are we sure to see what really occurs during n atural processes? J. Geochem. Explor. 2014, 141, 1–5.
13. Meier, M.F.; Mildenberger, T.; Locher, R.; Rausch, J.; Zünd, T.; Neururer, C.; Ruckstuhl, A.; Grobèty, B. A model based t wo-stage classifier for airborne particles analyzed with Computer Controlled Scanning Electron Microscopy. J. Aerosol Sci. 2018, 123, 1–16.
14. Templ, M.; Templ, B. Analysis of Chemical Compounds in Beverages- Guidance for Establishing a Compositional Analy sis. Food Chem. 2020, 325, 126755.
15. Greenacre, M. Compositional Data Analysis in Practice; CRC Press: Boca Raton, FL, USA, 2018.
16. van den Boogaart, G.K.; Tolosana-Delgado, R. Analyzing Compositional Data with R; Use R! Book Series; Springer: Be rlin/Heidelberg, Germany, 2013.
17. Pawlowsky-Glahn, V.; Egozcue, J.; Tolosana-Delgado, J. Lecture Notes on Compositional Data Analysis. 2007. Availabl e online: http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf (accessed on 3 September 2021).
18. Hron, K.; Templ, M.; Filzmoser, P. Estimation of a proportion in survey sampling using the logratio approach. Metrika 20 13, 76, 799–818.
19. Cayuela-Sánchez, J.A.; Palarea-Albaladejo, J.; Zira, T.P.; Moriana-Correro, E. Compositional method for measuring the nutritional label components of industrial pastries and biscuits based on Vis/NIR spectroscopy. J. Food Compos. Anal. 2020, 92, 103572.
20. Cayuela-Sánchez, J.A.; Palarea-Albaladejo, J.; García-Martín, J.F.; del Carmen Pérez-Camino, M. Olive oil nutritional l abeling by using Vis/NIR spectroscopy and compositional statistical methods. Innov. Food Sci. Emerg. Technol. 2019, 5 1, 139–147.
21. Parent, L.; Dafir, M. A Theoretical Concept of Compositional Nutrient Diagnosis. J. Am. Soc. Hortic. Sci. 1992, 117, 239 –242.
22. Parent, L.E. Diagnosis of the nutrient compositional space of fruit crops. Rev. Bras. Frutic. 2011, 33, 321–334.
23. Parent, L.E.; Rozane, D.E.; de Deus, J.A.L.; Natale, W. Diagnosis of nutrient composition in fruit crops: Major developm ents. In Fruit Crops; Srivastava, A., Hu, C., Eds.; Elsevier: Amsterdam, The Netherlands, 2020; Chapter 12; pp. 145–15 6.
24. Neto, A.; Deus, J.; Filho, V.; Natale, W.; Parent, L.E. Nutrient Diagnosis of Fertigated prata and Cavendish banana (Mu sa spp.) at Plot-Scale. Plants 2020, 9, 1467.
25. Rozane, D.E.; Mattos, D., Jr.; Parent, S.É.; Natale, W.; Parent, L.E. Meta-analysis in the Selection of Groups in Varietie s of Citrus. Commun. Soil Sci. Plant Anal. 2015, 46, 1948–1959.
26. Wang, J.; Li, Q.X. Chapter 3—Chemical Composition, Characterization, and Differentiation of Honey Botanical and Geo graphical Origins. Adv. Food Nutr. Res. 2011, 62, 89–137.
27. Santos-Buelga, C.; González-Paramás, A.M. Chemical Composition of Honey. In Bee Products-Chemical and Biologica l Properties; Alvarez-Suarez, J.M., Ed.; Springer: Cham, Switzerland, 2017; pp. 43–82.
28. Maggi, L.; Carmona, M.; Kelly, S.D.; Marigheto, N.; Alonso, G.L. Geographical origin differentiation of saffron spice (Cro cus sativus L. stigmas)—Preliminary investigation using chemical and multi-element (H, C, N) stable isotope analysis. Food Chem. 2011, 128, 543–548.
29. Wakefield, J.; McComb, K.; Ehtesham, E.; Van Hale, R.; Barr, D.; Hoogewerff, J.; Frew, R. Chemical profiling of saffron for authentication of origin. Food Control 2019, 106, 106699.
30. da Silva, P.M.; Gauche, C.; Gonzaga, L.V.; Costa, A.C.O.; Fett, R. Honey: Chemical composition, stability and authentic ity. Food Chem. 2016, 196, 309–323.
31. Escuredo, O.; Dobre, I.; Fernández-González, M.; Seijo, M.C. Contribution of botanical origin and sugar composition of honeys on the crystallization phenomenon. Food Chem. 2014, 149, 84–90.
32. Se, K.W.; Wahab, R.A.; Syed Yaacob, S.N.; Ghoshal, S.K. Detection techniques for adulterants in honey: Challenges a nd recent trends. J. Food Compos. Anal. 2019, 80, 16–32.
33. Soares, S.; Amaral, J.S.; Oliveira, M.B.P.; Mafra, I. A Comprehensive Review on the Main Honey Authentication Issues: Production and Origin. Compr. Rev. Food Sci. Food Saf. 2017, 16, 1072–1100.
34. Hagh-Nazari, S.; Keifi, N. Saffron and Various Fraud Manners in Its Production and Trades. In Acta Horticulturae; Intern ational Society for Horticultural Science (ISHS): Leuven, Belgium, 2007; pp. 411–416.
35. Rodionova, O.; Oliveri, P.; Pomerantsev, A. Rigorous and compliant approaches to one-class classification. Chemom. I ntell. Lab. Syst. 2016, 159, 89–96.

36. Wold, S.; Sjöström, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In Chemometrics: Theory and Application; ACS Symposium Series; American Chemical Society: Washington, DC, USA, 1977; Volume 52, pp. 243–282.
37. Branden, K.V.; Hubert, M. Robust classification in high dimensions based on the SIMCA Method. Chemom. Intell. Lab. Syst. 2005, 79, 10–21.
38. Templ, M.; Hron, K.; Filzmoser, P. Exploratory tools for outlier detection in compositional data with structural zeros. J. Appl. Stat. 2017, 44, 734–752.
39. Templ, M.; Hron, K.; Filzmoser, P.; Gardlo, A. Imputation of rounded zeros for high-dimensional compositional data. Chemom. Intell. Lab. Syst. 2016, 155, 183–190.
40. Templ, M. Artificial Neural Networks to Impute Rounded Zeros in Compositional Data. In Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn; Filzmoser, P., Hron, K., Martín-Fernández, J.A., Palarea-Albaladejo, J., Eds.; Springer: Cham, Switzerland, 2021; pp. 163–187.
41. Filzmoser, P.; Walczak, B. What can go wrong at the data normalization step for identification of biomarkers? J. Chromatogr. A 2014, 1362, 194–205.
42. Malyjurek, Z.; de Beer, D.; Joubert, E.; Walczak, B. Working with log-ratios. Anal. Chim. Acta 2019, 1059, 16–27.