Analyzing Wildland Fire Smoke Emissions Data Using Compositional Data Techniques

David Weise[1], Javier Palarea-Albaladejo[2], Timothy J. Johnson[3], and Heejung Jung[4]

[1]USDA Forest Service
[2]Biomathematics and Statistics Scotland
[3]Pacific Northwest National Laboratory
[4]University of California, Riverside

November 22, 2022

Abstract

By conservation of mass, the mass of wildland fuel that is pyrolyzed and combusted must equal the mass of smoke emissions, residual char and ash. For a given set of conditions, these amounts are fixed. This places a constraint on smoke emissions data which violates key assumptions for many of the statistical methods ordinarily used to analyze these data such as linear regression, analysis of variance, and t-tests. These data are inherently multivariate, relative, and non-negative parts of a whole and are then characterized as so-called compositional data. This paper introduces the field of compositional data analysis to the biomass burning emissions community and provides examples of statistical treatment of emissions data. Measures and tests of proportionality, unlike ordinary correlation, allow one to coherently investigate associations between parts of the smoke composition. An alternative method based on compositional linear trends was applied to estimate trace gas composition over a range of combustion efficiency which reduced prediction error by 4 percent while avoiding use of modified combustion efficiency as if it were an independent variable. Use of log-ratio balances to create meaningful contrasts between compositional parts definitively stressed differences in smoke emissions from fuel types originating in the southeastern and southwestern U.S. Application of compositional statistical methods as an appropriate approach to account for the relative nature of data about the composition of smoke emissions and the atmosphere is recommended.

Hosted file

`agusupporting-information weise et al.docx` available at https://authorea.com/users/537272/articles/599279-analyzing-wildland-fire-smoke-emissions-data-using-compositional-data-techniques

Analyzing Wildland Fire Smoke Emissions Data Using Compositional Data Techniques

David R. Weise [1], Javier Palarea-Albaladejo[2], Timothy J. Johnson[3], and Heejung Jung[4]

[1] USDA Forest Service, Pacific Southwest Research Station, Riverside, CA, 92507, USA

[2] Biomathematics and Statistics Scotland, Edinburgh, Scotland, UK

[3] Pacific Northwest National Laboratories, Richland, WA, USA

[4] Department of Mechanical Engineering, University of California, Riverside, CA, USA

Corresponding author: David R. Weise (david.weise@usda.gov)

Key Points:

- Emissions data involve multiple and interrelated elements that can be more effectively analyzed using multivariate statistical techniques.
- By mass conservation, the range of emissions data is inherently constrained, making techniques for compositional data appropriate.
- Relating trace gas emissions to modified combustion efficiency using compositional linear regression accounts for these features.

Abstract

By conservation of mass, the mass of wildland fuel that is pyrolyzed and combusted must equal the mass of smoke emissions, residual char and ash. For a given set of conditions, these amounts are fixed. This places a constraint on smoke emissions data which violates key assumptions for many of the statistical methods ordinarily used to analyze these data such as linear regression, analysis of variance, and t-tests. These data are inherently multivariate, relative, and non-negative parts of a whole and are then characterized as so-called compositional data. This paper introduces the field of compositional data analysis to the biomass burning emissions community and provides examples of statistical treatment of emissions data. Measures and tests of proportionality, unlike ordinary correlation, allow one to coherently investigate associations between parts of the smoke composition. An alternative method based on compositional linear trends was applied to estimate trace gas composition over a range of combustion efficiency which reduced prediction error by 4 percent while avoiding use of modified combustion efficiency as if it were an independent variable. Use of log-ratio balances to create meaningful contrasts between compositional parts definitively stressed differences in smoke emissions from fuel types originating in the southeastern and southwestern U.S. Application of compositional statistical methods as an appropriate approach to account for the relative nature of data about the composition of smoke emissions and the atmosphere is recommended.

1 Introduction

Wildland fire is a complex phenomenon of chemical and physical processes. Two of the chemical processes which are key to wildland fire are pyrolysis and combustion (Shafizadeh,

1984; Ward, 2001). During pyrolysis, a solid wildland fuel is heated and breaks down into constituent parts consisting of gases, tars and a solid material called char (Di Blasi, 2008). During combustion, pyrolysis products react with oxygen releasing energy and a large assortment of gaseous and solid chemical compounds (e.g. Akagi et al., 2011; Andreae & Merlet, 2001; May et al., 2014). Oxidation reactions involving atmospheric gases such as nitrogen occur (Crutzen & Brauch, 2016; Lobert et al., 1990). By conservation of mass, the mass of the products is equal to the sum of the masses of the reactants. For the moment we assume that all products can be measured with complete accuracy. The measured masses of the individual products cannot exceed the total mass and are thus numerically related. Measuring a subset of the complete list of products simply makes the total mass unknown but does not change the inherent numerical dependency. For example, a simplified balanced global reaction describing combustion of wood containing water and no inorganic content, shows 1 kg dry wood (M=0) produces 1.82 kg CO_2 and 0.32 kg H_2O for a total mass of products of 2.14 kg (Byram, 1959)

$$\begin{aligned} &4\,C_6H_9O_4 + 25\,O_2 + \left[0.322M\ H_2O + 94\,N_2\right] \rightarrow \\ &\quad 18\,H_2O + 24\,CO_2 + \left[0.322M\ H_2O + 94\,N_2\right] + energy \end{aligned} . \tag{1}$$

Because of the chemical complexity of wood, Byram approximated the proportion of C, H, and O atoms in wood by $C_6H_9O_4$. Complete combustion with no dissociation is an idealized situation which explains the maximum product mass possible. Incomplete combustion and thermal dissociation will yield additional products and less CO_2. The foliage of woody plants has a different chemical composition from the wood component which can affect both combustion and combustion products (Hough, 1969; Jolly et al., 2016; Rogers et al., 1986). The addition of elements such as N and S, and a host of other elements (many inorganic) to the wood, along with inclusion of a) the oxidation of atmospheric N by fire (Paul et al., 2008) as well as b) incomplete

combustion changes Eq.(1) but does not change conservation of mass. The total mass of the products (*T*) can be partitioned to consist of CO, CO_2, particulate matter (*PM*), other gases, char and ash

$$\mathrm{T} = \mathrm{CO} + \mathrm{CO_2} + \text{other gases} + \mathrm{PM} + \text{char} + \text{ash} \quad . \tag{2}$$

If the masses in Eq. (2) are transformed into mass ratios by dividing by the total of CO and CO_2 as in

$$\begin{aligned} \frac{\mathrm{T}}{\mathrm{CO}+\mathrm{CO_2}} &= \frac{\mathrm{CO_2}+\mathrm{CO}+\text{other gases}+\mathrm{PM}+\text{char}+\text{ash}}{\mathrm{CO}+\mathrm{CO_2}} \\ &= \frac{\mathrm{CO_2}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\mathrm{CO}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\text{other gases}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\mathrm{PM}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\text{char}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\text{ash}}{\mathrm{CO}+\mathrm{CO_2}}, \\ &= MCE + \frac{\mathrm{CO}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\text{other gases}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\mathrm{PM}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\text{char}}{\mathrm{CO}+\mathrm{CO_2}} + \frac{\text{ash}}{\mathrm{CO}+\mathrm{CO_2}} \end{aligned} \tag{3}$$

it can be easily seen that modified combustion efficiency $MCE = \mathrm{CO_2}/(\mathrm{CO}+\mathrm{CO_2})$ is part of the total and numerically dependent on the other parts because *T* is fixed. The above example places the masses on a relative basis to the amount of CO and CO_2 produced by a fire, demonstrably two of the three primary products (water being the third). Emissions data have been expressed as relative measures such as emission ratios and emission factors, concentrations, mixing ratios, mole fractions, mass fractions, and volume fractions for a very long time (e.g. Darley et al., 1966; Gerstle & Kemnitz, 1967). However, the statistical properties of relative data have not typically been considered when these data have been analyzed. Compositional data analysis (CoDA) is an approach that explicitly considers the statistical properties of relative data (Aitchison, 1986). Compositional data are contained in the positive orthant of multidimensional real space (Barceló-Vidal et al., 2001). [An orthant is the multidimensional analogue of a quadrant in the more familiar two-dimensional Cartesian space.] A recent paper presented analysis of emissions data using positive matrix factorization which recognized the non-negative

nature of emissions data (Sekimoto et al 2018). From a compositional point of view, stoichiometric equations such as (1) have characteristics that discourage, for example, measuring association (correlation) between parts of the composition in the ordinary way (Egozcue et al., 2014).

Individual gases produced during pyrolysis and combustion have long been associated with the different phases or conditions (pre-ignition, flaming, smoldering, mixed phase) under which the pyrolysis and combustion have occurred (Lobert & Warnatz, 1993; Tangren et al., 1976). Combustion efficiency (CE) and MCE are indices developed to describe the completeness of the conversion of the carbon contained in the fuel to CO_2 (Ward et al., 1980; Ward & Hao, 1991; Yokelson et al., 1997). Theoretically CE includes all carbon produced; however, the challenge to account for all products and the predominance of CO_2 and CO in smoke emissions led to the development of MCE (Ward & Hao, 1991). It has become common practice to correlate emission factors of products other than CO_2 and CO with MCE (e.g. Amaral et al., 2014; Burling et al., 2010; Ferek et al., 1998; Goode et al., 2000; Janhäll et al., 2010; McMeeking et al., 2009; Shen et al., 2013; Urbanski, 2013; Ward & Hao, 1991; Yokelson et al., 2013) using ordinary linear regression. While linking combustion products to CE and MCE is physically sound, the approach to statistical analysis has to date often ignored the intrinsic multivariate and relative nature of these data. As shown above by Eq. (3), MCE as an explanatory variable is automatically correlated to every other wildland fire emission (response variable) by its formulation, not because of physical causation.

Aitchison (2003) showed for illustration how two scientists examining the correlation between animal, vegetable, mineral, and water proportions of a sample can arrive at very different correlations (and conclusions) if one scientist dried the sample removing all water

before calculating the correlations between the components. We present a similar example (Table 1) using emission factors previously reported (Radke et al., 1988). The original emission factors (g/kg) comprise a composition of *D* parts which were put on a consistent relative scale by applying the closure operation

$$C(\mathbf{x}) = \frac{[x_1, x_2, \ldots, x_D]}{\sum_{i=1}^{D} x_i} \tag{4}$$

which divides the emission for each gas (x_i) by the sum of emissions and express the data as fractions of a fixed total, this being 1 by default to be expressed in proportions but in general any other total by simple multiplication (e.g $C(\mathbf{x})\cdot 10^2$ for percentages or $C(\mathbf{x})\cdot 10^6$ for parts per million). After closure, the CO_2 emission factor of 1664 g/kg becomes the proportion 0.948.

Table 1. Example illustrating how ordinary correlation of relative data such as emission factors produces spurious results. Data from Radke et al. (1988).

Emission factors (g/kg)								
	CO	CO_2	CH_4	C_3H_6	C_2H_6	C_3H_8	C_2H_3	PM
	74	1664	2.4	0.58	0.35	0.21	0.32	13.5
	75	1650	3.6	0.46	0.55	0.32	0.21	23.0
	106	1626	3.0	0.70	0.60	0.25	0.22	6.1
	89	1637	2.6	0.08	0.56	0.42	0.19	20.2
Full composition (after closure *C*)								
	4.22E-02	9.48E-01	1.37E-03	3.30E-04	1.99E-04	1.20E-04	1.82E-04	7.69E-03
	4.28E-02	9.41E-01	2.05E-03	2.62E-04	3.14E-04	1.83E-04	1.20E-04	1.31E-02
	6.08E-02	9.33E-01	1.72E-03	4.02E-04	3.44E-04	1.43E-04	1.26E-04	3.50E-03
	5.09E-02	9.35E-01	1.49E-03	4.57E-05	3.20E-04	2.40E-04	1.09E-04	1.15E-02
No CO_2 subcomposition								
	8.10E-01		2.63E-02	6.35E-03	3.83E-03	2.30E-03	3.50E-03	1.48E-01
	7.27E-01		3.49E-02	4.46E-03	5.33E-03	3.10E-03	2.04E-03	2.23E-01
	9.07E-01		2.57E-02	5.99E-03	5.13E-03	2.14E-03	1.88E-03	5.22E-02
	7.87E-01		2.30E-02	7.08E-04	4.95E-03	3.72E-03	1.68E-03	1.79E-01
Pearson correlation								
CO_2	Full (F)		-0.25	0.18	-0.93	-0.49	0.82	0.25
CO	F	-0.88	-0.01	0.19	0.69	0.06	-0.44	-0.68
	Sub (S)		-0.59	0.40	-0.10	-0.67	-0.03	-1.00
CH_4	F			0.19	0.58	0.12	-0.53	0.31
	S			0.29	0.37	0.00	0.01	0.53
C_3H_6	F				-0.17	-0.92	0.51	-0.71
	S				-0.36	-0.95	0.61	-0.47
C_2H_6	F					0.54	-0.93	-0.01
	S					0.38	-0.92	0.09
C_3H_8	F						-0.81	0.65
	S						-0.53	0.72
C_2H_3	F							-0.34
	S							0.01

Regardless of the total, it is important to note that the resulting relativized data vector is equivalent to the original emission factors and lives in what is known as a *D*-part simplex and, hence, statistical analysis on any equivalent representation of the data should provide the same results. In the example in Table 1, $D = 8$ for the full composition. The familiar Pearson correlation coefficient (r) was then calculated for all pairs of gases. In the full composition, CO was negatively correlated with CO_2 (r=-0.88), positively correlated with C_2H_6 (0.69) and not correlated with C_3H_8 (0.06). The emission factor for CO_2 was then removed from the full

composition as if it had not been measured and the closure operation was performed on the subcomposition (subset), producing the second set of values, and correlation between the pairs was calculated. Correlation coefficients which changed appreciably are highlighted. Note that CO is now negatively correlated with C_3H_8 (-0.67) and not correlated with C_2H_6 (-0.10). Similarly, CH_4 was negatively correlated with C_2H_3 in the full composition (-0.53) and not correlated (0.01) when CO_2 is not in the composition. The point of this illustration is that this index (Pearson correlation coefficient) that is generally trusted as a measure of pairwise association can produce different results depending on something that should not affect it. This is an artifact not related to the actual relationship between the variables. Hence, it is not a reliable measure with this type of data, regardless of the magnitude of the difference in a particular case, which will be arbitrarily big or small. That the interpretation of these changing correlations could lead to spurious conclusions is well known (Pearson, 1896). This simple example illustrates the problem using correlation with relative data. Linear regression utilizes correlation so it is also affected by this constraint: measuring associations in terms of proportionality is a coherent and meaningful alternative to ordinary correlation for compositional data (Egozcue et al., 2014, 2018; Lovell et al., 2015).

The use of predictions resulting from correlations and linear regressions that do not account for the relative nature of emissions data may produce misleading estimates in emissions inventories developed by various regulatory agencies. This is true for operational tools such as the Fire Emissions Production Simulator and its successors (Anderson et al., 2004) as well as the First Order Fire Effects Model and its successors. It also represents just one of many sources of potential error in emissions calculations (Ottmar et al., 2008; Surawski et al., 2016). In this paper we therefore propose a different approach to analyze emissions data that reflects their

compositional nature. A well-principled methodological body to analyze compositional data has been developed in the past 30 years and this is an active area of statistical research so we have chosen to apply it to fire emissions data, in particular gas-phase emissions. Interestingly, it has been successfully applied in varied fields, but appears to have seldom been applied to combustion or emissions data (Bandeen-Roche, 1994; Billheimer, 2001; Buccianti & Pawlowsky-Glahn, 2006; Speranza et al., 2018). A recent analysis that ercognized

The relative nature of emissions data defines emissions data as compositional data which are coherently analyzed using CoDA (Aitchison, 1986; Barceló-Vidal et al., 2001; Lovell et al., 2015; Pawlowsky-Glahn et al., 2015b). CoDA methodology has three underlying principles. The first is scale invariance—vectors with proportional positive components represent the same composition, and form what is known as an equivalence class. This means changing the units should not change relative relationships between the parts nor affect results and scientific conclusions. The second is that inferences about subcompositions, i.e. smaller compositions formed from subsets of parts, must not contradict the inferences from the full composition (as in the example in Table 1). The third principle states that the order of the parts of the composition must not affect the inferences. While the initial work on CoDA explicitly assumed in the definition of a composition that the parts sum (are closed) to a constant total, theory has developed to show that this is only a particular representation of the data in a simplex and equivalent non-closed compositions carry the same relative information (Barcelo-Vidal & Martín-Fernández, 2016). So the “conservation of mass” argument presented earlier is not necessary for emissions data to be considered compositional as shown by Egozcue (2009). The structural relationship and interdependence between MCE and other emissions as shown in (3) still holds.

Aitchison (1982) showed that a meaningful approach to compositional data is to analyze log-ratios of the parts which carry the relative information. Aitchison (1986) defined the two basic operations of perturbation ($\oplus$, analogous to addition or translation with ordinary real-valued data)

$$\mathbf{z} = \mathbf{x} \oplus \mathbf{p} = C\left[x_1 \cdot p_1, \ldots, x_d \cdot p_d\right] \tag{5}$$

and power transformation ($\otimes$, analogous to multiplication by a scalar)

$$\mathbf{z} = \lambda \otimes \mathbf{x} = C\left[x_1^{\lambda}, \ldots, x_d^{\lambda}\right] \tag{6}$$

where $\mathbf{x}$ is the initial composition, $\mathbf{p}$ is a perturbation vector and λ is a constant. These operations are foundational to CoDA.

In CoDA today, in order to use familiar statistical techniques such as exploratory data analysis, linear regression, multivariate analysis of variance and other multivariate techniques, the mainstream approach is to transform the parts from the simplex to real numbers using isometric log-ratio (ilr) coordinates (van den Boogaart & Tolosana-Delgado, 2013; Egozcue & Pawlowsky-Glahn, 2005; Mateu-Figueras et al., 2011; Pawlowsky-Glahn et al., 2015b). The linear algebra theory supporting these transformations also provides the underpinnings for “standard” or “classical” statistics routinely used in the sciences (Graybill, 2002). Several texts describe the theory and methods of compositional data analysis (Aitchison, 1986; van den Boogaart & Tolosana-Delgado, 2013; Filzmoser et al., 2018; Pawlowsky-Glahn et al., 2015b; Pawlowsky-Glahn & Buccianti, 2011). Software to perform compositional data analysis is available, including the stand-alone point-and-click *CoDaPack* package (Thió-Henestrosa & Comas, 2016) (http://www.compositionaldata.com/codapack.php) and comprehensive libraries on the open-source R statistical computing system (R Core Team, 2018): *compositions* (van den

Boogaart & Tolosana-Delgado, 2013), *robCompositions* (Templ et al., 2011) and *zCompositions* (Palarea-Albaladejo et al., 2014; Palarea-Albaladejo & Martín-Fernández, 2015).

The above considerations borne in mind, the objectives of this manuscript are to re-analyze the Burling et al. (2010) emissions factors data to 1) determine if parts (individual gases) were proportional to each other (in place of correlated in the usual way), 2) determine if a compositional linear trend can be used to model the data as combustion efficiency changes (in place of ordinary linear regression using MCE), and 3) determine if the composition of the gases differed between fuel types using analysis of variance within a CoDA framework. We hope to demonstrate that the CoDA approach can shed as much or more light on the relationships in and between the emissions data by applying techniques consistent with the nature of the data instead of using simple linear regression with MCE thus avoiding artifacts derived from the very own nature of the data.

2 Statistical Methods

Burling et al. (2010) reported emission factors for 18 gases measured using an open-path FTIR spectrometer (Burling et al., 2010) and characteristics of the combustion (fuel moisture content, fuel consumption, MCE) for 65 laboratory fires (observations) in 15 different wildland fuel types. Each of the 65 observations comprised a vector $\mathbf{x}_i = [x_1 \ldots x_{18}]_i$ where parts $x_1 \ldots x_{18}$ were the measured emission factors for CO_2, CO, CH_4, C_2H_2, C_2H_4, C_3H_6, CH_3OH, HCOOH, CH_3COOH, HCHO, C_4H_4O, NH_3, NO, NO_2, HONO, HCN, HCl, and SO_2, respectively. In actuality, these 18 gases were a subcomposition of a much larger composition of gaseous and solid emissions sampled from these experimental fires by a variety of methods and instruments described elsewhere (Burling et al., 2010; Chang-Graham et al., 2011; Gilman et al., 2015;

Hosseini et al., 2010, 2013; Roberts et al., 2010; Veres et al., 2010; Warneke et al., 2011; Weise et al., 2015; Yokelson et al., 2013). The full composition for this experiment consisted of over 100 parts. The data set resulted from a laboratory study at the United States Forest Service's Missoula Fire Sciences Lab (USFS-FSL). This FSL study characterized smoke emissions related to prescribed burning of 15 different shrub and woodland fuel types from the southeastern and southwestern U.S. and described the composition of gas and particulate matter in detail. In the present paper, 18 gases were re-analyzed which had previously been measured using an open-path FTIR spectrometer (Burling et al., 2010) and then adjusted to field values using MCE (Yokelson et al., 2013) While it is possible to measure H_2O (gas) in smoke emissions using FTIR and while it is a significant product of combustion that can influence flame processes (Ferguson et al., 2013), emission factors for H_2O are not typically reported. The gas-phase data analyzed here are originally available as supplementary information to the Yokelson et al. (2013) paper. Furan (C_4H_4O) and hydrochloric acid (HCl) had one and four (1.5 and 6 %) instances of below-detection limit (BDL) values, respectively. In order to facilitate statistical analysis, the log-ratio EM algorithm included in the *zCompositions* package was used to impute the BDL values with realistically small values while accounting for their compositionality (Palarea-Albaladejo & Martín-Fernández, 2015). In the following, we introduce the basic compositional analyses conducted using these data which are analogous to those commonly conducted on emission data.

2.1 Summary statistics and proportionality associations

The data were closed and compositional summary statistics consisting of the center (geometric mean) and the variation array (Aitchison, 1986) were estimated. The center ($\bar{\mathbf{g}}$) is the closed vector of geometric means for each part estimated as

$$\hat{\mathbf{g}} = C\left[\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_D\right] \tag{7}$$

where $\hat{g}_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{1/n}$, $j = 1, 2, \ldots, D$. While there are different measures for the variability of compositional data (van den Boogaart & Tolosana-Delgado, 2013), a common summary is given by the variance $\left(\tau_{ij}\right)$ of the log-ratio of parts *i* and *j*; the variation matrix **V** is a *D* x *D* symmetric matrix containing elements estimated by

$$\hat{\tau}_{ij} = \mathrm{var}\left(\ln \frac{x_i}{x_j}\right) \tag{8}$$

where *i* and *j* range from 1 to *D* and var is the usual variance. The total (or metric) variance (Mvar) can be obtained from them as (Egozcue & Pawlowsky-Glahn, 2011; Pawlowsky-Glahn & Egozcue, 2001):

$$\mathrm{Mvar}\left(\mathbf{x}\right) = \sum_i \sum_j \hat{\tau}_{ij} = \sum_{i=1}^{D} \mathrm{var}\left[\mathrm{clr}_i\left(\mathbf{x}\right)\right] \tag{9}$$

where the centered log-transformation (clr) and its inverse are

$$\begin{aligned} &\mathrm{clr}\left(\mathbf{x}\right) = \left[\ln \frac{x_1}{g_m\left(\mathbf{x}\right)}, \ln \frac{x_2}{g_m\left(\mathbf{x}\right)}, \ldots, \ln \frac{x_D}{g_m\left(\mathbf{x}\right)}\right], \quad g_m\left(\mathbf{x}\right) = \left(\prod_{i=1}^{D} x_i\right)^{1/D} \\ &\mathrm{clr}^{-1}\left(\mathbf{x}\right) = C\left[\exp\left(\mathrm{clr}\left(\mathbf{x}\right)\right)\right] \end{aligned} \tag{10}$$

and the metric standard deviation (Mstd) is $\sqrt{\frac{\mathrm{Mvar}}{D-1}}$. In general, the smaller the values of $\hat{\tau}_{ij}$, the more proportional are the parts involved.

For non-compositional data, relationships between variables are ordinarily explored by examining correlation (parametric or nonparametric) between the variables. For compositional data, proportionality is the preferred measure to examine relationships between parts of a

composition in accordance with their relative scale (Aitchison, 1986; Lovell et al., 2015). As stated in Lovell et al (2015), *"measures of association produce results regardless of the data they are applied to-it is up to the analyst to ensure that the measures are appropriate to the data."* They further state that proportional relative abundances imply that the absolute abundances are proportional. Balance association (or b-association for short) was developed as a consistent statistical concept of proportionality (Egozcue et al., 2018). A measure of b-association, ϕ, has been defined and used to formulate a statistical hypothesis test of equality to ±1 of the slope coefficient of a major or standardized major axis regression model (Warton et al., 2006) relating one log-contrast of parts to another log-contrast. This is the so-called unitary slope test, with significance based on a standard F distribution (Egozcue et al., 2018; Lovell et al., 2015). Rejection of the hypothesis suggests that the data are not compatible with b-association (or proportionality) between the parts; however, it does not distinguish whether the slope is positive or negative. Following Egozcue et al. (2018), we rejected the hypothesis when the estimated slope was negative and reported the p-value as a minus sign (-).

2.2 Compositional linear trend analysis

Since its introduction *MCE* has frequently been correlated to single emission factors (*EF*) by linear regression $EF = \beta_0 + \beta_1 \left(MCE \right)$ (Ward & Hao, 1991) which assumed that the predictor variable *MCE* could be treated separately from the response variable *EF*. As shown above in eq. 3 this is not the case, so an alternative method to estimate emission factors developed for compositional data was applied instead. Deriving principal components from multivariate data is a common practice. They were used in von Eynatten et al. (2003) to develop the compositional linear trend method which we apply to our data using recently developed R code (Rockwell et

al., 2014). The basic idea is that compositional data can be projected onto the first principal component to produce a linear trend provided that the first principal component explains a large proportion of the total variance. The projected (or fitted) composition can be transformed back to the original units (Pawlowsky-Glahn et al., 2015a). We chose to apply this method because it has been wellestablished that the composition of smoke changes as the combustion efficiency changes (Byram, 1957). If the linear trend method works, it could potentially be applied to other data sets to predict composition.

When originally developed, the linear trend method was applied to geological data to model the compositional changes in granitic rocks as they weathered from fresh parent material. The unweathered parent material served as the starting point of the linear trend which is formulated in the simplex as

$$\mathbf{x} = \mathbf{a} \oplus \left(k \otimes \mathbf{p} \right) = \mathrm{C}\left[a_1 p_1^k, \ldots, a_d p_d^k \right] \tag{11}$$

and estimated by

$$\hat{\mathbf{x}} = \mathbf{a} \oplus \left(k \otimes \mathrm{clr}^{-1}\left[\mathbf{v}_1 \right] \right) \tag{12}$$

where **a** is the starting composition, *k* is the latent trend, **p** is a perturbation vector estimated by $\mathrm{clr}^{-1}\left(\mathbf{v}_1\right)$, $\mathbf{v}_1$ is the first eigenvector derived by noncentral principal component analysis of the original *n* observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ after they were adjusted to the starting point **a** by $\mathrm{clr}\left(\mathbf{x}_1\right) - \mathrm{clr}\left(\mathbf{a}\right), \ldots, \mathrm{clr}\left(\mathbf{x}_n\right) - \mathrm{clr}\left(\mathbf{a}\right)$. Noncentral principal component analysis indicates the variances were maximized relative to **a** instead of to the mean (von Eynatten, 2004). The full mathematical development can be found elsewhere (von Eynatten et al., 2003). The linear trend method describes changes within a compositional data set that are not attributed to variables

external to the composition. In the present study, we evaluated three compositions as the potential starting composition **a**: the highest (High) and lowest (Low) MCE compositions and the center composition of the data set. Linear regressions were not fit for CO and CO_2 since they form MCE. In order to compare the performance of the linear trend to the ordinary linear regression, common error measures were calculated using the observed and predicted emission factors for each gas for each linear regression and linear trend. The estimated values were scaled to the original units (Pawlowsky-Glahn et al., 2015a). The normalized mean absolute error (NMAE) and root mean squared error (RMSE) of the observed and estimated values were calculated using the *modStats* function in R (Carslaw, 2015; Carslaw & Ropkins, 2012). Since multivariate linear regression is the extension of linear regression to data with correlated response variables (Fox & Weisberg, 2018, 2019), we performed a multivariate linear regression with all trace gases except CO and CO_2 as the dependent variables and MCE as the predictor variable. The fitted values from the multivariate linear regression were identical to the individual linear regressions so the error measures were identical and are not presented. Multivariate linear regression, even though it takes into the account the correlation structure between the parts of the composition, was still subject to the fact that the MCE ratio was not independent of the other gases in the mixture. To examine the fits of the linear trend and multivariate linear regression for the entire composition, a coefficient of determination (R^2_{CLT}) for the linear trend (van den Boogaart & Tolosana-Delgado, 2013; Cayuela-Sánchez et al., 2019) and the squared multiple correlation coefficient (R^2_{LR}) for the multivariate linear regression (Mardia et al., 1979) were estimated. The metric standard deviation in the original emission factor units $\exp(\text{Mstd})$ and the root mean squared error of the multivariate linear regression (calculated as the square root of the mean of the trace of the residual matrix) estimated $RMSE_{CLT}$ and $RMSE_{LR}$, respectively.

2.3 Multivariate analysis of variance

Analysis of variance was used to test for differences in the composition of the gases according to fuel type. Note that because of experimental design deficiencies resulting in a singular design matrix, it was not possible to consider fuel type and region simultaneously. Egozcue and Pawlowsky-Glahn (2005) devised a procedure to obtain sets of ilr coordinates by sequential binary partitioning that can be used for them to represent comparisons between scientifically meaningful subsets of parts of a composition. These ilr coordinates known as compositional balances $\tilde{z}_k$ are defined as

$$\tilde{z}_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} \ln \frac{\left(x_{i_1} x_{i_2} \ldots x_{i_{r_k}}\right)^{1/r_k}}{\left(x_{j_1} x_{j_2} \ldots x_{j_{s_k}}\right)^{1/s_k}}, \quad k = 1, \ldots, D-1 \tag{13}$$

where the log-ratio compares the geometric mean of r_k parts in one subset with the geometric mean of s_k parts in another subset. Sequential binary partitioning of a composition containing D parts results in D-1 balances $\tilde{z}_k$, k = 1, …, D-1. The emissions data were then transformed into balance coordinates that partitioned the composition into various meaningful subsets of parts. The matrix used to define the subsets contains 1, -1, or a blank to indicate that the part is in subset 1 (numerator), subset 2 (denominator) or absent from the log-ratio, respectively. Once the data were transformed into balance coordinates, analysis of variance was used on them to test for differences in mean according to fuel type. Given the large number of statistical tests performed in this analysis, we chose to adjust the p-values to control for false discovery rate (Benjamini & Hochberg, 1995). Statistical significance was assessed at the usual 5% level.

3 Results and Discussion

3.1 Summary statistics

MCE ranged from 0.91 for the *lit* fuel type to over 0.98 for the *oas* fuel type. The geometric mean of MCE for the data set was 0.96. [Details of the fuel types and fuel consumption in the individual fires have been presented in the original publications (Burling et al., 2010; Hosseini et al., 2010, 2013)]. Numerical differences in the MCE values reported by Burling et al. (2010) and Yokelson et al. (2013) due to adjustment for field measurements are described in the latter reference. Unsurprisingly, the chemical compositions changed as combustion efficiency decreased from high to low (Table 2). Examination of the compositional makeup of the low MCE, geometric mean MCE, and high MCE revealed that relative abundance of all gases except CO_2, NO, and NO_2 increased as the MCE decreased. By definition, CO_2 increases and CO decreases as the MCE increases. Previously reported gas species associated with flaming combustion (higher MCE) included CO_2, NO, NO_2, HCl, SO_2 and HONO; those usually associated with smoldering combustion (lower MCE) include CO, CH_4, NH_3, HCN, C_3H_6, CH_3OH, CH_3COOH, and C_4H_4O (Burling et al., 2010; Goode et al., 2000). Of the species we measured, the remaining ones have been associated with both flaming and smoldering combustion.

Large variation in the emission factor for HCl for this data set was previously reported (Burling et al., 2010). The large log-ratio variance associated with HCl was also readily apparent in a biplot (Aitchison & Greenacre, 2002) produced from the first two principal components of the data set (not shown). This can be also seen with its clr-variance (Table 2). The clr-variance for HCl (1.50) was approximately 25 percent of the total variance (6.09). The clr-variances for

the remaining 17 parts were similar in size, yet small compared to HCl. While there were several low values of $\hat{\tau}_{ij}$ that suggested proportionality, only a few of the unitary slope tests were not statistically significant (Table 2) suggesting that some of the gases might be pairwise proportional. Potential pairwise proportionality relationships include: 1) propene (C_3H_6) with acetic acid (CH_3COOH), formaldehyde (HCHO), furan (C_4H_4O), ammonia (NH_3), and HCN, all of which have been associated with smoldering combustion; and 2) acetic acid with furan, nitrous acid (HONO) and hydrocyanic acid (HCN). The mean log-ratios of the five gases potentially proportional with propene ranged from -0.62 (HCN) to 1.61 (acetic acid) suggesting relatively less propene than HCN consistently in the smoke samples and relatively more propene than acetic acid, consistently. While the log-ratio variances for these five gases with propene were similar in size (0.25 to 0.33), the probabilities associated with the F-tests ranged considerably (0.07 to 0.94); the higher probability levels provide better support the potential proportionality of propene with the other hydrocarbons and less support for proportionality with the two N gases (NH_3, HCN). Note that while the log-ratio variance of CH_3OH with propene was lower than the log-ratio variances for CH_3COOH, HCHO, furan, ammonia and HCN, the slope test rejected potential pairwise proportionality. Other smoldering compounds for which results were compatible with proportionality included 3) formaldehyde with methanol, acetic acid, and furan; 4) ethene (C_2H_4) with methane and ammonia; and 5) HCN with methane and furan. Of the gases associated with flaming combustion, CO_2 was potentially proportional with NO_2, C_2H_2 and HONO. It was interesting to note that some gases normally associated with flaming show some level of proportionality with smoldering gases: HONO with acetic acid, SO_2 with CO and ammonia. Because of its large variability, HCl exhibited the lowest proportionality to any other gases in the composition; all tests were rejected since the slope values were negative (J.J.

Egozcue et al., 2018). All log-ratio means for HCl were negative which indicated that the proportion of HCl in the compositions was less than the proportions of the other gases.

Table 2. Variation array and center of smoke emissions. The upper right triangular matrix contains estimates of log-ratio variance $\left(\hat{\tau}_{ij}\right)$ and (probability of F-statistic for slope test of proportionality). The minus (-) sign denotes a negative slope and rejection of the null hypothesis of perfect proportionality. Shading indicates that the hypothesis was not rejected at the 5% significance level suggesting potential proportionality between parts. The lower left triangular matrix contains the log-ratio means. The three bottom rows are the compositional makeup (as proportions) for the lowest, highest, and center MCE values (x10^3).

	CO_2	CO	CH_4	C_2H_2	C_2H_4	C_3H_6	CH_3OH	HCOOH	CH_3COOH	HCHO	C_4H_4O	NH_3	NO	NO_2	HONO	HCN	HCl	SO_2
CO_2		0.12 (0.00)	0.37 -	0.50 (0.11)	0.43 -	0.69 -	0.65 -	0.92 -	0.73 -	0.71 -	0.63 -	0.24 (0.02)	0.06 (0.00)	0.16 (0.23)	0.23 (0.11)	0.95 -	2.88 (0.00)	0.14 (0.00)
CO	-3.33		0.18 -	0.33 (0.00)	0.20 -	0.41 -	0.33 -	0.58 -	0.44 -	0.36 -	0.33 -	0.17 (0.02)	0.22 (0.00)	0.21 (0.00)	0.21 (0.00)	0.54 -	2.66 (0.00)	0.10 (0.29)
CH_4	-7.05	-3.72		0.36 (0.00)	0.08 (0.45)	0.17 (0.00)	0.17 (0.03)	0.51 -	0.31 -	0.27 -	0.35 -	0.15 (0.03)	0.53 -	0.38 -	0.32 -	0.29 (0.00)	2.92 -	0.18 -
C_2H_2	-9.61	-6.28	-2.56		0.28 (0.00)	0.64 -	0.70 -	1.05 -	0.88 -	0.67 -	0.85 -	0.42 (0.00)	0.63 -	0.67 -	0.48 (0.00)	0.87 -	2.78 (0.00)	0.44 (0.00)
C_2H_4	-8.28	-4.95	-1.23	1.33		0.12 (0.00)	0.19 (0.15)	0.49 -	0.33 -	0.24 (0.01)	0.35 -	0.18 (0.15)	0.56 -	0.42 -	0.27 -	0.30 (0.00)	3.13 -	0.21 -
C_3H_6	-9.43	-6.10	-2.37	0.18	-1.14		0.16 (0.04)	0.48 (0.03)	0.29 (0.94)	0.25 (0.49)	0.33 (0.81)	0.30 (0.08)	0.86 -	0.58 -	0.43 -	0.26 (0.07)	3.47 -	0.40 -
CH_3OH	-8.48	-5.15	-1.43	1.13	-0.20	0.94		0.21 (0.00)	0.07 (0.00)	0.08 (0.11)	0.17 (0.02)	0.35 -	0.83 -	0.44 -	0.37 -	0.16 (0.00)	3.31 -	0.31 -
HCOOH	-10.11	-6.78	-3.05	-0.50	-1.83	-0.68	-1.63		0.13 (0.00)	0.12 (0.00)	0.22 (0.01)	0.74 -	1.14 -	0.61 -	0.49 (0.00)	0.44 (0.56)	3.81 -	0.59 -
CH_3COOH	-7.82	-4.49	-0.76	1.79	0.47	1.61	0.66	2.29		0.09 (0.27)	0.20 (0.85)	0.52 -	0.94 -	0.43 -	0.36 (0.35)	0.28 (0.09)	3.53 -	0.38 -
HCHO	-8.43	-5.10	-1.38	1.18	-0.15	1.00	0.05	1.68	-0.61		0.17 (0.28)	0.51 -	0.93 -	0.48 -	0.36 -	0.28 (0.01)	3.34 -	0.41 -
C_4H_4O	-10.17	-6.84	-3.12	-0.56	-1.89	-0.75	-1.69	-0.06	-2.35	-1.74		0.45 -	0.81 -	0.52 -	0.46 -	0.33 (0.14)	3.51 -	0.38 -
NH_3	-8.04	-4.71	-0.99	1.57	0.24	1.38	0.44	2.07	-0.22	0.39	2.13		0.30 (0.00)	0.32 -	0.31 -	0.44 (0.00)	2.93 -	0.15 (0.13)
NO	-6.70	-3.37	0.35	2.91	1.58	2.73	1.78	3.41	1.12	1.73	3.47	1.34		0.28 (0.00)	0.34 (0.00)	1.14 -	3.10 -	0.21 (0.00)

NO_2	-7.85	-4.52	-0.79	1.76	0.44	1.58	0.63	2.26	-0.03	0.58	2.32	0.19	-1.15		0.19 (0.59)	0.72 -	2.87 -	0.18 (0.00)
HONO	-8.87	-5.54	-1.82	0.74	-0.59	0.55	-0.39	1.24	-1.05	-0.44	1.30	-0.83	-2.17	-1.02		0.73 -	3.33 -	0.21 (0.02)
HCN	-10.05	-6.72	-2.99	-0.44	-1.76	-0.62	-1.56	0.06	-2.23	-1.62	0.12	-2.01	-3.35	-2.20	-1.17		3.49 -	0.48 -
HCl	-10.46	-7.13	-3.41	-0.85	-2.18	-1.04	-1.98	-0.36	-2.65	-2.03	-0.29	-2.42	-3.76	-2.62	-1.59	-0.42		3.01 -
SO_2	-7.74	-4.41	-0.69	1.87	0.54	1.68	0.74	2.37	0.07	0.69	2.43	0.30	-1.04	0.11	1.13	2.30	2.72	
Low	900.	87.4	2.960	0.084	0.565	0.227	1.160	0.375	1.950	1.570	0.345	0.460	1.020	0.214	0.170	0.382	0.036	0.926
Center	961.	34.4	0.830	0.064	0.243	0.077	0.199	0.039	0.387	0.210	0.037	0.309	1.180	0.375	0.135	0.042	0.027	0.417
High	981.	15.5	0.287	0.020	0.064	0.017	0.060	0.018	0.181	0.063	0.016	0.110	1.610	0.316	0.093	0.005	0.001	0.363
clr(variance)	0.29	0.21	0.21	0.35	0.22	0.27	0.24	0.35	0.28	0.26	0.28	0.24	0.36	0.26	0.25	0.32	1.50	0.22

3.2 Compositional linear trend as an alternative to MCE linear regression

For all gases except CO_2, NO, and NO_2, the proportions decreased from the Low MCE fire to the High MCE fire (Table 2, Table 3) which is consistent with previously reported findings (Burling et al., 2010) suggesting the possibility of fitting a compositional linear trend (Eq. (11)). The three possible starting compositions (**a**) for the linear trend and the estimated perturbation vector (**p**) for each are contained in Table 3. The goodness of fit of the linear trend to the data (determined by the first eigenvalue λ_1 as a percentage of the total variation) was 72.9, 77.5, and 50.1 percent for the High, Low and GM starting points, respectively. Of the three linear trends, using the Low MCE composition as the starting point produced the smallest absolute errors, lowest RMSE, and highest correlation between the observed and predicted values for 14 of the 18 gases (Table 4). While Low MCE produced the lowest mean bias (NMB) for 8 of the gases and GM for 6 gases, NMB was similar for several gases for all three linear trends. Using the GM as the starting point provided a better fit for NO_2, HONO and HCl. There was no correlation between observed and predicted values for C_2H_2 from any of the three linear trends. The observed and predicted values for NO_2 and HCl were negatively correlated for the Low MCE trend.

When compared to the fitted ordinary linear regressions for each gas, the Low MCE compositional trend had smaller errors (NMAE) for 12 of the 18 gases, lower RMSE for 8 of the 18 and higher correlation (r) for 13 of the 18 gases (Table 5). Overall, the geometric mean NMAE for the CLT (0.26) was less than LR (0.30) indicating the linear trend estimates had less error than the linear regression estimates; however, the CLT was negatively biased. Bias for linear regression was 0 in all cases because the residuals $\left(\hat{y}_i - y_i\right)$ sum to 0 for every linear

regression with an intercept term (Draper & Smith, 1981). Mean RMSE for the CLT and LR were equal. The correlations between observed and estimated values for most of the trace gases were similar between the linear trend and the linear regression models. For the linear trend, correlation between observed and predicted values was not significant for C_2H_2 and NO; for linear regression there was no significant correlation between observed and predicted values for NO_2 and HCl. In some cases (such as HONO), the EF was fairly constant (Figure 1). For 9 of the 16 gases, correlation for the CLT was larger than for the LR. For the overall measures, the coefficient of determination R^2_{CLT} for the Low MCE linear trend (0.283) was lower than the comparable measure R^2_{LR} for the multivariate linear regression (0.933). It is interesting to note that R^2_{CLT} GM linear trend was 0.501. Because R^2_{CLT} is a measure of the entire composition projected onto the first principal component, parts of the composition (gases) more strongly associated with other principal components would contribute to the lack of fit of the linear trend. Preliminary analysis of the data using a biplot (Aitchison & Greenacre, 2002) suggested that HCl was not strongly associated with the first principal component (not shown) which decreased the amount of variability that the linear trend would account for thus reducing its predictive ability. Recall that the clr-variance for HCl was nearly 25% of the total variance. The fitted values of HCL by both the linear trend and the linear regression were relatively constant (Figure 1). The $RMSE_{CLT}$ ranged from 0.426 to 0.511 for the three starting points which was much smaller than $RMSE_{LR}$ (2.658). It is important to note that because compositional data are restricted to positive upper orthant, predictions from the linear trend were always positive unlike predictions from the linear regression. The predictions from the linear regression in Figure 1 are the fitted values, not predictions made outside the range of the data. The fitted linear regressions for 6 of the 18 gases

produced values below zero. Scatterplots for all 18 gases are available in the supplementary information. Generally, the CLT performed equal to or superior to the LR based on MCE. This coupled with the fact that the compositional nature of emissions data were analyzed using techniques appropriate to the type of data indicates the value of this analytical approach.

The data set used to demonstrate this technique was one that was readily available to the authors. There are at several compilations of emission factor data that could be used to explore if smoke composition changes linearly as efficiency of a fire changes (Akagi et al., 2011; Lincoln et al., 2014; Yokelson et al., 2013). If a linear trend can be successfully fit for a larger data set producing better goodness of fit measures, such a linear trend could be used to reliably estimate emissions of gases not typically measured if the log-ratio variance of the parts is relatively low.

3.3 Testing the effects of fuel type on compositional data

In the original report (Burling et al., 2010), effects of fuel type on subsets of the emissions were inferred from bar plots and error bars, but no formal hypothesis testing was presented. The fuel types were representative of major local vegetation types in the southwestern and southeastern U.S. The pine litter fuel type was the sole type composed of only dead pine needles (*Pinus* spp.) and branches. All other fuel types included live foliage and branches in addition to dead fuels. In the present study compositional balances 1 to 7 were designed to test meaningful observations reported in Burling et al. (2010); whereas balances 8-17 were necessary to obtain the full projection of the compositions from the simplex into real-valued ilr coordinates but are not defined according to any particular scientific relevance (see supplemental data). The intercept term, which is the mean for the 1-year rough fuel type due to ANOVA parameterization, was statistically significant for 5 of the 7 balances of interest (Table 6). For the

significant intercepts, a positive value indicated that the numerator of the balance was on average relatively larger than the denominator, while a negative value indicated it was relatively smaller. For example, there was relatively less N compared to the other compounds (hydrocarbons, C oxides, etc.; balance 1) but relatively more NOx than other N compounds (balance 2) for the 1-year rough fuel type. The estimated effect for a fuel type is the sum of the intercept and the fuel type value. Thus the estimate of balance 1 for pine litter was -2.02 (-1.18 + -0.84); because the fuel type value was significant, the composition of the emissions for pine litter had relatively even less N compared to other compounds; similarly the oak savanna and woodland fuel types had relatively more N compared to other compounds since the balance estimate was close to -0.2 for these fuels. For balance 1, the relative amount of N versus the other compounds for the other fuel types was not significantly different from the 1 yr rough. In addition to the smoke containing relatively more NOx compared to the other N compounds (balance 2 intercept), eight of the southwestern fuel types had relatively more NOx compared to other N compounds than five of the six southeastern fuel types. Not surprisingly, there were relatively more C oxides than organic C compounds (balance 5). We observed CH_4 in relatively higher quantities compared to the non-methane organic compounds (NMOC) (balance 6). While overall NMHC and OVOC relative abundances were statistically comparable for the 1-year fuel type (balance 7 is not significantly different from zero), a significant negative difference with respect to this baseline was concluded for seven of the other fuel types, suggesting that relatively more OVOC compared to NMHC was observed on average for those fuel types. The relative amount of NH_3 to NO_x (balance 3) did not differ significantly between all fuel types except for the oak savanna fuel type. The overall amounts of HCl and SO_2 observed were much smaller in comparison to the

quantities of C compounds (balance 4); however, eight of the fuel types significantly reduced this difference in relative amounts.

Table 3. Starting point compositions (a) in original units (g kg^{-1}) and estimated perturbation vectors (p) for a compositional linear trend fit to a data set of smoke emissions from wildland fuels. The starting points after closure are contained in Table 2.

Gas	Starting point (**a**)			Perturbation (**p**)		
	Low MCE (0.911)	Center	High MCE (0.984)	Low MCE	Center	High MCE
CO_2	1584.6780	1739.8798	1745.3486	0.0716	0.0572	0.0722
CO	153.7918	62.2174	27.5268	0.0552	0.0558	0.0569
CH_4	5.2066	1.5030	0.5105	0.0500	0.0521	0.0525
C_2H_2	0.1473	0.1167	0.0358	0.0654	0.0583	0.0510
C_2H_4	0.9946	0.4395	0.1144	0.0553	0.0503	0.0491
C_3H_6	0.3993	0.1402	0.0299	0.0515	0.0472	0.0464
CH_3OH	2.0335	0.3601	0.1060	0.0434	0.0472	0.0506
HCOOH	0.6595	0.0708	0.0317	0.0380	0.0442	0.0572
CH_3COOH	3.4240	0.7000	0.3222	0.0448	0.0457	0.0574
HCHO	2.7553	0.3799	0.1128	0.0408	0.0471	0.0507
C_4H_4O	0.6071	0.0666	0.0282	0.0387	0.0465	0.0563
NH_3	0.8104	0.5600	0.1948	0.0627	0.0539	0.0534
NO	1.7903	2.1428	2.8557	0.0736	0.0572	0.0785
NO_2	0.3762	0.6791	0.5613	0.0795	0.0548	0.0680
HONO	0.2986	0.2439	0.1652	0.0651	0.0503	0.0651
HCN	0.6733	0.0754	0.0086	0.0386	0.0468	0.0389
HCl	0.0631	0.0497	0.0025	0.0689	0.1327	0.0273
SO_2	1.6302	0.7546	0.6450	0.0569	0.0528	0.0686

Table 4. Ordinary measures of fit based on observed and predicted emission factors for compositional linear trends that started at the High, Low or geometric mean (GM) value of MCE for gases associated with smoke from wildland fire. Highlighted values indicate the best value for each measure by gas.

Gas	NMB			NMAE[1]			RMSE			r		
	High	Low	GM	High	Low	GM	High	Low	GM	High	Low	GM
CO_2	0.00	0.00	0.00	0.01	0.01	0.01	26.16	16.65	26.60	0.94	0.98	0.94
CO	-0.07	0.04	-0.05	0.26	0.20	0.24	22.35	16.26	22.77	0.38	0.78	0.08^{N}
CH_4	-0.15	-0.04	-0.14	0.40	0.30	0.42	1.08	0.68	1.10	0.32	0.79	0.24
C_2H_2	-0.23	-0.24	-0.24	0.52	0.53	0.54	0.15	0.15	0.15	0.16^{N}	0.12^{N}	-0.10^{N}
C_2H_4	-0.21	-0.19	-0.18	0.50	0.37	0.47	0.49	0.42	0.47	0.17^{N}	0.58	0.30
C_3H_6	-0.28	-0.26	-0.23	0.59	0.41	0.55	0.17	0.14	0.17	0.24	0.74	0.32
CH_3OH	-0.29	-0.12	-0.24	0.63	0.27	0.55	0.51	0.23	0.50	0.25	0.92	0.29
HCOOH	-0.40	-0.11	-0.32	0.73	0.28	0.68	0.14	0.06	0.14	0.17^{N}	0.92	0.27
CH_3COOH	-0.33	-0.21	-0.26	0.67	0.34	0.60	1.11	0.66	1.08	0.20^{N}	0.87	0.28
HCHO	-0.31	-0.07	-0.27	0.65	0.26	0.59	0.58	0.24	0.56	0.19^{N}	0.91	0.32
C_4H_4O	-0.33	0.02	-0.27	0.63	0.23	0.56	0.12	0.03	0.12	0.22^{N}	0.96	0.27
NH_3	-0.12	-0.12	-0.09	0.37	0.33	0.36	0.30	0.27	0.29	0.29	0.43	0.12^{N}
NO	0.00	-0.01	-0.02	0.17	0.16	0.18	0.48	0.47	0.50	0.31	0.18^{N}	-0.21^{N}
NO_2	-0.08	-0.17	-0.07	0.32	0.37	0.32	0.32	0.39	0.31	0.18^{N}	-0.43	0.21^{N}
HONO	-0.16	-0.16	-0.09	0.41	0.35	0.35	0.16	0.14	0.14	-0.02^{N}	0.58	0.41
HCN	-0.38	-0.14	-0.37	0.73	0.36	0.68	0.17	0.08	0.17	0.27	0.92	0.28
HCl	-0.17	-0.47	-0.02	0.42	0.77	0.35	0.08	0.14	0.06	0.85	-0.32	0.90
SO_2	-0.05	-0.01	-0.05	0.27	0.20	0.28	0.31	0.20	0.30	-0.03^{N}	0.75	0.25

1. NMB is normalized mean bias, NMAE is normalized mean average error, RMSE is root mean squared error, r is Pearson correlation coefficient – N indicates that r is not significantly different from 0 at 5% significance level based on t-test.

$$\mathrm{NMB} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)\Big/\bar{y}$$

$$\mathrm{NMAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|\Big/\bar{y}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}$$

where $y_i, \hat{y}_i, \bar{y}$ are observed, predicted, and mean emission factor (g/kg); n is number of observations.

Table 5. Measures of goodness of fit[1] of estimates for the Low MCE compositional linear trend (CLT) and linear regression (LR) with observed smoke emissions from wildland fire.

	NMAE		NMB	RMSE		r	
	CLT	LR	CLT	CLT	LR	CLT	LR
CO_2	0.01		0.00	16.65		0.98	
CO	0.20		0.04	16.26		0.78	
CH_4	0.30	0.30	-0.04	0.68	0.66	0.79	0.80
C_2H_2	0.53	0.55	-0.24	0.15	0.14	0.12^N	0.35
C_2H_4	0.37	0.34	-0.19	0.42	0.39	0.58	0.59
C_3H_6	0.41	0.45	-0.26	0.14	0.12	0.74	0.69
CH_3OH	0.27	0.41	-0.12	0.23	0.26	0.92	0.86
HCOOH	0.28	0.55	-0.11	0.06	0.08	0.92	0.81
CH_3COOH	0.34	0.50	-0.21	0.66	0.70	0.87	0.76
HCHO	0.26	0.42	-0.07	0.24	0.30	0.91	0.85
C_4H_4O	0.23	0.52	0.02	0.03	0.06	0.96	0.84
NH_3	0.33	0.32	-0.12	0.27	0.25	0.43	0.47
NO	0.16	0.15	-0.01	0.47	0.45	0.18^N	0.35
NO_2	0.37	0.31	-0.17	0.39	0.31	-0.43	0.24^N
HONO	0.35	0.41	-0.16	0.14	0.14	0.58	0.34
HCN	0.36	0.51	-0.14	0.08	0.09	0.92	0.83
HCl	0.77	0.82	-0.47	0.14	0.13	-0.32	0.04^N
SO_2	0.20	0.21	-0.01	0.20	0.21	0.75	0.73
Mean	0.26	0.30	-0.13	0.33	0.33		

1. NMB is normalized mean bias, NMAE is normalized mean average error, RMSE is root mean squared error, r is Pearson correlation coefficient – N indicates that r is not significantly different from 0 at probability = 0.05 based on t-test.

$$\mathrm{NMB} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)\Big/\bar{y}$$

$$\mathrm{NMAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|\Big/\bar{y}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}$$

where $y_i, \hat{y}_i, \bar{y}$ are observed, predicted, and mean emission factor (g/kg); *n* is number of observations. Mean value of NMAE calculated as geometric mean, mean RMSE calculated square root of mean of squared RMSE.

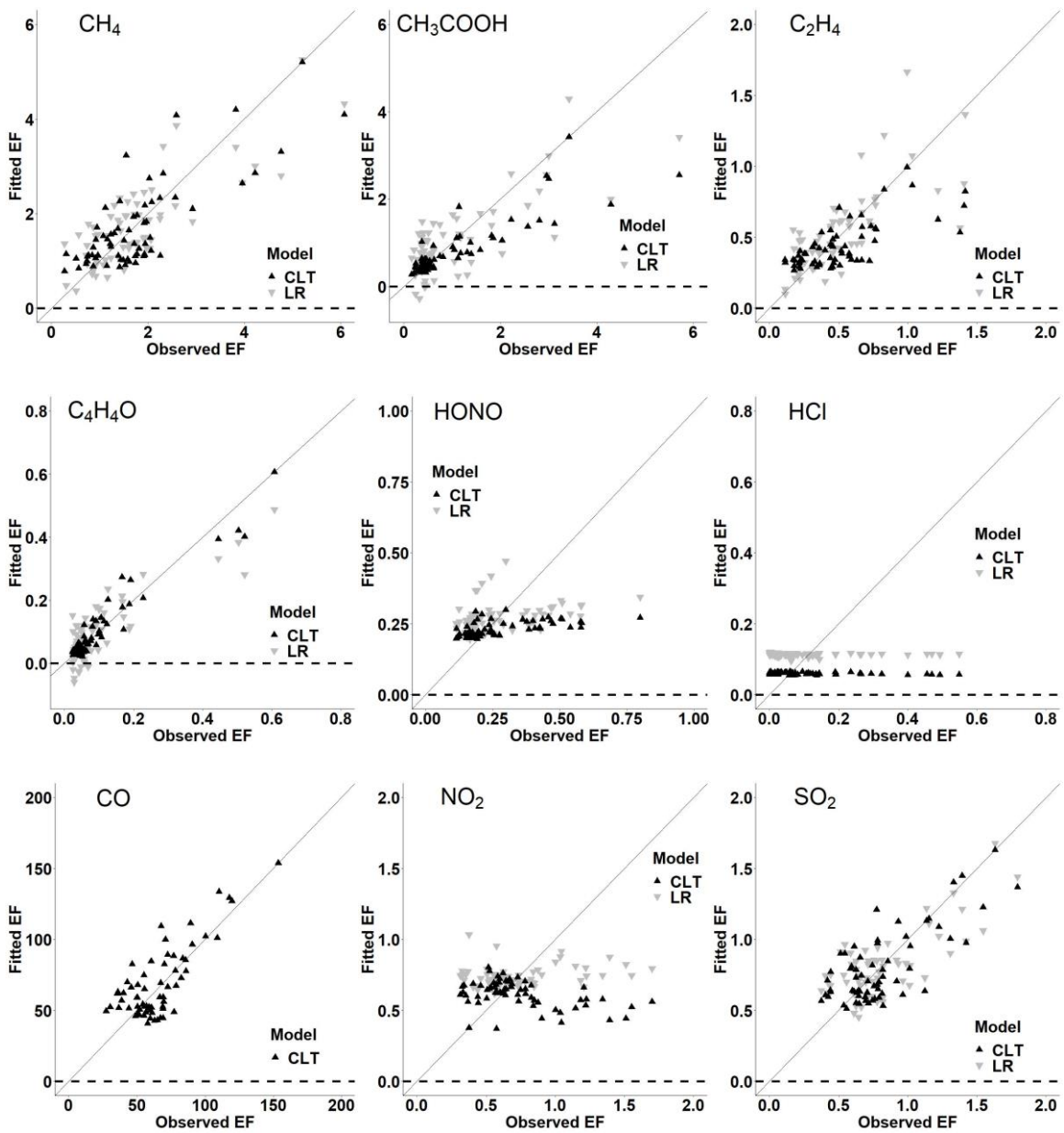


Figure 1. Comparison of observed emission factors (EF) for several wildland fire gases with fitted EF from a compositional linear trend (CLT) and a linear regression (LR) using modified combustion efficiency as the predictor variable.

Table 6. Significance of mean differences in compositional balances of smoke emissions from wildland fire according to fuel type. * = p-value < 0.05. Significant differences are shown in grey text. Dark grey denotes the balances of interest, light grey denotes the other balances. P-values adjusted to control for false discovery rate.

ID	Balance	Intercept (1 year rough)[1]	2 year rough	California sagebrush	Ceanothus	Chamise/scrub oak	Coastal sage scrub	Chipped understory hardwood	Pine litter	Manzanita	Maritime chaparral	Masticated mesquite	Oak savanna	Oak woodland	Pocosin	Understory hardwood
1	N vs other	-1.18*	0.12	-0.08	0.25	0.02	-0.01	0.08	-0.84*	0.15	0.10	0.52	0.97*	0.99*	-0.39	0.39
2	NOx vs other N	1.17*	0.02	0.48	1.06*	0.70*	0.58*	0.99*	-0.20	1.22*	0.68*	1.01*	1.61*	0.88*	0.53	0.50
3	NH_3 vs NO_x	-0.34	-0.04	0.08	-0.55	-0.44	-0.11	-0.43	0.09	-0.55	0.02	-0.26	-0.94*	-0.23	-0.45	-0.39
4	HCl & SO_2 vs C	-3.61*	0.59	2.51*	2.10*	0.79	1.96*	1.80*	0.89	2.67*	2.97*	2.26*	0.86	0.15	1.94*	0.42
5	C oxide vs OC	8.17*	0.01	1.01*	1.10*	0.83*	1.34*	1.57*	-0.23	1.83*	1.20*	1.25*	1.75*	1.30*	0.44	0.59
6	CH_4 vs NMOC	1.65*	0.04	0.42*	-0.01	-0.02	0.58*	0.77*	-0.14	0.09	0.55*	0.63*	0.27	0.35*	-0.15	-0.14
7	NMHC vs OVOC	0.40	-0.90*	0.39	-0.97*	-0.74*	0.55*	-0.27	-1.90*	-0.48	0.37	-0.47	-0.95*	-0.06	-1.64*	-1.27*
8	*Alkene vs alkyne*	*1.08**	*-0.06*	*-0.26*	*-0.18*	*0.03*	*-0.23*	*-0.11*	*0.05*	*-0.16*	*-0.12*	*0.00*	*0.04*	*-0.02*	*-0.06*	*-0.04*
9	*Furan vs OVOC*	*-1.80**	*0.07*	*0.54**	*0.38*	*0.45**	*0.44**	*0.57**	*0.74**	*0.96**	*0.67**	*0.24*	*0.59**	*0.70**	*0.20*	*0.37*
10	*Formaldehyde vs OVOC*	*0.44**	*-0.28*	*-0.02*	*-0.15*	*-0.08*	*0.07*	*-0.05*	*-0.17*	*-0.01*	*0.04*	*-0.18*	*-0.37**	*-0.17*	*-0.22*	*-0.12*
11	*HCl vs SO_2*	*-2.99**	*0.28*	*1.84**	*2.12**	*0.78*	*1.28**	*1.52**	*0.96*	*2.05**	*2.52**	*1.37**	*-0.61*	*-0.90*	*1.82**	*-0.08*
12	*NO vs NO_2*	*0.48**	*-0.04*	*0.57**	*0.17*	*0.48**	*0.70**	*0.08*	*0.07*	*0.54**	*0.50**	*0.37**	*0.65**	*0.76**	*-0.59**	*0.04*
13	*CH3 vs HCOOH*	*1.47**	*0.12*	*0.30*	*-0.13*	*0.07*	*0.26*	*0.28*	*0.02*	*-0.03*	*0.46*	*0.37*	*-0.08*	*0.27*	*-0.10*	*0.12*
14	*CH_3COOH vs CH_3OH*	*-0.53**	*-0.06*	*0.24**	*-0.04*	*0.07*	*0.20*	*0.25**	*0.11*	*0.05*	*0.18*	*0.25**	*-0.19*	*0.11*	*-0.28**	*-0.11*
15	*HONO vs HCN*	*0.70**	*-0.38*	*0.19*	*0.62**	*0.73**	*0.05*	*0.19*	*-1.40**	*0.48**	*-0.12*	*-0.13*	*0.82**	*0.30*	*0.30*	*0.35*
16	*CO vs CO_2*	*-2.20**	*0.08*	*-0.11*	*-0.15*	*-0.05*	*-0.07**	*-0.34**	*0.37*	*-0.16**	*-0.22**	*-0.25**	*-0.59**	*-0.46**	*-0.24**	*-0.26**
17	*Acetylene vs Propene*	*-0.18*	*-0.40*	*0.60*	*0.19*	*-0.33**	*1.14*	*-0.18*	*-0.70*	*0.52*	*-0.11*	*0.10*	*0.11*	*-0.30*	*-0.18*	*-0.24**

1. Standard ANOVA parameterization on balance ilr coordinates–the intercept term estimates the mean of the 1 year rough fuel type and the value shown for the other types is the difference between the fuel type mean and the intercept (*1yr* fuel mean).

4 Conclusions

Smoke emissions data are inherently multivariate and relative in nature. While this has been recognized for many years, the statistical techniques commonly used to analyze the data ignore these features. This not only applies to the composition of the emissions, but also to the different fuel types which burn to produce the emissions. Since emissions and fuel composition represent parts of a whole, the measured values of the individual parts (elements, chemical species, fuel component, etc.) are intrinsically not independent from each other. The measured values are relative and are only meaningful in relation to each other. Such constraints violate many of the underlying assumptions of ordinary statistical methods. Alternatively, compositional data analysis as a well-developed body of statistical methodology provides models and methods equivalent to traditional ones yet accounts for these special constraining features of relative data. The approach has been used for decades to analyze analogous types of data in the geosciences (Buccianti et al., 2006) and, more recently, in other disparate areas such as molecular biology to analyze sequencing data (Quinn et al., 2018)or physical activity epidemiology for the analysis of daily time-use patterns (Chastin et al., 2015; McGregor et al., 2019). While the statistical theory may be unfamiliar and not typically taught in most statistics courses, recent publications and software have made the use of these techniques both feasible and accessible.

The expression of emissions data as ratio data has long been reported. Even in this early work, conversion of the composition of emissions between different units by simple multiplication was presented, reinforcing the idea that emissions data form an equivalence class. Transformation of data using the arc-sine and square root transformations for count data to enable use of the normal distribution or to stabilize variance in linear regression and the log-odds

transformation used in logistic regression are familiar statistical techniques routinely used in the atmospheric sciences and other fields. Linear transformation of data is used to code data to simplify analysis for a variety of statistical calculations. The compositional data approach based on log-ratio coordinates essentially maps the data onto the ordinary real space so that familiar statistical techniques can be appropriately applied. This approach matches analysis techniques to the data type thus reducing the possibility of the reporting of spurious results that may or may not reflect the underlying relationships.

The linear regression approach as it has been typically applied uses one portion of the composition (expressed as MCE) to predict other parts of the composition, which ignores the intrinsic interplay between smoke emissions and can produce predictions beyond the domain of their possible values, e.g. negative values. The compositional data analysis approach recognizes the inherently positive-valued nature of the data thus eliminating the need for an analyst to ignore or discount when a fitted model produces negative values. Robust methods have been developed to allow inclusion of parts of a composition that are known to exist but fall below detection limits thus permitting a more complete analysis. We have illustrated how the use of a compositional linear trend to describe changes in the composition of smoke emissions as combustion efficiency changed yielded predicted emission factors with error (difference between observed and predicted) comparable to and, in some cases, superior to predictions from linear regression models that used modified combustion efficiency as a predictor variable. Moreover, the use of compositional balances to form log-ratios contrasting subsets of parts of interest enabled the use of analysis of variance and hypothesis testing to examine differences in meaningful trade-offs between smoke components with more formal rigor than was previously presented by respecting the very relative nature of the data as derived from underlying natural

laws like conservation of mass. We have definitively shown that fuel type affected several different ratios of groups of emissions and are assured that the results are not an artifact of the analysis which can then be used to make various inferences and decisions. Near the end of the article by Burling et al. (2010), there is discussion and hypothesis formation about the impacts of wildland fuel management activities and influence of ocean proximity on the composition of observed emissions. These impacts and hypotheses could be rigorously tested with the techniques presented here. More complex analyses of log-ratios of gas pairs or groupings as functions of external fire behavior variables such fire intensity (heat release rate) and flame residence time are possible. More rigorous time series and spatial analysis to examine aging smoke composition within the smoke plume and in response to atmospheric processes are possible with compositional data. It is our view and recommendation that future analysis of the composition of smoke emissions and other mixtures of atmospheric pollutants as well as general atmospheric composition should consider the use of compositional data analysis methods to provide more statistically rigorous and consistent results.

Acknowledgments, Samples, and Data

The data used in this paper resulted from projects the DOD/DOE/EPA Strategic Environmental Research and Development Program projects RC-1648 and 1649. The senior author appreciates the guidance and R scripts provided by Prof. Girty at San Diego State University to estimate linear trends by perturbation. J. P.-A. was supported by the Spanish Ministry of Science, Innovation and Universities under the project CODAMET (RTI2018-095518-B-C21, 2019-2021). The data used in this study have been previously published and are available in the original publications. DRW conceived the initial manuscript (70 percent) and performed the bulk of the data analysis. JPA provided statistical guidance and compositional data

expertise and contributed 20 percent of the manuscript. TJJ and HJ were extensively involved in the study that provided the data. TJJ provide smoke emissions expertise and HJ provided combustion expertise. The authors declare that they have no conflict of interest. The use of trade or firm names in this publication is for reader information and does not imply endorsement by the U.S. Department of Agriculture of any product or service.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*(2), 139–177.

Aitchison, J. (1986). *The statistical analysis of compositional data*. London ; New York: Chapman and Hall.

Aitchison, J. (2003). A concise guide to compositional data analysis. In *CDA Workshop, Girona*.

Aitchison, J., & Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *51*(4), 375–392. https://doi.org/10.1111/1467-9876.00275

Akagi, S. K., Yokelson, R. J., Wiedinmyer, C., Alvarado, M. J., Reid, J. S., Karl, T., et al. (2011). Emission factors for open and domestic biomass burning for use in atmospheric models. *Atmospheric Chemistry and Physics*, *11*(9), 4039–4072. https://doi.org/10.5194/acp-11-4039-2011

Amaral, S. S., de Carvalho, J. A., Jr., Costa, M. A. M., Neto, T. G. S., Dellani, R., & Leite, L. H. S. (2014). Comparative study for hardwood and softwood forest biomass: Chemical characterization, combustion phases and gas and particulate matter emissions. *Bioresource Technology*, *164*, 55–63. https://doi.org/10.1016/j.biortech.2014.04.060

Anderson, G. K., Sandberg, D. V., & Norheim, R.A. (2004, January). Fire Emission Production Simulator (FEPS) User's Guide 1.0. USDA Forest Service, Pacific Northwest Research Station.

Andreae, M. O., & Merlet, P. (2001). Emission of trace gases and aerosols from biomass burning. *Global Biogeochemical Cycles*, *15*(4), 955–966. https://doi.org/10.1029/2000GB001382

Bandeen-Roche, K. (1994). Resolution of additive mixtures into source components and contributions: a compositional approach. *Journal of the American Statistical Association*, *89*(428), 1450–1458. https://doi.org/10.1080/01621459.1994.10476883

Barcelo-Vidal, C., & Martín-Fernández, J.-A. (2016). The Mathematics of Compositional Analysis. *Austrian Journal of Statistics*, *45*(4), 57. https://doi.org/10.17713/ajs.v45i4.142

Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01* (Vol. CD-ROM, p. 20). Cancun, MX: International Association for Mathematical Geosciences. Retrieved from http://ima.udg.edu/~barcelo/index_archivos/Cancun.pdf

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Billheimer, D. (2001). Compositional receptor modeling. *Environmetrics*, *12*(5), 451–467. https://doi.org/10.1002/env.472

van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing compositional data with R*. Heidelberg: Springer.

Buccianti, A., & Pawlowsky-Glahn, V. (2006). Statistical evaluation of compositional changes in volcanic gas chemistry: a case study. *Stochastic Environmental Research and Risk Assessment*, *21*(1), 25–33. https://doi.org/10.1007/s00477-006-0041-x

Buccianti, A., Mateu-Figueras, G., & Pawlowsky-Glahn, V. (Eds.). (2006). *Compositional data analysis in the geosciences: from theory to practice*. London: The Geological Society.

Burling, I. R., Yokelson, R. J., Griffith, D. W. T., Johnson, T. J., Veres, P., Roberts, J. M., et al. (2010). Laboratory measurements of trace gas emissions from biomass burning of fuel types from the southeastern and southwestern United States. *Atmospheric Chemistry and Physics*, *10*(22), 11115–11130. https://doi.org/10.5194/acp-10-11115-2010

Byram, G. M. (1957). Some principles of combustion and their significance in forest fire behavior. *Fire Control Notes*, *18*(2), 47–57.

Byram, G. M. (1959). Combustion of forest fuels. In K. P. Davis (Ed.), *Forest Fire: Control and Use* (1st ed., pp. 61–89). New York: McGraw-Hill.

Carslaw, D. C. (2015). *The openair manual— open-source tools for analysing air pollution data* (Manual No. 1.1-4) (p. 287). London, UK: King's College. Retrieved from http://www.openair-project.org/PDF/OpenAir_Manual.pdf

Carslaw, D. C., & Ropkins, K. (2012). openair — An R package for air quality data analysis. *Environmental Modelling & Software*, *27–28*, 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008

Cayuela-Sánchez, J. A., Palarea-Albaladejo, J., García-Martín, J. F., & Pérez-Camino, M. del C. (2019). Olive oil nutritional labeling by using Vis/NIR spectroscopy and compositional

statistical methods. *Innovative Food Science & Emerging Technologies*, *51*, 139–147. https://doi.org/10.1016/j.ifset.2018.05.018

Chang-Graham, A. L., Profeta, L. T. M., Johnson, T. J., Yokelson, R. J., Laskin, A., & Laskin, J. (2011). Case study of water-soluble metal containing organic constituents of biomass burning aerosol. *Environmental Science & Technology*, *45*(4), 1257–1263. https://doi.org/10.1021/es103010j

Chastin, S. F. M., Palarea-Albaladejo, J., Dontje, M. L., & Skelton, D. A. (2015). Combined Effects of Time Spent in Physical Activity, Sedentary Behaviors and Sleep on Obesity and Cardio-Metabolic Health Markers: A Novel Compositional Data Analysis Approach. *PLOS ONE*, *10*(10), e0139984. https://doi.org/10.1371/journal.pone.0139984

Crutzen, P. J., & Brauch, H. G. (Eds.). (2016). *Paul J. Crutzen: a pioneer on atmospheric chemistry and climate in the anthropocene*. Zürich: Springer.

Darley, E. F., Burleson, F. R., Mateer, E. H., Middleton, J. T., & Osterli, V. P. (1966). Contribution of Burning of Agricultural Wastes to Photochemical Air Pollution. *Journal of the Air Pollution Control Association*, *16*(12), 685–690. https://doi.org/10.1080/00022470.1966.10468533

Di Blasi, C. (2008). Modeling chemical and physical processes of wood and biomass pyrolysis. *Progress in Energy and Combustion Science*, *34*(1), 47–90. https://doi.org/10.1016/j.pecs.2006.12.001

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2d ed). New York: Wiley.

Egozcue, J.J. (2009). Reply to “On the Harker Variation Diagrams; …” by J.A. Cortés. *Mathematical Geosciences*, *41*(7), 829–834. https://doi.org/10.1007/s11004-009-9238-0

Egozcue, J.J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, *37*(7), 795–828. https://doi.org/10.1007/s11004-005-7381-9

Egozcue, J.J., & Pawlowsky-Glahn, V. (2011). Basic Concepts and Procedures. In V. Pawlowsky-Glahn & A. Buccianti (Eds.), *Compositional Data Analysis* (pp. 12–28). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch2

Egozcue, J.J., Lovell, D., & Pawlowsky-Glahn, V. (2014). Testing compositional association. In K. Hron, P. Filzmoser, & M. Templ (Eds.), *Proceedings of the 5th International Workshop on Compositional Data Analysis* (pp. 28–36). Vorau, Austria. Retrieved from http://www.statistik.tuwien.ac.at/CoDaWork/CoDaWork2013Proceedings.pdf

Egozcue, J.J., Pawlowsky-Glahn, V., & Gloor, G. B. (2018). Linear Association in Compositional Data Analysis. *Austrian Journal of Statistics*, *47*(1), 3. https://doi.org/10.17713/ajs.v47i1.689

von Eynatten, H. (2004). Statistical modelling of compositional trends in sediments. *Sedimentary Geology*, *171*(1–4), 79–89. https://doi.org/10.1016/j.sedgeo.2004.05.011

von Eynatten, H., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Modelling compositional change: the example of chemical weathering of granitoid rocks. *Mathematical Geology*, *35*(3), 231–251. https://doi.org/10.1023/A:1023835513705

Ferek, R. J., Reid, J. S., Hobbs, P. V., Blake, D. R., & Liousse, C. (1998). Emission factors of hydrocarbons, halocarbons, trace gases and particles from biomass burning in Brazil. *Journal of Geophysical Research: Atmospheres*, *103*(D24), 32107–32118. https://doi.org/10.1029/98JD00692

Ferguson, S. C., Dahale, A., Shotorban, B., Mahalingam, S., & Weise, D. R. (2013). The role of moisture on combustion of pyrolysis gases in wildland fires. *Combustion Science and Technology*, *185*, 435–453. https://doi.org/10.1080/00102202.2012.726666

Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis: with worked examples in R*. New York, NY: Springer Berlin Heidelberg.

Fox, J., & Weisberg, S. (2018, September 21). Multivariate linear models in R: an appendix to "An R Companion to Applied Regression", 3rd edition. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Multivariate-Linear-Models.pdf

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third edition). Thousand Oaks, California: Sage Publications, Inc.

Gerstle, R. W., & Kemnitz, D. A. (1967). Atmospheric Emissions from Open Burning. *Journal of the Air Pollution Control Association*, *17*(5), 324–327. https://doi.org/10.1080/00022470.1967.10468988

Gilman, J. B., Lerner, B. M., Kuster, W. C., Goldan, P. D., Warneke, C., Veres, P. R., et al. (2015). Biomass burning emissions and potential air quality impacts of volatile organic compounds and other trace gases from fuels common in the US. *Atmospheric Chemistry and Physics*, *15*(24), 13915–13938. https://doi.org/10.5194/acp-15-13915-2015

Goode, J. G., Yokelson, R. J., Ward, D. E., Susott, R. A., Babbitt, R. E., Davies, M. A., & Hao, W. M. (2000). Measurements of excess O_3, CO_2, CO, CH_4, C_2H_4, C_2H_2, HCN, NO, NH_3, HCOOH, CH_3COOH, HCHO, and CH_3OH in 1997 Alaskan biomass

burning plumes by airborne Fourier transform infrared spectroscopy (AFTIR). *Journal of Geophysical Research*, *105*(D17), 22147. https://doi.org/10.1029/2000JD900287

Graybill, F. A. (2002). *Matrices with applications in statistics*. Belmont: Duxbury.

Hosseini, S., Li, Q., Cocker, D., Weise, D., Miller, A., Shrivastava, M., et al. (2010). Particle size distributions from laboratory-scale biomass fires using fast response instruments. *Atmospheric Chemistry and Physics*, *10*(16), 8065–8076. https://doi.org/10.5194/acp-10-8065-2010

Hosseini, S., Urbanski, S. P., Dixit, P., Qi, L., Burling, I. R., Yokelson, R. J., et al. (2013). Laboratory characterization of PM emissions from combustion of wildland biomass fuels: particle emissions from biomass burning. *Journal of Geophysical Research: Atmospheres*, *118*(17), 9914–9929. https://doi.org/10.1002/jgrd.50481

Hough, W. A. (1969). *Caloric value of some forest fuels of the southern United States* (Research Note No. SE-120) (p. 6). Asheville, NC: USDA Forest Service, Southeastern Forest Experiment Station. Retrieved from http://www.treesearch.fs.fed.us/pubs/2778

Janhäll, S., Andreae, M. O., & Pöschl, U. (2010). Biomass burning aerosol emissions from vegetation fires: particle number and mass emission factors and size distributions. *Atmospheric Chemistry and Physics*, *10*(3), 1427–1439. https://doi.org/10.5194/acp-10-1427-2010

Jolly, W. M., Hintz, J., Linn, R. L., Kropp, R. C., Conrad, E. T., Parsons, R. A., & Winterkamp, J. (2016). Seasonal variations in red pine (*Pinus resinosa*) and jack pine (*Pinus banksiana*) foliar physio-chemistry and their potential influence on stand-scale wildland

fire behavior. *Forest Ecology and Management*, *373*, 167–178. https://doi.org/10.1016/j.foreco.2016.04.005

Lincoln, E., Hao, W., Weise, D. R., & Johnson, T. J. (2014). *Wildland fire emission factors database* (Archived data No. RDS-2014-0012). Fort Collins, CO: USDA Forest Service Research Data Archive. Retrieved from http://dx.doi.org/10.2737/RDS-2014-0012

Lobert, J. M., & Warnatz, J. (1993). 2 - Emissions from the combustion process in vegetation. In P. J. Crutzen & J. G. Goldammer (Eds.), *Fire in the Environment: The Ecological Atmospheric, and Climatic Importance of Vegetation Fires* (pp. 15–37). John Wiley & Sons Ltd.

Lobert, J. M., Scharffe, D. H., Hao, W. M., & Crutzen, P. J. (1990). Importance of biomass burning in the atmospheric budgets of nitrogen-containing gases. *Nature*, *346*(6284), 552–554. https://doi.org/10.1038/346552a0

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS Computational Biology*, *11*(3), e1004075. https://doi.org/10.1371/journal.pcbi.1004075

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London ; New York: Academic Press.

Mateu-Figueras, G., Pawlowsky-Glahn, V., & Egozcue, J. J. (2011). The Principle of Working on Coordinates. In V. Pawlowsky-Glahn & A. Buccianti (Eds.), *Compositional Data Analysis* (pp. 31–42). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch3

May, A. A., McMeeking, G. R., Lee, T., Taylor, J. W., Craven, J. S., Burling, I., et al. (2014). Aerosol emissions from prescribed fires in the United States: A synthesis of laboratory and aircraft measurements: Aerosols from US prescribed fires. *Journal of Geophysical Research: Atmospheres*, *119*(20), 11,826-11,849. https://doi.org/10.1002/2014JD021848

McGregor, D., Palarea-Albaladejo, J., Dall, P., Hron, K., & Chastin, S. (2019). Cox regression survival analysis with compositional covariates: Application to modelling mortality risk from 24-h physical activity patterns. *Statistical Methods in Medical Research*, 096228021986412. https://doi.org/10.1177/0962280219864125

McMeeking, G. R., Kreidenweis, S. M., Baker, S., Carrico, C. M., Chow, J. C., Collett, J. L., et al. (2009). Emissions of trace gases and aerosols during the open combustion of biomass in the laboratory. *Journal of Geophysical Research*, *114*(D19). https://doi.org/10.1029/2009JD011836

Ottmar, R. D., Miranda, A. I., & Sandberg, D. V. (2008). Characterizing Sources of Emissions from Wildland Fires. In *Developments in Environmental Science* (Vol. 8, pp. 61–78). Elsevier. Retrieved from http://www.treesearch.fs.fed.us/pubs/34248

Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2015). zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, *143*, 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019

Palarea-Albaladejo, J., Martín-Fernández, J. A., & Olea, R. A. (2014). A bootstrap estimation scheme for chemical compositional data with nondetects. *Journal of Chemometrics*, *28*(7), 585–599. https://doi.org/10.1002/cem.2621

Paul, K. T., Hull, T. R., Lebek, K., & Stec, A. A. (2008). Fire smoke toxicity: The effect of nitrogen oxides. *Fire Safety Journal*, *43*(4), 243–251. https://doi.org/10.1016/j.firesaf.2007.10.003

Pawlowsky-Glahn, V., & Buccianti, A. (Eds.). (2011). *Compositional data analysis: theory and applications*. Chichester, West Sussex, U.K.: Wiley.

Pawlowsky-Glahn, V., & Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, *15*(5), 384–398. https://doi.org/10.1007/s004770100077

Pawlowsky-Glahn, V., Egozcue, J. J., Olea, R. A., & Pardo-Igúzquiza, E. (2015a). Cokriging of compositional balances including a dimension reduction and retrieval of original units. *Journal of the Southern African Institute of Mining and Metallurgy*, *115*, 59–72.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015b). *Modelling and analysis of compositional data*. Chichester, West Sussex, U.K.: Wiley.

Pearson, K. (1896). Mathematical contributions to the theory of evolution.--on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London (1854-1905)*, *60*(1), 489–498. https://doi.org/10.1098/rspl.1896.0076

Quinn, T. P., Erb, I., Richardson, M. F., & Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, *34*(16), 2870–2878. https://doi.org/10.1093/bioinformatics/bty175

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0). Vienna, Austria: R Foundation for Statistical Computing.

Radke, L., Hegg, D., Lyons, J., Brock, C., Hobbs, P., Weiss, R., & Rasmussen, R. (1988). Airborne measurements on smokes from biomass burning. In P. V. Hobbs & M. P. McCormick (Eds.), *Aerosols and Climate* (pp. 411–422).

Roberts, J. M., Veres, P., Warneke, C., Neuman, J. A., Washenfelder, R. A., Brown, S. S., et al. (2010). Measurement of HONO, HNCO, and other inorganic acids by negative-ion proton-transfer chemical-ionization mass spectrometry (NI-PT-CIMS): application to biomass burning emissions. *Atmospheric Measurement Techniques*, *3*(4), 981–990. https://doi.org/10.5194/amt-3-981-2010

Rockwell, B. G., Girty, G. H., & Rockwell, T. K. (2014). A statistical framework for calculating and assessing compositional linear trends within fault zones: a case study of the NE block of the Clark Segment, San Jacinto Fault, California, USA. *Pure and Applied Geophysics*, *171*(11), 2919–2935. https://doi.org/10.1007/s00024-014-0851-6

Rogers, J. M., Susott, R. A., & Kelsey, R. G. (1986). Chemical composition of forest fuels affecting their thermal behavior. *Can. J. For. Res.*, *16*(4), 721–726. https://doi.org/10.1139/x86-129

Sekimoto, K., Koss, A. R., Gilman, J. B., Selimovic, V., Coggon, M. M., Zarzana, K. J., et al. (2018). High- and low-temperature pyrolysis profiles describe volatile organic compound emissions from western US wildfire fuels. *Atmospheric Chemistry and Physics*, *18*(13), 9263–9281. https://doi.org/10.5194/acp-18-9263-2018

Shafizadeh, F. (1984). The Chemistry of Pyrolysis and Combustion. In R. Rowell (Ed.), *The Chemistry of Solid Wood* (Vol. 207, pp. 489–529). Washington, D.C.: American Chemical Society.

Shen, G., Xue, M., Wei, S., Chen, Y., Zhao, Q., Li, B., et al. (2013). Influence of fuel moisture, charge size, feeding rate and air ventilation conditions on the emissions of PM, OC, EC, parent PAHs, and their derivatives from residential wood combustion. *Journal of Environmental Sciences*, *25*(9), 1808–1816. https://doi.org/10.1016/S1001-0742(12)60258-7

Speranza, A., Caggiano, R., Pavese, G., & Summa, V. (2018). The study of characteristic environmental sites affected by diverse sources of mineral matter using compositional data analysis. *Condensed Matter*, *3*(2), 16. https://doi.org/10.3390/condmat3020016

Surawski, N. C., Sullivan, A. L., Roxburgh, S. H., Meyer, C. P. M., & Polglase, P. J. (2016). Incorrect interpretation of carbon mass balance biases global vegetation fire emission estimates. *Nature Communications*, *7*, 11536. https://doi.org/10.1038/ncomms11536

Tangren, C. D., McMahon, C. K., & Ryan, P. W. (1976). Chapter II - Contents and effects of forest fire smoke. In *Southern forestry smoke management guidebook* (pp. 9–22). Asheville, NC: USDA Forest Service, Southeastern Forest Experiment Station. Retrieved from http://www.treesearch.fs.fed.us/pubs/683

Templ, M., Hron, K., & Filzmoser, P. (2011). robCompositions: An R-package for robust statistical analysis of compositional data. In V. Pawlowsky-Glahn & A. Buccianti (Eds.), *Compositional Data Analysis* (pp. 341–355). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch25

Thió-Henestrosa, S., & Comas, M. (2016, April 20). CoDaPack v2 User’s Guide. University of Girona, Dept. of Computer Science and Applied Mathematics. Retrieved from http://ima.udg.edu/codapack/assets/codapack-manual.pdf

Urbanski, S. P. (2013). Combustion efficiency and emission factors for wildfire-season fires in mixed conifer forests of the northern Rocky Mountains, US. *Atmospheric Chemistry and Physics*, *13*(14), 7241–7262. https://doi.org/10.5194/acp-13-7241-2013

Veres, P., Roberts, J. M., Burling, I. R., Warneke, C., de Gouw, J., & Yokelson, R. J. (2010). Measurements of gas-phase inorganic and organic acids from biomass fires by negative-ion proton-transfer chemical-ionization mass spectrometry. *Journal of Geophysical Research*, *115*(D23), D23302. https://doi.org/10.1029/2010JD014033

Ward, D. E. (2001). Combustion chemistry and smoke. In E. A. Johnson & K. Miyanishi (Eds.), *Forest Fires: Behavior and Ecological Effects* (pp. 55–77). San Diego, CA: Academic Press. Retrieved from http://www.doi.org/10.1016/B978-012386660-8/50006-3

Ward, D. E., & Hao, W. M. (1991). Projections of emissions from burning of biomass for use in studies of global climate and atmospheric chemistry (p. 19). Presented at the Annual Meeting of the Air and Waste Management Association, Vancouver, British Columbia, Canada: Air and Waste Management Association. Retrieved from http://www.treesearch.fs.fed.us/pubs/43258

Ward, D. E., Clements, H. B., & Nelson, R. M., Jr. (1980). Particulate matter emission factor modeling for fire in southeastern fuels. In *Sixth Conference on Fire and Forest Meteorology* (pp. 276–284). Seattle, WA: American Meteorological Society.

Warneke, C., Roberts, J. M., Veres, P., Gilman, J., Kuster, W. C., Burling, I., et al. (2011). VOC identification and inter-comparison from laboratory biomass burning using PTR-MS and PIT-MS. *International Journal of Mass Spectrometry*, *303*(1), 6–14. https://doi.org/10.1016/j.ijms.2010.12.002

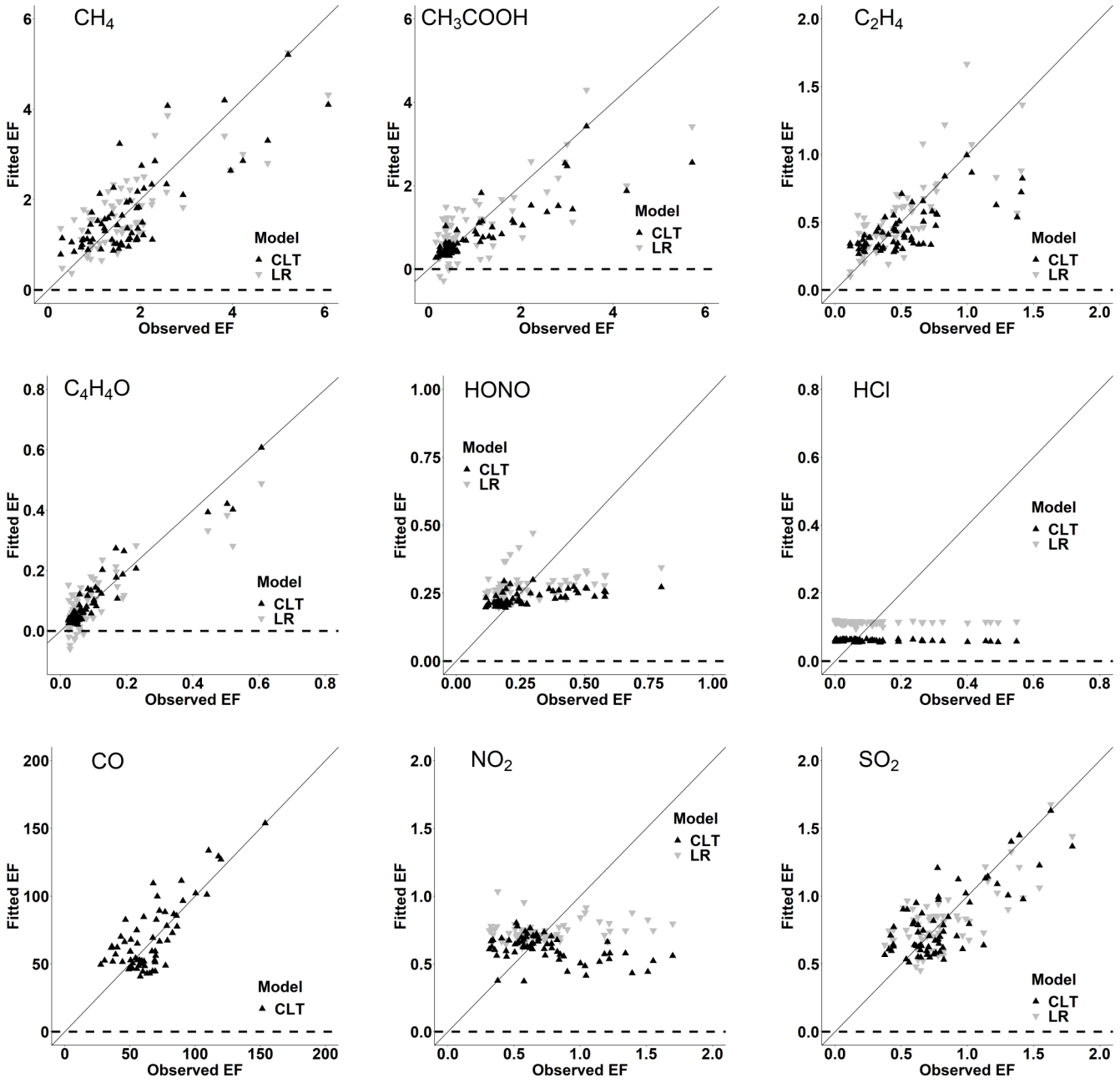
Warton, D. I., Wright, I. J., Falster, D. S., & Westoby, M. (2006). Bivariate line-fitting methods for allometry. *Biological Reviews*, *81*(02), 259. https://doi.org/10.1017/S1464793106007007

Weise, D. R., Johnson, T. J., & Reardon, J. (2015). Particulate and trace gas emissions from prescribed burns in southeastern U.S. fuel types: Summary of a 5-year project. *Fire Safety Journal*, *74*, 71–81. https://doi.org/10.1016/j.firesaf.2015.02.016

Yokelson, R. J., Susott, R., Ward, D. E., Reardon, J., & Griffith, D. W. T. (1997). Emissions from smoldering combustion of biomass measured by open-path Fourier transform infrared spectroscopy. *Journal of Geophysical Research-Atmospheres*, *102*(D15), 18865–18877. https://doi.org/10.1029/97JD00852

Yokelson, R. J., Burling, I. R., Gilman, J. B., Warneke, C., Stockwell, C. E., de Gouw, J., et al. (2013). Coupling field and laboratory measurements to estimate the emission factors of identified and unidentified trace gases for prescribed fires. *Atmospheric Chemistry and Physics*, *13*(1), 89–116. https://doi.org/10.5194/acp-13-89-2013

Figure 1.


CH4
CH3COOH
C2H4
C4H4O
HONO
HCl
CO
NO2
SO2
Fitted EF
Observed EF
Model
CLT
LR