Journal Pre-proof

Univariate statistical analysis of gas chromatography – mass spectrometry fingerprints analyses

Tamires Oliveira Melo , Luziane Franciscon , George Brown , Joachim Kopka , Luis Cunha , Federico Martinez-Seidel , Luiz Augusto dos Santos Madureira , Fabricio Augusto Hansel , TPI Network

PII: S2405-8300(21)00073-2
DOI: https://doi.org/10.1016/j.cdc.2021.100719
Reference: CDC 100719

To appear in: *Chemical Data Collections*

Received date: 10 December 2020
Revised date: 15 April 2021
Accepted date: 3 May 2021

Please cite this article as: Tamires Oliveira Melo , Luziane Franciscon , George Brown , Joachim Kopka , Luis Cunha , Federico Martinez-Seidel , Luiz Augusto dos Santos Madureira , Fabricio Augusto Hansel , TPI Network, Univariate statistical analysis of gas chromatography – mass spectrometry fingerprints analyses, *Chemical Data Collections* (2021), doi: https://doi.org/10.1016/j.cdc.2021.100719

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier B.V.

Title:

Univariate statistical analysis of gas chromatography – mass spectrometry fingerprints analyses

Authors:

Tamires Oliveira Melo[a,b,c], Luziane Franciscon[b], George Brown[b,d], Joachim Kopka[c], Luis Cunha[e,f], Federico Martinez-Seidel[c,g], Luiz Augusto dos Santos Madureira[h], TPI Network[i], and Fabricio Augusto Hansel[b]

Affiliations:

[a] Departamento de Química, Universidade Federal do Paraná, Curitiba 81531 980, PR, Brazil.

[b] Embrapa Florestas, Estrada da Ribeira, km 111, C.P. 319, Guaraituba, Colombo 88411 000, PR, Brazil.

[c] Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany.

[d] Soil Science Department, Federal University of Paraná, Curitiba, 80035-050, PR, Brazil

[e] University of Coimbra, Centre for Functional Ecology, Department of Life Sciences, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal.

[f] School of Applied Sciences, University of South Wales, Pontypridd, Wales, CF37 4BD UK.

[g] School of BioSciences, The University of Melbourne, Victoria 3010, Australia.

[h] Departamento de Química, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina 88040 900, Brazil.

i School of BioSciences, Cardiff University, Cardiff, CF103AT, United Kingdom.

Contact email: Fabricio.hansel@embrapa.br

Abstract

Gas Chromatography - Mass Spectrometry (GC-MS) has been used for a long time in fingerprint analysis. We present a workflow of univariate statistical treatment of compound by considering their type of response variables. Two data sources were used: (i) comparative data from two Brazilian Amazon soils, and (ii) the Nitrogen-dose response experiment involving two Ilex paraguariensis clones. During type of response variables selection, the following assumptions were tested: normality and homogeneity of variances. After defining a strategy to select the type of response variables, the compounds were classified according to the statistical test that must be used to evaluate them: analysis of variance (ANOVA, LM), generalized linear model (GLM), and a non-parametric (NP) test. The developed workflow allows individual compound and class comparisons, and a couple examples that illustrate a wider range of similar datasets are open to the readers to test either their own data or ours.

Keywords analytical chemistry, biological markers, plant physiology, biogeochemistry.

Specifications Table

Subject area	*Analytical chemistry*
Compounds	*Analyte peak areas or heights of Gas Chromatography - Mass Spectrometry (GC-MS) fingerprint analysis*
Data category	*Numerical data matrices of semi-quantitative compounds*
Data acquisition format	mass-to-charge features
Data type	*Raw, filtered, and analyzed*
Procedure	*mass-to-charge aligned molecular feature intensities (height or area) were normalized to IS, and sample amount. Threshold were stablished according to percentage distribution.*
Data accessibility	*https://github.com/FAHansel/Stats_Fp_GC*

1. Rationale

Gas Chromatography - Mass Spectrometry (GC-MS) has been used for long time as fingerprint analysis in many different fields of science, such as organic geochemistry and metabolomics [1,2,3ς6]. In such profiling studies, information on most response variables in statistical terminology (i.e., a data item that can be measured or counted), or in the terminology of metabolomic analyses (i.e., a molecular features) is qualitative, e.g., present or absent, or a relative change compared to a reference condition [7]. As a routine, relative quantification (i.e. semi-quantification) of the variables is supported by addition of at least one internal standard (IS), such as an isotope-labelled or xenobiotic reference compound. Analyte peak areas or heights are normalised proportional to the IS and corrected for a measure of initial sample amount [5,8]. In the case of GC-MS profiling partially automated mass spectral deconvolution and matching software are key to the annotation task, such as AMDIS [9] or Tagfinder [10]. After annotation the semi- quantification and statistical analysis process can be carried out based on numerical data matrices of chromatographically and mass-to-charge aligned molecular features [8,10].

For screening and data reduction purposes multivariate analyses (MVAs), such as principal components analysis (PCA) and hierarchical cluster analyses (HCA), are frequently used to differentiate between groups of analysed samples in order to cluster groups of alike response variables according to a predefined similarity or dissimilarity metric. Such MVAs provide a rough overview of general sample distribution at the cost of data/variable reduction [11]. Univariate analyses of response variables and their distributions are an extension to MVA and allow class comparison, and may even be seen as the basic toolbox of data mining.

Univariate statistics including t-tests and analyses of variance (ANOVA) for pairwise comparisons of control and treatment conditions, in combination with post-hoc and its non-parametric test (NP) versions are the most popular statistical methods for metabolomics and biomarkers analyses [11,12]. T-tests and ANOVA are used when the response variable has a normal distribution (LM). Modified non-parametric procedures (NP) are available for cases of with non-normal distributed response variables [13,14]. Interestingly, the generalized linear model (GLM) as an alternative to NP testing has been applied to GC-MS data that are non-normally distributed [15].

According to Vinaixa et al. (2012) [16] response variables, including GC-MS technology, may include both parametric (normally) and non-parametric (non-normally) variables. In this paper we propose a workflow or in other words a practical guide to the statistical treatment of single response variables using LM, GLM, or NP approaches. The statistical data analysis workflow is illustrated using two experiments that differ in nature and scope: (i) resulting from a biomarker analysis dataset of different soil aggregates from two amazonian soils macroaggregates and, [illegible]

2. Procedure

2.1 Biomarkers in two amazonian soils macroaggregates

2.1.1 Soil Samples

The Amazonian Dark Earth (ADEs) soils are characterized by a high content of nutrients and organic matter [17]. The origin of the fertility ADE soils has been attributed in part to the incorporation of biochar into the mineral soil [18], and the bioturbation by ecosystem engineers (e.g. earthworms) that may play an important role in such incorporation [19]. Briefly, the experiment was conducted in order to understand the distinction of below ground behaviour of ecosystem engineers and their interactions with physical and root aggregates considering two contrasting amazonian soils, ADE and control soil. Soil samples were taken from ADE and a control soil (non-ADE adjacent soil) in the Brazilian Amazon Basin (Teotônio, Porto Velho). Five 10 x 10 cm blocks were collected from the 0-10 cm topsoil horizon in each soil using an X-shaped sampling design, with four samples in each corner (distant 60 m from each other) and one in the center. The soils at the site were classified as Pretic Clayic Anthrosol (ADE) and Xanthic Dystric Plinthosol (control soil) according to the World Resource Base/Food and Agriculture Organization classification [20, 21]. Each soil sample was divided into three aggregates fractions: (i) fauna-produced (n=10), physical (n=10) and, root (n=9), resulting in a total 20 ADE samples and 19 control soil samples [22].

2.1.2 Sample preparation, solvent extraction, and derivatization

[illegible] *-cholestane (IS) was extracted with a solution of chloroform: methanol (2: 1 v / v, 5 mL, 3 x 15 min) in an ultrasound bath. Each extract was centrifuged for 10 min at 3000 rpm and the supernatants combined in a glass flask. A rotary evaporator was used to remove the excess solvent, while the remaining solvent was dried with a micro-column with 2 g of anhydrous sodium sulfate, and the eluate transferred to a glass vial (2.0 mL). Nitrogen was used to remove the remaining solvent. The extracts were stored at -20 ° C until analysis on GC-MS. Dry samples were trimethylsil*[illegible] *-MS analysis.*

2.1.3 Gas chromatography [illegible] *mass spectrometry*

The samples were analyzed using an Agilent Technologies GC (7890B) coupled to the MS from Agilent Technologies (5977A), with a PAL System (CombiPAL RSI model) autosampler, with splitless injection mode into a HP-5ms column , 30 m, 0.25 mm, 0.25 µm film thickness (J&W Ultra Inert column), with the following oven program: 50 °C (3 min), heating rate at 10 °C min^{-1} to 300 °C (3 min), and the injector temperature at 280 °C. Helium was used as the carrier gas (1.0 ml min^{-1}). The GCçMS interface and the ion source temperatures were 300 °C and 200 °C, respectively. The mass spectrometer was operated in the positive electronic mode at 70 eV with the range of 50ç650 Da. The program used for CG-MS data analysis was the MassHunter (version B.07.00.1413), while the semi-quantification was performed using MSD ChemStation (version E.02.02.1431), both of Agilent Technologies. The mass spectral deconvolution and the automated calculation of retention indices (RI) were performed using AMDIS (v. 2.72 build 140.24, NIST / DTRA / OPCW). Compounds were identified from deconvoluted mass spectra in comparison to mass spectra and RI provide by the NIST MS software (version 2.2). Guidelines for manually supervised analyte identification were the presence of at least three specific mass fragments per compound and a RI deviation of less than 1.5% [23]. Analytes intensities were normalized to IS peak area, and dried soil sample weights.

2.2 Nitrogen nutrition Ilex experiment

2.2.1 Biological material, plant nutrition, and harvesting

Ilex paraguariensλă ! φ {ΰφ I λftφ óάŜℕØI· Y΄I·ΰŜέύ tŜI·ŧ ΰλăăάdzŜ KI·ă ōŜŜγ΄2dzΰăΰI·yŘλyΞ I·ă ℕI·ø Y΄I·ΰŜℕλI·t λγ΄ŘλŧŧŜℕŜγΰ sectors, due to the production of bioactive compounds [24ç26]. Comparing different genotypes to evaluate the distribution of bioactive substances in relation to their nutrition is an important factor to support the breeding programs [27ç29]. Thus, the changes in plant metabolite levels were evaluated in relation to N nutrition level of Ilex paraguariensis [30]. Epicormic shoots were collected and rooted from two 10-year-old mother plants from Ivaí-PR-Brazil (genotypes EC22 and EC40) [31]. The cuttings were established as stumps after ca. 120 days with 15 cm height, and maintained in a semi-hydroponics system in a non-acclimatized greenhouse. Five biological replicate samples of each clone (EC22 and EC40) were fertilized with five concentrations of inorganic N (NH4+ + NO3-) 114, 206, 380, 761, and 1142 mg L-1. The metabolite analysis of mature leaves was performed after 8 months of plant establishment.

2.2.2 Sample preparation, solvent extraction, and derivatization

The extraction and profile analysis of metabolites were performed as previously described [35]. In detail, the extraction was carried out at 450 rpm for 15 min at 70 ° C, with 20 mg (± 5 mg) of fresh frozen and ΞℕŽdzyŘ ØŜΞŜΰI·ōfŜ Y΄I·ΰŜℕλI·f I·yŘ олл >[2ŧ LJℕŜ-cooled methanol, previously transferred to a Y΄λŎℕŽŎŜγΰℕλŧdzΞŜ ΰdzōŜ 2ŧ н Y΄[I·yŘ Ø2ℕΰŜËŜŘ I·yŘ ŜγℕλŎKŜŘ øλΰK ол >[2ŧ $^{13}C_6$-sorbitol (IS, 0.2 mg mL^{-1} in Y΄ŜΰKI·y2fύ I·yŘ ол >[2ŧ y-nonadecanoic acid, extraction control (2 mg mL^{-1} in chloroform). The samples øŜℕŜ ΰKŜγ΄Ŏ22fŜŘ ΰŽ ℕ22Y΄ ΰŜ Y΄LJŜℕI·ΰdzℕŜ I·yŘ нлл >[2ŧ ŎKf2ℕŽŧŽℕY΄ I·ŘŘŜŘφ ¢KŜ ă2fdzΰλ2γ΄øI·ă Ø2ℕΰŜËŜŘ I·yŘ ΰKŜγ΄KŜI·ΰŜŘ I·ΞI·λγ΄λγ΄ΰKŜ ΰKŜℕY΄2Y΄λÈŜℕ I·ΰ прл ℕLJY΄ŧ2ℕ p Y΄λγ΄I·ΰ om c /Σ плл >[2ŧ dzfΰℕI·LJdzℕŜ øI·ΰŜℕ øI·ă added and the ΰdzōŜă øŜℕŜ Ø2ℕΰŜËŜŘ ŧ2ℕ нл ăŜŎ2yŘă I·yŘ ŎŜγΰℕλŧdzΞŜŘ ŧ2ℕ m Y΄λγ΄I·ΰ мнΣллл ℕLJYφ ¢ø2 мсл >[aliquots of the polar phase were transferred to two 1.5 ml conical microcentrifuge tubes (that is, one aliquot for each tube), and the extracts were vacuum dried. Methoximation was performed in a ΰKŜℕY΄2Y΄ÈŜℕ I·ΰ фрл ℕLJY΄ŧ2ℕ фл Y΄λγ΄I·ΰ ол c /Σ I·ŘŘλyΞ пл >[2ŧ ă2fdzΰλ2γ΄Ŏ2yΰI·λγλyΞ п- (dimethylamino) -

pyridine (5 mg) and methoxyamine hydrochloride (40 mg) in pyridine (1 ml) to the dry extract. Trimethylsilylation was achieved õè l·ŘŘλyƎ mл >[2₸ . {¢C !Σ Ø2ℕűŝÈλyƎ l·yŘ λyÕʤōl·űλyƎ l·ű om c / λy˝l· thermomixer at 950 rpm for 30 min. Soon after, the eppendorf tubes were centrifuged at 14000 rpm for p Yλy˝l·yŘ l·y˝yл >[l·fλljʤż́ ǿl·ă űℕl·yãŦŜℕℕŜŘ ű2 l· Ŏ2yλŎl·f ō2ű́ű2Y˝ Øλal and the sample was analyzed by GC-TOF/MS.

2.2.3 Gas chromatography ς mass spectrometry
The profiling of polar metabolite by GC-TOF/MS was performed by an 6890N24 gas chromatograph (Agilent Technologies, Germany), with splitless injection onto a capillary column VF-5 MS, 30 m, 0.25 mm, 0.25 µm film thickness (Agilent Technologies, Santa Clara, CA, USA), connected to a Pegasus III TOF (LECO Instrumente GmbH, Mönchengladbach, Germany), with the following oven program: 70 °C (1 min), heating rate at 9 °C min^{-1} to 350 °C (5 min)), and the injector temperature at 230 °C. Helium was used as the carrier gas (0.6 ml min^{-1}). The GCçMS interface and the ion source temperatures were 250 °C and 200 °C, respectively. The mass spectrometer was operated in the positive electronic mode at 70 eV, with the range of 70ç600 Da. Chromatograms were acquired, visually controlled, baseline corrected and exported in NetCDF file format using ChromaTOF (v. 4.22; LECO, St. Joseph, USA). Compounds were identified by mass spectral and RI matching to the reference collection of the Golm Metabolome Database (GMD [32,33]) under manual supervision using TagFinder software [10]. Guidelines for manually supervised metabolite identification were as described above [23]. Metabolite intensities were normalized to IS peak height, and sample fresh weight.

2.3 Interaction experiments and data treatment
Interaction experiments were considered to assess the relationship between two independent variables (factors - F) whereby the magnitude of one variable moderates the effect of the other. The workflow for univariate statistical data treatment was illustrated in two data sources that include interaction experiments: (i) the comparative data from the Brazilian Amazon, including two soil types (F1; i.e., ADE and control) and three types of aggregates (F2; i.e., fauna-produced, roots and physical), and (ii) the N-dose response experiment involving two Ilex clones (F1; EC22 and EC40) and five inorganic N doses (F2) (dataset in https://github.com/FAHansel/Stats_Fp_GC). Compounds at concentration higher than of 0.1% in soil aggregates, and 0.01% in N nutrition in the replicates were included in the quantification analyses. The data were pre-processed following the guide of Yang et al., 2015 [34] and Wei et al., 2018 [35].

The variables were statistically tested using R and Rstudio (Version 1.1.453 ς © 2009-2018 RStudio, Inc.). The packages used were stats [36], xlsx [37], multcomp [38], lattice [39], agricolae [40], hnp [41], ExpDes [42], effects [43] and, imputeMissings [44]. Variables with normal distributions were analysed using ANOVA with F-test ($p > 0.05$), and in the sequence GLM with a CHISQ test ($p < 0.05$) were applied. During selection of the GLM distribution model, the shapes of the response variable distributions were plotted (i.e., kurtosis and square of skewness) for both data sources (i.e., the comparative and dose dependent experiments). Simultaneously, several random vectors with specific distributions (i.e. normal, logis, beta and gamma) were constructed and the same parameters were plotted in lines to see how the observations approximate the model distributions (Fig 1, R function in

https://github.com/MSeidelFed/RandodiStats). The parameterizing of the response variables with the link function negative inverse (μ = -ó · ĵ ú-1) of the gamma or beta family in the GLM appeared to be the best mathematical decision. Therefore, in this study the GLM gamma family was used in the exemplified variables, and if the response variables differed from a gamma distribution needed a NP treatment, and hence were analysed using the Kruskal-Wallis test (dataset in https://github.com/FAHansel/Stats_Fp_GC). The post-hoc tests adopted the critical probability value of 5% for statistical significance, and regression analyses were tested using 1st, 2nd, and 3rd order equations, using excel® to plot the equations provided by R software.

3. Data, value and validation

In all univariate statistical tests the three assumptions must be satisfied to apply parametric tests (e.g. analysis of variance (ANOVA)): normality, homogeneity of variances (homoscedasticity) and independence assumptions (often addressed during study design) [12], and if not satisfied nonparametric tests (NP) may be applied (e.g. Kruskal-Wallis test). However, considering that parametric tests are more powerful than NP tests [16], before the use of KruskalçWallis test, we tried to satisfy variable normality and homogeneity of variances assumptions by parametrizing the mean and variances using GLM with gamma distribution [15].

Thus, it is important to define a strategy to select how well the response values of a given variable fit an expected/assumed analytical non-systematic error distribution, and for such purpose the four residual plots provided by R software were used (Fig. 2) [36]. Residuals vs fitted plots are used to verify the distribution of the residuals and indicate if they are randomly distributed around zero (with no obvious patterns), further indicating the homoscedasticity (Fig. 2A). In the case of n-eicosanoic acid, a conical pattern of the residuals was detected confirming the lack of homoscedasticity (Fig. 2G). The residuals that follow a straight line as in the case of the shikimic acid, quantile-quantile plot (normal Q-Q plot; Fig. 2B) indicate normal distribution, contrasting with the Q-Q plot for 2-hydroxy-glutaric acid, that has residual values moving off the line considerably, resembling an outlier behaviour (Fig. 2E). A half normal plot with simulation envelopes is an alternative plot to check if the residual are normally distributed: if a large number of residuals are off the two solid lines, it indicates that they do not follow a normal distribution (Fig. 2F). A scale-location plot is a classical plot to test the assumption homoscedasticity, and residuals with equal variance are distributed along the horizontal line without significant distortion (Fig. 2C). An example for data with lack of homoscedasticity (heteroscedasticity) in the scale location plot is depicted for n-eicosanoic acid (Fig. 2H). The residual vs. leverage plot is used to find influential cases in the dataset, and their inclusion or exclusion should be considered in the analysis. A sample with high residuals and low leverage does not fit the model well. The third case is a sample with high residual and fŜØŜℕI·ƎŜ I·yŘ ℕŜLJℕŜăŜyűă Y2ăű fλ]Ŝfë I· úℕdzŜ 2dzúfλŜℕT λű I·LJLJŜI·ℕă λyãλŘŜ űKŜ fλyŜ űKI·ű λyŘλŎI·ŰŜă űKŜ /22]Ûă distance, and its exclusion from the analysis must be considered. Figure 2D shows that all samples fit the model well, thus that no outlier was detected during analysis of shikimic acid.

Compounds can be classified in three types of response variables according to the statistical test that must be used to evaluate them: LM, GLM and NP (Fig. 3A). Three types of response variables may be

observed in GC-MS fingerprint univariate statistical analysis with different proportions (Fig. 3A). The flow diagram for univariate statistical data treatment using is summarized in Fig. 3B. In the first step the normality assumption was evaluated using a Shapiro-Wilk test. Secondly, for each dataset it was always necessary to evaluate the four residual plots to assess normality, heteroscedasticity, and outliers. This prompted the decision of whether to use parametric (LM or GLM) or NP tests, and to follow the same approach throughout the rest of the data analysis procedure. It is worth mentioning that NP methods were not used to analyse interaction between factors [45]; rather during the evaluation of NP type response variables the main effects (F1 and F2) were analysed separately, and if differences occurred in both ($p<0.05$) the tests were further dissected into the levels.

Results of comparative experiment (Amazonian soils x aggregates) are showed in the Table 1 and Figure 4. Seven compounds were exemplified: two with interaction (F1 x F2), two only aggregates were significant (F2), three with only soils were different (F1). The NP and GLM response variables were depicted.

The compounds regression curves of N-dose fertilization study are depicted in figure 5, two with interaction (F1 x F2) and two dose dependent (F2) and include both LM and GLM type of response variables. Ilex clones should be compared when only the F1 main effect is significant. Tyramine (GLM) and myo-inositol (LM) showed an interaction effect, and in the clone EC40 a linear regression curve fitted both compounds (Fig. 5 A,B). For tyramine a third order regression curve fitted data of the clone EC22 (Fig. 5A). For the compounds shikimic and succinic acids, no difference was detected between the clones; only the N doses were significant, so that mean values of the clones were plotted against the N doses (Fig. 5C,D). Shikimic acid distribution followed a second order curve (Fig. 5C), similar to that of tyramine in the clone EC22 (Fig. 5A). A linear increase on succinic acid with increase of N fertilization was observed (Fig. 5D).

4. Conclusion

A simple guideline for univariate statistical treatment of GC-MS fingerprint analyses is reported. The process of data analysis is applicable to comparative and dose dependent study designs. The criterion comprises a defined strategy to select and guide the statistical treatment according to the nature of the response variables that are evaluated by LM, GLM or NP statistical tests. The addition of GLM represent a gain to the statistical analysis of data, mainly in respect to the dose dependent experiment in which the number of NP type of response variables limits the number of regression curves. The developed workflow allows chemical classes and individual variables comparisons, and it can be considered as an essential and basic toolbox for data mining. In order to facilitate and standardize the criterion application all statistical analyses should be carried out using R free software.

Conflicts of Interest:

The authors declare no conflict of interest.

Acknowledgements

We thank Jocinei Dognini e Ilene Crestani (Senai, Escola Agrícola de Blumenau) for the GC-MS equipment, MSc. Gustavo Karsten for the analysis and the dataset of compounds from the Amazonian soil samples, Alexander Erban for the helping in the metabolome dataset, and the anonymous reviewer for his comments. We thank Embrapa Rondônia and Embrapa Florestas for logistical support and the farmers and landowners for permitting soil sampling on their properties. Jéssica de Cássia Tomasi and Ivar Wendling are thanked for the conduction of ilex paraguariensis mini-stumps in a semi-hydroponics system. This work was supported by grants from EMBRAPA (02.09.01.014.00.00 and 20.18.01.002.00.03), CNPq (401824/2013-6), Fundação Araucária (45166.460.32093.02022015), NERC-UK (NE/M017656/1), Newton Fund (NE/N000323/1), as well as an U-Horizon 2020 Marie Curie grant to LC (MSCA-IF-2014-GF-660378), a DAAD-Capes grant to TOM (Edital CAPES nº 15/2017), and a CNPq fellowship to GGB (307486/2013-3).

References

[1] S. Alseekh, A.R. Fernie, Metabolomics 20 years on: what have we learned and what hurdles remain?, Plant J. 94 (2018) 933ç942. https://doi.org/10.1111/tpj.13950.
[2] and J.M.M. Peters, K. E., C. C. Walters, The biomarker guide: v. 1 τ Biomarkers and isotopes in the environment and human history, : Cambridge, United Kingdom, Cambridge University Press, 2005.
[3] P.A. Meyers, R. Ishiwatari, Lacustrine organic geochemistry-an overview of indicators of organic matter sources and diagenesis in lake sediments, Org. Geochem. 20 (1993) 867ç900. https://doi.org/10.1016/0146-6380(93)90100-P.
[4] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: Improving the biological information content of metabolomics data, BMC Genomics. 7 (2006) 1ç15. https://doi.org/10.1186/1471-2164-7-142.
[5] R.P. Evershed, Organic residue analysis in archaeology: The archaeological biomarker revolution, Archaeometry. 50 (2008) 895ç924. https://doi.org/10.1111/j.1475-4754.2008.00446.x.
[6] Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L., Metabolite profiling for plant functional genomics, Nat. Biotechnol. 18 (2000) 1157ç61. http://www.nature.com/nbt/journal/v18/n11/full/nbt1100_1157.html.
[7] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.-M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, Proposed minimum reporting standards for chemical analysis, Metabolomics. 3 (2007) 211ç221.
[8] J. Kopka, A. Fernie, W. Weckwerth, Y. Gibon, M. Stitt, Metabolite profiling in plant biology: Platforms and destinations, Genome Biol. 5 (2004) 1ç9. https://doi.org/10.1186/gb-2004-5-6-109.
[9] S.E. Stein, P. Ausloos, C.L. Clifton, J.K. Klassen, S.G. Lias, A.I. Mikaya, O.D. Sparkman, D. V Tchekhovskoi, V. Zaikin, D. Zhu, Evaluation of the NIST/EPA/NIH Mass Spectral Library., (1999).
[10] A. Luedemann, K. Strassburg, A. Erban, J. Kopka, TagFinder for the quantitative analysis of gas chromatography - Mass spectrometry (GC-MS)-based metabolite profiling experiments, Bioinformatics. 24 (2008) 732ç737. https://doi.org/10.1093/bioinformatics/btn023.

[11] F. de A. Lima, L. Leifels, Z. Nikoloski, Regression-Based Modeling of Complex Plant Traits Based on Metabolomics Data, in: Plant Metabolomics, Springer, 2018: pp. 321ς327.
[12] {ǿ [ƛ·Y ÌǾKKƛ·yŜΣ tǿ {ŜyΣ !ǿaǿ 5ÌǾǀŜyãΣ ¢ǿ ƛèƛűèfŅÌyŜyΣ aǿ hŜǾÌőΣ !y hǿŜǾÌŜǿ ƻŦ aŜűƛ·ōƻfƻY ÌǾă Data Analysis: Current Tools and Future Perspectives, Compr. Anal. Chem. 82 (2018) 387ς413. https://doi.org/10.1016/bs.coac.2018.07.001.
[13] A. Khan, G. Rayner, Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem, J. Appl. Math. Decis. Sci. 7 (2003) 187ς206. https://doi.org/10.1207/s15327612jamd0704_1.
[14] S. Ren, A.A. Hinzman, E.L. Kang, R.D. Szczesniak, L.J. Lu, Computational and statistical analysis of metabolomics data, Metabolomics. 11 (2015) 1492ς1513. https://doi.org/10.1007/s11306-015-0823-6.
[15] N. Austel, J. Schubert, S. Gadau, H. Jungnickel, L.T. Budnik, A. Luch, Influence of fumigants on sunflower seeds: Characteristics of fumigant desorption and changes in volatile profiles, J. Hazard. Mater. 337 (2017) 138ς147. https://doi.org/10.1016/j.jhazmat.2017.04.070.
[16] M. Vinaixa, S. Samino, I. Saez, J. Duran, J.J. Guinovart, O. Yanes, A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data, Metabolites. 2 (2012) 775ς795. https://doi.org/10.3390/metabo2040775.
[17] B. Glaser, J.J. Birk, State of the scientific knowledge on properties and genesis of Anthropogenic Dark Earths in Central Amazonia (terra preta de Índio), Geochim. Cosmochim. Acta. 82 (2012) 39ς51.
[18] B. Glaser, L. Haumaier, G. Guggenberger, W. Zech, Black carbon in soils: the use of benzenecarboxylic acids as specific markers, Org. Geochem. 29 (1998) 811ς819.
[19] M. Blouin, M.E. Hodson, E.A. Delgado, G. Baker, L. Brussaard, K.R. Butt, J. Dai, L. Dendooven, G. Pérès, J.E. Tondoh, A review of earthworm impact on soil function and ecosystem services, Eur. J. Soil Sci. 64 (2013) 161ς182.
[20] I.W.G. WRB/FAO, World reference base for soil resources 2014, update 2015: International soil classification system for naming soils and creating legends for soil maps, World Soil Resour. Reports No. 106. (2015) 192.
[21] W.C. Demetrio, Soil macroinvertebrates and soil quality in Amazonian Dark Earths and adjacent soils, Federal University of Paraná (UFPR), 2018.
[22] E. Velasquez, C. Pelosi, D. Brunet, M. Grimaldi, M. Martins, A.C. Rendeiro, E. Barrios, P. Lavelle, This ped is my ped: visual separation and near infrared spectra allow determination of the origins of soil macroaggregates, Pedobiologia (Jena). 51 (2007) 75ς87.
[23] N. Strehmel, J. Hummel, A. Erban, K. Strassburg, J. Kopka, Retention index thresholds for compound matching in GC-MS metabolite profiling, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 871 (2008) 182ς190. https://doi.org/10.1016/j.jchromb.2008.04.042.
[24] C.I. Heck, E.G. De Mejia, Yerba Mate Tea (Ilex paraguariensis): a comprehensive review on chemistry, health implications, and technological considerations, J. Food Sci. 72 (2007) R138ςR151.
[25] R. Filip, S.B. Lotito, G. Ferraro, C.G. Fraga, Antioxidant activity of Ilex paraguariensis and related species, Nutr. Res. 20 (2000) 1437ς1446.
[26] M.F. Matei, R. Jaiswal, M.A. Patras, N. Kuhnert, LC-MSn study of the chemical transformations of hydroxycinnamates during yerba maté (Ilex paraguariensis) tea brewing, Food Res. Int. 90 (2016) 307ς312.

[27] V.H. Techio, A. Cagliari, P.A. Floss, D.M. da Croce, Morfometria e nervação foliar em procedências de erva-mate (Ilex paraguariensis A. St. Hill.)(Aquifoliaceae), Acta Sci. Biol. Sci. 31 (2009) 433ς437.
[28] J.S.C. FERNANDES, S. Ushiwata, R. de Mattos Daminelli, J. Gabardo, M. Kobiyama, A.M. Junior, C. Prevedello, R.M.S. Resende, M.D.V. RESENDE, J.A. Sturion, Estimativas de parâmetros relacionados ao melhoramento genético da erva-mate: possibilidade de seleção precoce, Sci. Agrária. 1 (2000) 45ς53.
[29] I φ ¢ăǑK2ŜLJΣ ¸φ Dλō2yΣ tφ / I·ℕΛff2Σ tφ ! ℕYΣŷ∃I·ȡzŘΣ a φ {ÎŜǑ2ǿ⌉I·Σ ! φ b ¦ b9{πb9{LΣ ! φw φ CŜℕyΛŜΣ Y φ Koehl, M. StittΣ ! Řƨdzǎű Y Σŷű 2Ŧ ∃ℕ2ǿűK I·yŘ ǑŜyűℕI·f Y ŜűI·ō2fλǎY ű2 I· Y ΛfŘ ōdzű ǎdzǎűI·λyŜŘ yΛúℕ2∃Ŝyπf Λ Y λűI·ûΛ2y in Arabidopsis, Plant. Cell Environ. 32 (2009) 300ς318.
[30] L.G. Riachi, C.A.B. De Maria, Yerba mate: An overview of physiological effects in humans, J. Funct. Foods. 38 (2017) 308ς320.
[31] F.P. de Sá, D.C. Portes, I. Wendling, K.C. Zuffellato-Ribas, Miniestaquia de erva-mate em quatro épocas do ano, Ciência Florest. 28 (2018) 1431ς1442.
[32] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, P. Dörmann, W. Weckwerth, Y. Gibon, M. Stitt, GMD@ CSB. DB: the Golm metabolome database, Bioinformatics. 21 (2005) 1635ς1638.
[33] N. Schauer, D. Steinhauser, S. Strelkov, D. Schomburg, G. Allison, T. Moritz, K. Lundgren, U. Roessner-Tunali, M.G. Forbes, L. Willmitzer, A.R. Fernie, J. Kopka, GC-MS libraries for the rapid identification of metabolites in complex biological samples, FEBS Lett. 579 (2005) 1332ς1337. https://doi.org/10.1016/j.febslet.2005.01.029.
[34] J. Yang, X. Zhao, X. Lu, X. Lin, G. Xu, A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis, Front. Mol. Biosci. 2 (2015) 1ς9. https://doi.org/10.3389/fmolb.2015.00004.
[35] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, Y. Ni, Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data, Sci. Rep. 8 (2018) 1ς10. https://doi.org/10.1038/s41598-017-19120-0.
[36] R.C. Team, R: a language and environment for statistical computing [online]. R Foundation for Statistical Computing, Vienna, Austria, (2018).
[37] A.D. Adrian, A. Cole, xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files R package version 0.6.1, (2018).
[38] T. Hothorn, F. Bretz, P. Westfall, Simultaneous inference in general parametric models, Biometrical J. J. Math. Methods Biosci. 50 (2008) 346ς363.
[39] D. Sarkar, Lattice: multivariate data visualization with R, Springer Science & Business Media, New York, 2008.
[40] F.A. de Mendibru, Statistical procedures for agricultural research. R package version 1.2-8., (2017).
[41] R. de Andrade Moral, J. Hinde, C. Garcia Borges Demétrio, Half-normal plots and overdispersed models in R: The hnp package, J. Stat. Softw. 81 (2017).
[42] E.B. Ferreira, P.P. Cavalcanti, D.A. Nogueira, ExpDes: an R package for ANOVA and experimental designs, Appl. Math. 5 (2014) 2952.
[43] J. Fox, Effect displays in R for generalised linear models, J. Stat. Softw. 8 (2003) 1ς27.

[44] V. den P.D. Meire M, Ballings M, Compute missing values on a training data set and impute them on a new data set. Current available options are median/mode and random forest., R Packag. Version 3.5.3. (2016).
[45] C. Leys, S. Schumann, A nonparametric method to analyze interactions: The adjusted rank transform test, J. Exp. Soc. Psychol. 46 (2010) 684ς688. https://doi.org/10.1016/j.jesp.2010.02.007.

Figure captions

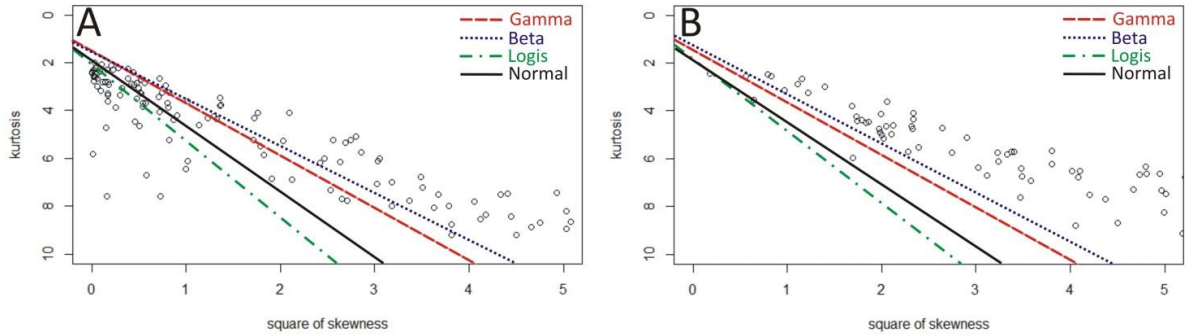

Figure 1. Random response variables with specific distribution shapes, as determined by kurtosis and square of skewness parameters, were used to determine the regression family model for a GLM evaluating both N dose response (A, n=7617) and comparative Amazonian soils (B, n=3393) datasets. Kurtosis was plotted in the ordinates and square of skewness in the abscissa and the random vectors with specific distributions (i.e. normal, logis, beta and gamma) were simplified by average-linearization and plotted in the curves. The R function used to build the distributions is reported in RandodiStats GitHub repository (https://github.com/MSeidelFed/RandodiStats).

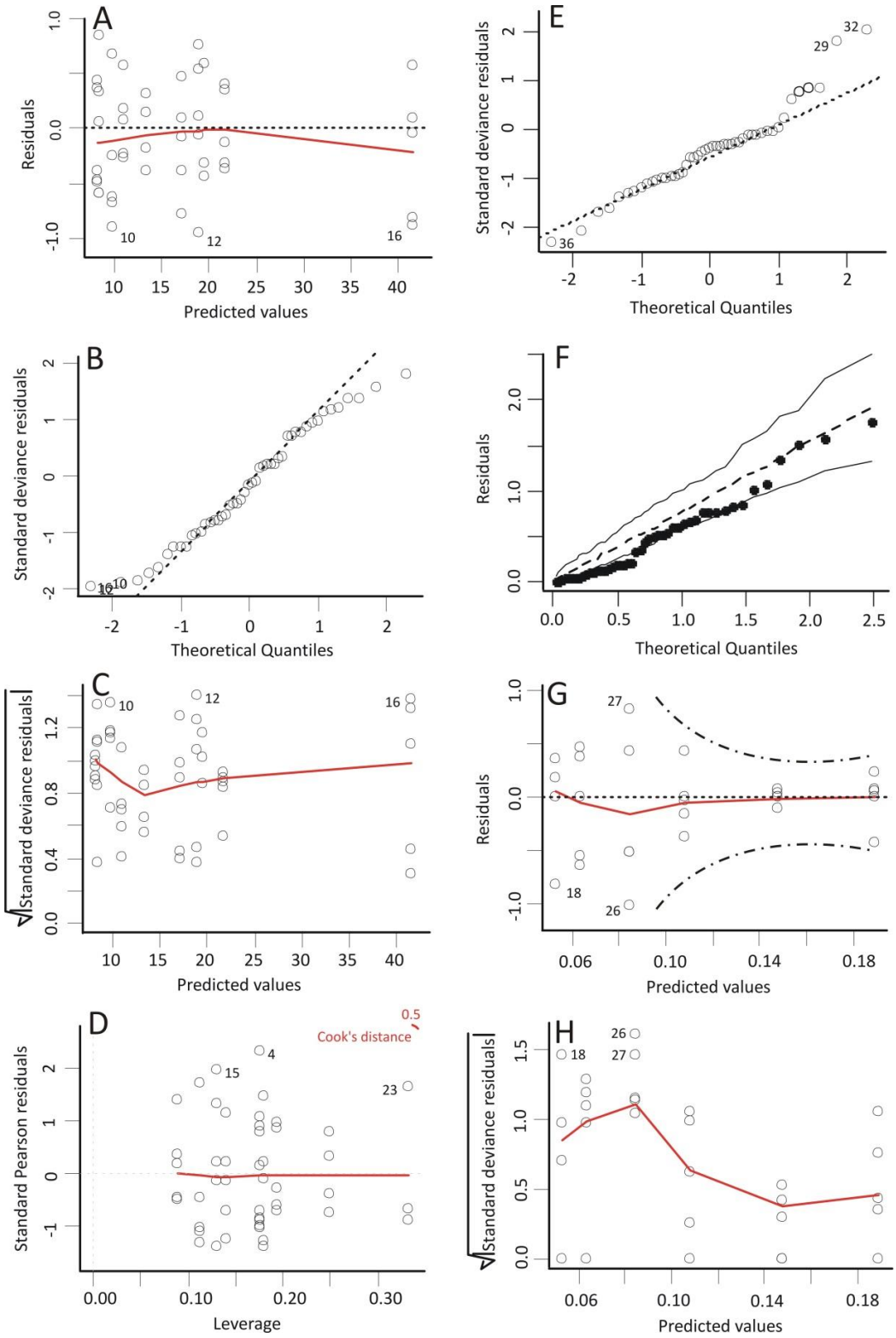

Figure 2. Diagnostic plots; residual vs fitted (A), quantile-quantile (B), scale-location (C), residual vs leverage (D) quantile-quantile (E), and half-normal plot with simulation envelopes (F) residual vs fitted (G), and scale-location (H). Featuring shikimic acid (C17) (A-D) plots that fulfils the underlying

assumptions for GLM evaluation with a gamma family distribution. Secondly, 2-hydroxy-glutamic acid (C8) (E-F) plots that shown a violation of the assumption of normal distribution of the residuals and hence must be evaluated with a non-parametric test (NP). The last ones, eicosanoic acid (C15) (G-H) plots indicating that the variable should be treated as NP since the residual violates the assumption of homoscedasticity of variance across experimental conditions. Numbers in the plots indicate the extreme variables in the model, according to variable position in dataset (row number). Dotted lines indicates the ideal distribution of the residues, and solid line show the real distribution of the residues.

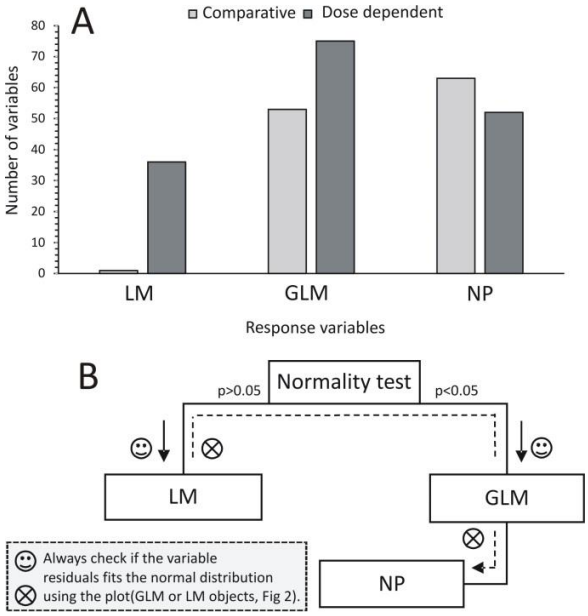

Figure 3. Histogram of the types of response variables (i.e. LM, GLM and, NP) counted in the comparative (Amazonian soils) and the N-dose fertilizer datasets (A). Flowchart showing how to select response variable to be used in the statistical analysis (B). The R scripts used are reported in Stats_Fp_GC GitHub repository (https://github.com/FAHansel/Stats_Fp_GC).

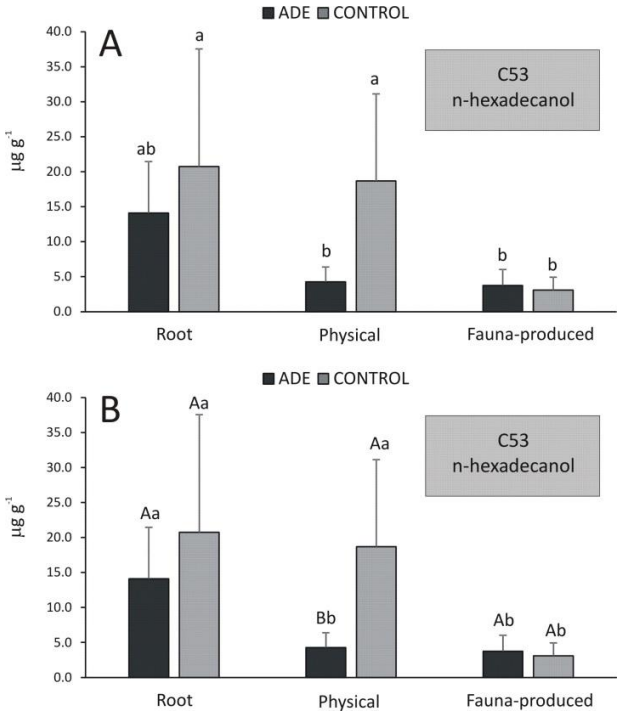

Figure 4. Histograms of selected GLM response variable compound exemplifying the results of the comparative study of Amazonian soils (ADE and control) and aggregates (root, physical and fauna-produced), comparing all samples (A), and analysing the main effects (F1 and F2) separately with comparisons within the levels (B). According to the Tukey contrasts ($p<0.05$), mean values followed by the same letter do not differ statistically, in B plot capital letters refer to F1 (soil, ADE x control) comparison and, lowercase letters are associated to F2 (aggregates root x physical x fauna-produced) comparison in F1 separately (ADE or control). Interestingly, for ADE soils, no differences was detected in the aggregates (A), but when the main effects were analysed separately (soil and aggregates), and the comparison was made between soils and between aggregates in the same soil a differences was observed in the ADE soil aggregates (B).

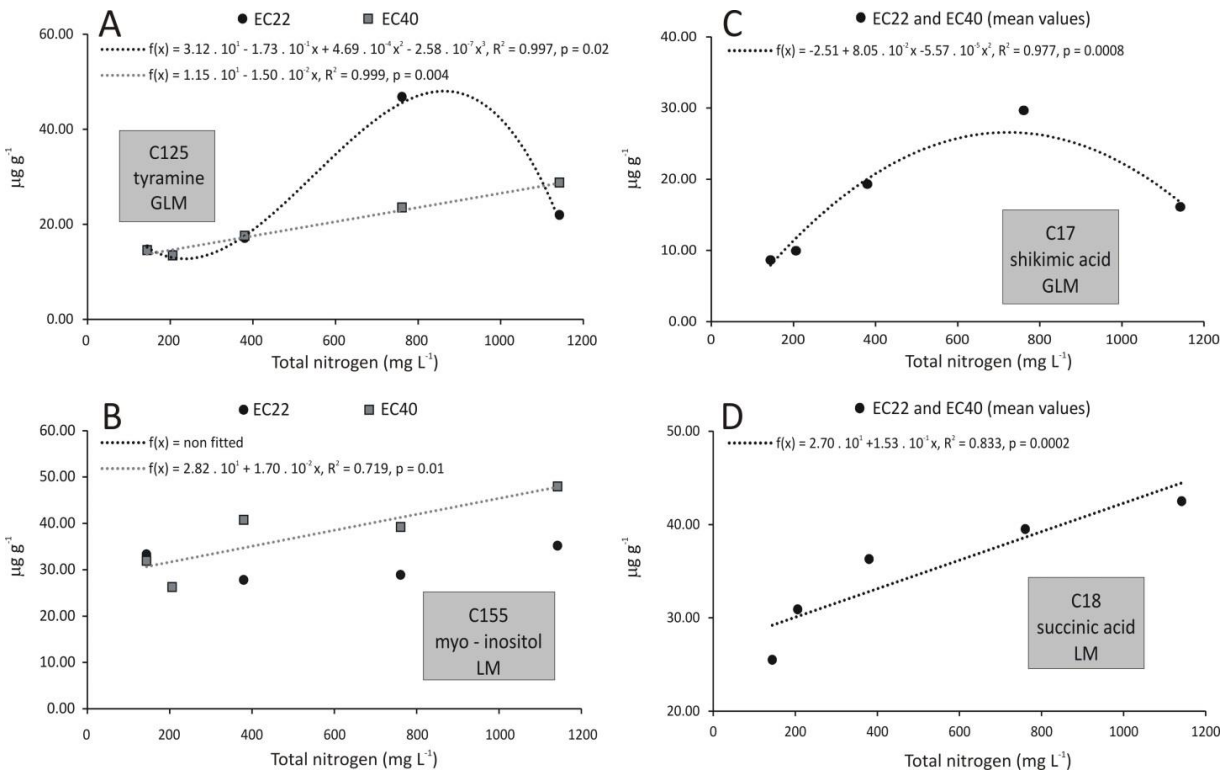

Figure 5. Regression curves (dot lines) and mean values (symbols) of selected compounds exemplifying the Ilex N-fertilizer dose study using interaction statistic treatment (F1 (clones x F2 (N-dose). GLM and LM response variables with interaction (F1xF2, A and B) with both clones plotted separately, and no-interaction (only F2 significant, C and D) with the mean values of both clones considered. Cx refers to the compound abbreviations used in R software.

Table 1. Selected compounds exemplifying the results[a] of the Amazonian soils comparative dataset using interaction statistic treatment.

				F2				
				Root	Physical	Fauna-produced		
Compound	R code[b]	Response variable[c]	F1		µg g^{-1}		Main effect	Significance[d]
n-eicosanoic acid	C15	NP	-	17.2 ± 7.6^{a}	8.1 ± 2.3^{b}	8.6 ± 6.9^{b}	F2	**
n-docosanol	C51	NP	ADE	16.3 ± 7.8^{Aa}	5.6 ± 3.3^{Bb}	4.1 ± 0.7^{Bb}	F1 & F2	**
			control	24.3 ± 5.9^{Aa}	34.0 ± 15.6^{Aa}	9.4 ± 2.8^{Ab}		*
n-hexadecanol	C53	GLM	ADE	14.1 ± 7.3^{ab}	3.8 ± 2.2^{b}	3.8 ± 2.3^{b}	F1xF2	*
			control	20.7 ± 16.8^{a}	18.7 ± 12.5^{a}	3.1 ± 1.8^{b}		
Compound	R code	Response variable	F1	Soils (µg g^{-1})				
phosphoric acid	C5	NP	ADE	9.5 ± 5.9^{A}	-	-	F1	***
			control	0.8 ± 0.0^{B}	-	-		
hexadec-9-enoic acid	C8	GLM	ADE	90.2 ± 68.8^{A}	-	-	F1	**
			control	40.8 ± 30.1^{B}	-	-		
n-docosanoic acid	C13	GLM	ADE	8.8 ± 6.7^{B}	-	-	F1	*
			control	16.3 ± 9.3^{A}	-	-		

According to the post hoc tests, mean values followed by the same letter do not differ statistically, capital and lowercase letter refer to F1 (ADE x control, column) and, F2 (root x physical x fauna-produced, row) comparisons, respectively; a: the raw data were used for statistical treatment, in case of transformed data (e.g. $\log_{10}$) the same criteria presented in figure 2 must be followed after transformation; b: compound abbreviations used in R software; c: NP ς non-parametric, GLM ς general linear model, the response variable with normal distribution (LM) did not result in statistical differences; d: significance levels in the statistical analyses: *p < 0.05, **p < 0.01, ***p < 0.001.

Graphical abstract