CHEMICAL ENGINEERING TRANSACTIONS

VOL. 105, 2023

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: David Bogle, Flavio Manenti, Piero Salatino
Copyright © 2023, AIDIC Servizi S.r.l.
ISBN 979-12-81206-04-5; **ISSN** 2283-9216

DOI: 10.3303/CET23105078

Canonical Correlation Analysis to Biomass CHONS Prediction

Federico Moretta[a], Vincenzo Del Duca[b,*], Giulia Bozzano[a], Antonio Coppola[c], Fabrizio Scala[b,c]

[a]Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133, Milan (MI), Italy
[b]DICMaPI, Università degli Studi di Napoli Federico II, Piazzale Vincenzo Tecchio 80, 80125 Naples (NA), Italy
[c]STEMS, Consiglio Nazionale delle Ricerche, piazzale Vincenzo Tecchio 80, 80125, Napoli (NA), Italy
idelducavincenzo@gmail.com

Fermentation biomasses can be defined as a complex mixture of different natural components and microbes, having biodegradable and organic waste as the primary source. Its correct characterization is crucial to have proper processing in fermentative units. Firstly, proximate analysis is done to retrieve the content of specific compounds in the mixture, such as fat, proteins, and carbohydrates. However, this is often not enough to achieve the sufficient precision, since some low-concentration species are not easily found through this methodology (i.e., sulfate compounds, ethanol, caproic acid). Consequently, ultimate analysis is performed to evaluate the exact amount of every element in the mixture. For biomass-based compounds, atoms content can be synthesized in carbon, hydrogen, oxygen, nitrogen, and sulfur. The total content of these elements is also known as CHONS. From this, it is possible to derive the exact amount of the relative species in the biomass. However, the experimental procedure for its determination is rather time and budget-consuming. On the other hand, the amount of data collected in the literature, from both experimental and industrial analysis, can be exploited to build a numerical model, based on the multivariate statistical analysis and machine learning principles that predict the CHONS content for every type of biomass. In this work, a data-driven model has been developed to achieve this aim, having as input a set of relevant variables. Consequently, a dataset has been built to gather all these data. The multivariate statistical technique of Canonical Correlation Analysis (CCA) is used to find 'hidden' correlations and predict CHON content for 27 different biomass types. In future research, machine learning techniques will be applied to compare the results obtained.

1. Introduction

Nowadays, the scientific world is facing environmental burdens that can modify consistently the energy production and supply. In this scenario, local biomass for the production of green energy vectors (e.g., biomethane, biomethanol) can become a significant mitigation resource (World Energy Outlook 2022 – Analysis). Biomasses are very heterogeneous, and consequently, it is important to characterize them before use. Macromolecules concentration, such as carbohydrates, proteins and lipids, are necessary to predict the biodegradability of the system in digestion systems (Triolo et al., 2011). On the other hand, the elemental fraction of carbon, nitrogen, hydrogen and oxygen is necessary information to predict the theoretical biomethane potential through the Buswell formula (Achinas and Euverink, 2016). These characteristics are related to one another. For example, a high protein concentration averagely brings a higher nitrogen fraction due to the abundance of the nitrogen-rich functional group of the amino acids. It is of course possible to find an optimum between all these properties when mixing more biomasses, to regulate specific properties such as the C/N ratio (Guarino et al., 2016). However, the settings of the optimal parameters, for the desired biomass conversion unit, need a relative number of experimental analyses to find all the necessary properties. Proximate analysis is done to find ash, moisture, fixed carbon and volatile components amount of the biomass. The compositional analysis is performed to find the macromolecules, including lignin and cellulose. Elemental analysis instead evaluates carbon, hydrogen, oxygen, nitrogen and sulfur percentage (CHONS) in the biomass. All these procedures can become much time-consuming and budget-consuming when a lot of feedstocks have to be tested. Nowadays, experimental procedures can easily be overridden by data-driven models which, exploiting all the data coming

Paper Received: 25 February 2023; Revised: 27 June 2023; Accepted: 3 May 2023
Please cite this article as: Moretta F., Del Duca V., Bozzano G., Coppola A., Scala F., 2023, Canonical Correlation Analysis to Biomass CHONS Prediction, Chemical Engineering Transactions, 105, 463-468 DOI:10.3303/CET23105078

from past experiments, elaborate reliable predictions in the observed range (Chang et al., 2021). This work focused attention on the individuation of specific correlations related to the calculation of the CHONS fraction starting from other biomass attributes, representing the first attempt for building up a decision-support system seeking the biomass perfect amount and characteristic (Del Duca et al., 2022). For this aim, the dataset from Moretta et al. (2022) has been used, which gathers enough data to perform the study. The tool adopted for the application of multivariate statistical analysis is the Canonical Correlation Analysis (CCA). Moreover, correlation and other statistical indexes were used to find which of these attributes relate better to the evaluation. The database has been pre-processed according to the tool input-output system, generating a new ad-hoc dataset, which directly undergoes the mathematical procedure.

2. Methodology

Canonical Correlation Analysis (CCA) (Wilks, 2019) is widely used to extract the correlated patterns between two sets of variables, ***x*** and ***y***. ***x*** generally represents the input matrix with size (*n x m*), where *n* is the number of observations and *m* is the number of specific input variables; conversely***, y*** represents the output matrix with size (*n x J*), where *J* is the number of specific output variables.
CCA looks at two sets of variables for modes of maximum correlation between the two sets. Thus, CCA sits at the top of a hierarchy of regression models where it can manage multiple predictors (inputs) and multiple predictands (outputs). . If ***x*** is the set of predictors and ***y*** the predictands, then CCA can be used to predict ***y*** when new observations of ***x*** become available (Hsieh, 2000). The method finds linear combinations of the original variables by projecting them onto coefficient vectors a_m and b_m, which is chosen such that each pair of the new variables v_m and w_m, called *canonical variates*, exhibit maximum correlation, while being uncorrelated with the projections of the data onto any of the other identified patterns (Eq. 1a-1b).

$$v_m = a_m^T x' = \sum_{i=1}^{I} a_{m,i} x_i' \,, \quad m = 1, \dots, \min\,(I,J) \tag{1a}$$

$$w_m = b_m^T y' = \sum_{j=1}^{J} b_{m,j} y_j' \,, \quad m = 1, \dots, \min\,(I,J) \tag{1b}$$

In other words, CCA identifies new variables that maximize the interrelationships between two data sets in this sense. The vectors of linear combination weights a_m and b_m, are called the *canonical vectors*. The number of pairs, M, of canonical variates that can be extracted from the two data sets is equal to the smaller of the dimensions of ***x*** and ***y***. The canonical vectors a_m and b_m is the choices that result in the canonical variates (Eq. 2a-2d) having the following properties:

$$Corr(v_1, w_1) \geq Corr(v_2, w_2) \geq \cdots \geq Corr(v_M, w_M) \geq 0 \tag{2a}$$

$$Corr(v_k, w_m) = \begin{cases} r_{C_m}, & k = m \\ 0, & k \neq m \end{cases} \tag{2b}$$

$$Corr(v_k, v_m) = Corr(w_k, w_m) = 0, \;\; k \neq m \tag{2c}$$

$$Var(v_m) = Var(w_m) = 1, \;\; m = 1, \dots, M \tag{2d}$$

Equation 2a states that each of the M successive pairs of canonical variates exhibits no greater correlation than the previous pair and these correlations between the pairs of canonical variates are called the canonical correlations, r_C, where [R_C] is the diagonal matrix (Eq. 3).

$$[R_C] = \begin{bmatrix} r_{C_1} & 0 & 0 & \cdots & 0 \\ 0 & r_{C_2} & 0 & \cdots & 0 \\ 0 & 0 & r_{C_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r_{C_M} \end{bmatrix} \tag{3}$$

Equations 2b and 2c state that each canonical variate is uncorrelated with all the other canonical variates except its specific counterpart in the m^th^ pair, finally equation 2d states that each of the canonical variates has a variance equal to 1. The basic idea behind forecasting with CCA is straightforward: simple linear regressions are constructed that relate the predictand canonical variates w_m to the predictor canonical variates v_m (Eq. 4).

$$\hat{w}_m = \hat{\beta}_{0,m} + \hat{\beta}_{1,m} v_m, \;\; m = 1, \dots, M \tag{4}$$

Where $\hat{\beta}_{0,m}$=0. Because the CCA is calculated from the centred data ***x'*** and ***y'***, and $\hat{\beta}_{1,m}$=r_{C_m}, since the canonical variates are scaled to have unit variance, the regression slopes are simply equal to the corresponding canonical correlations.

The database used for the statistical analysis has been built up by the data presented in the scientific literature and coming from real plant analysis. A total of 215 observations have been analysed. The input data ***x*** has been constituted by a combination of ultimate and proximate analysis of the raw biomass, while the output data ***y*** reports the CHONS percentage of biomass. Table 1 summarizes the input and output variables used for CCA. Among a total number of 21 biomass attributes (i.e., total solids, macromolecules, etc.), only a small combination of them have been considered as input, looking for the minimum error achievable.

*Table 1: Variables of **x** and **y** used for CCA statistical analysis.*

X data	Definition	***y*** data	Definition
TS	Total solids in the biomass, as %w/w of the total mass	C	Carbon fraction, as %mol.
VS	Volatile solids of biomass, as %TSw	H	Hydrogen fraction, as %mol.
SU	Sugar concentration in biomass, as %TSw	O	Oxygen fraction, as %mol.
PR	Protein concentration in biomass, as %TSw	N	Nitrogen fraction, as %mol.
LP	Lipids concentration in biomass, as %TSw		
LG	Lignin concentration in biomass, as %TSw		
HCE	Hemicellulose concentration in biomass, as %TSw		
BD	Biodegradability index of the biomass		

%TSw = weight percentage of the total solids in the biomass.

The sulfur molar fraction in biomass (S) is evaluated as complementary to the other elements, since its average value is relatively lower than the other species and can therefore influence the analysis results increasing its stiffness. Sulfur content can be reliably predicted from the average content of cysteine and methionine. However, the database used in this study does not have this information. In Table 1, TS is the total solids content in the biomass, VS represents the actual amount of elements which can undergo digestion, and C/N is an important parameter since gives stability to the evaluation of the carbon and nitrogen output. Similarly, BD indicates the biodegradability of the biomass and is directly correlated to its bio-methane potential. Other input parameters (SU, PR, LP, LG, HCE) are the macromolecule concentration relative to that substrate under study. The CCA analysis has been implemented in MATLAB™ through the command *canoncorr(**x**,**y**)*, which computes the canonical vectors a_m and b_m, and the canonical variates v_m and w_m. The performance indicators MAE, MSE, SAE and MAPE have been calculated (Table 2) in the evaluation of the correlations between ***x*** and ***y.***

MAE = mean absolute error; MSE = mean square error; SAE = sum absolute error; MAPE = mean absolute percentage error.

3. Results

Equations from 5 to 8 report the linear correlations obtained by CCA from ***x*** and ***y*** data for the CHONS evaluation. As said, and as also shown in Table 1, the S fraction is not directly predicted, since it can be evaluated as a complementary component of the mixture. Despite the relatively high error obtained from the analysis (Table 2), it must be considered that the heterogenicity of the biomass and the goodness of the experimental analyses strictly influence the values in the dataset. This can be also seen in Figure 1, where it is possible to notice the presence of enough outliers in the CNHO estimation exiting the confidence interval, thus increasing the error. This is principally due to the variability of the data-taking process, since the biomasses are extremely heterogenous and the fraction of a certain type of it can consistently differ to another one. It is possible to say that the CHONS content is strictly related to the relative content of the biomass itself. Consequently, it is are obtained for the nitrogen and hydrogen fractions. However, the MAPE index put hydrogen in second place, valorising the carbon and nitrogen fraction, showing values lower than 16%.

Table 2: Results evaluated in terms of error performance indicators.

Element	*MAE*	*MAPE*	*SAE*	*MSE*
C	6.3	15.9	1351.6	85.2
H	1.5	78.0	315.8	4.7
N	0.9	15.0	201.6	2.5
O	7.7	135.1	1663.6	127.6

The oxygen is the output which has the highest error in the study. This can be caused by the high number of outlier values in the database. This depicts the difficulty to find both experimentally, in terms of reliable analysis, and numerically the amount of oxygen present in the biomass solids fraction.

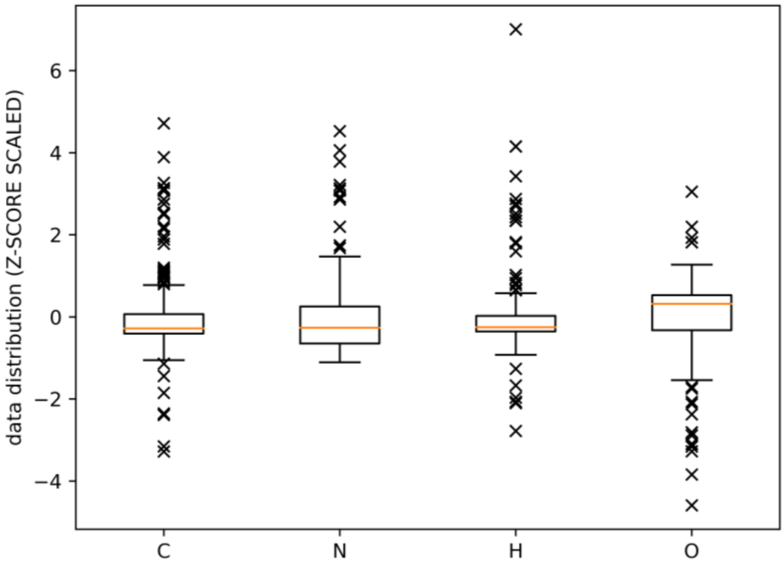


Figure 1: boxplot and outliers' representation of CNHO values from the database, scaled with the z-score algorithm.

Figure 2 shows a comparison between the database and predicted values (Eq. 5-8) for all outputs considered. Though this comparison it is easier to look at the goodness of the prediction. CCA predictions well represent the trends of the molar fraction of the species.

$$C_{RM} = 0.069 \times TS + 0.135 \times VS + 0.001 \times LP + 0.123 \times PR - 0.034 \times SU - 0.069 \times LG + 0.088 \times HCE + 2.715 \times BD + 25.154 \quad (5)$$

$$N_{RM} = 6.76 \times 10^{-5}\, TS + 0.012 \times VS + 0.030 \times LP + 0.174 \times PR + 0.032 \times SU + 0.023 \times LG - 0.006 \times HCE - 3.401 \times BD + 0.701 \quad (6)$$

$$H_{RM} = 0.003 \times TS + 0.025 \times VS + 0.015 \times LP + 0.011 \times PR + 6.06 \times 10^{-4} \times SU + 0.012 \times LG + 0.024 \times HCE - 0.170 \times BD + 3.261 \quad (7)$$

$$O_{RM} = -0.072 \times TS - 0.179 \times VS - 0.051 \times LP - 0.352 \times PR - 0.008 \times SU + 0.029 \times LG - 0.108 \times HCE + 1.491 \times BD + 71.351 \quad (8)$$

RM = Raw Material.

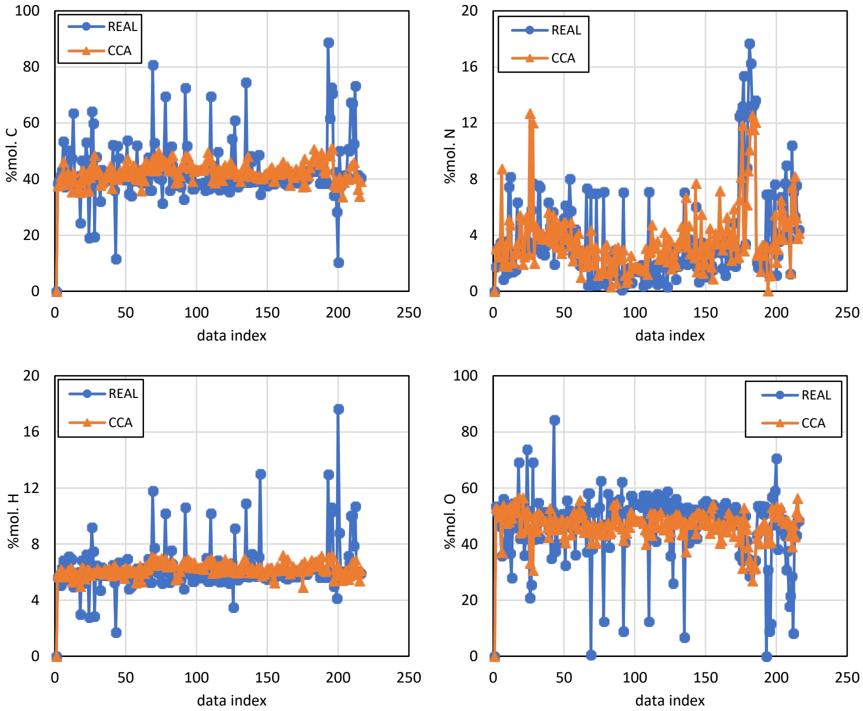
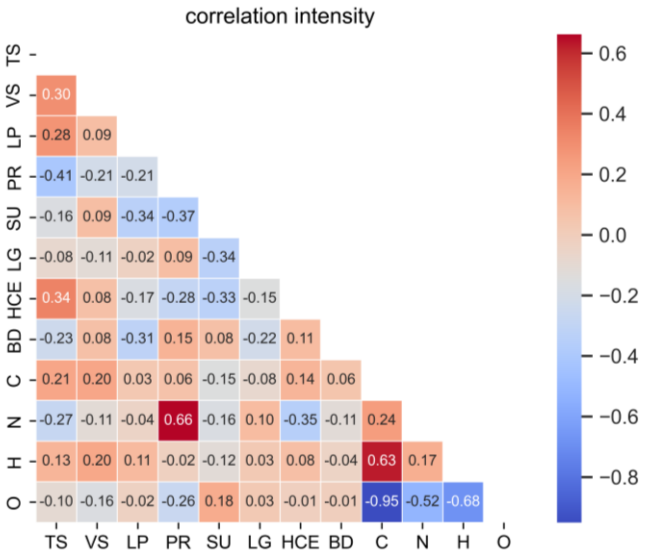


Figure 2: Comparison of the predicted value of the CCA model with database values.



	TS	VS	LP	PR	SU	LG	HCE	BD	C	N	H	O
TS												
VS	0.30											
LP	0.28	0.09										
PR	-0.41	-0.21	-0.21									
SU	-0.16	0.09	-0.34	-0.37								
LG	-0.08	-0.11	-0.02	0.09	-0.34							
HCE	0.34	0.08	-0.17	-0.28	-0.33	-0.15						
BD	-0.23	0.08	-0.31	0.15	0.08	-0.22	0.11					
C	0.21	0.20	0.03	0.06	-0.15	-0.08	0.14	0.06				
N	-0.27	-0.11	-0.04	0.66	-0.16	0.10	-0.35	-0.11	0.24			
H	0.13	0.20	0.11	-0.02	-0.12	0.03	0.08	-0.04	0.63	0.17		
O	-0.10	-0.16	-0.02	-0.26	0.18	0.03	-0.01	-0.01	-0.95	-0.52	-0.68	

Figure 3: Heatmap of the correlation matrix.

CCA has been used to find hidden linear correlations between the input selected and the output. Figure 3 shows the heatmap of the Pearson correlation intensity between all these parameters. From this figure, the higher the better the two features are correlated with one another. As it is possible to see, nitrogen and protein fractions are strongly correlated, due to the presence of the ammonium functional group in the amino acid. Furthermore, carbon and hydrogen fractions are also strongly correlated to each other, since many organic carbon-rich species have a high C/H ratio. On the other hand, both carbon and nitrogen have an inverse correlation with oxygen. Oxygenated species in the solid phase (i.e., carboxylic acids) are really poor, and the principal source of oxygen comes from sugars and ethanol. The oxygen path ends with the decomposition of these species into CO_2 and H_2O. The intensity of the other parameter correlation factor is high enough for the model, but not to make any mention in this specific discussion. Whereas Pearson correlations focus on relationships within a single dataset, canonical correlations assess the relationships between two distinct datasets, allowing one to identify the optimal linear combinations between the variables in the two sets that show the maximum *correlation.*

4. Conclusions

This study looked at using statistical analysis to find linear correlations between different variables. Specifically, a technique called CCA was used to uncover “hidden” linear correlations between input and output variables, and to predict the CHONS content for different types of biomasses. The results of the analysis showed that while there were advantages in using CCA, there were also some drawbacks. One advantage was that it was easy to implement, although the dataset needed to be carefully constructed. However, the main limitation of CCA was that it could only identify linear correlations between inputs and outputs, which resulted in some errors in the predictions made in this study. This limitation could be overcome by using machine learning techniques. In future research, machine learning techniques will be applied to the data to find stronger relations between the biomass features. Overall, the CCA approach represented an initial attempt to create simple tools for the prediction of CHONS content for every type of biomass.

References

Achinas, S., Euverink, G.J.W., 2016. Theoretical analysis of biogas potential prediction from agricultural waste. Resource-Efficient Technologies 2, 143–147. https://doi.org/10.1016/j.reffit.2016.08.001

Chang, J., Kim, J., Zhang, B.-T., Pitt, M.A., Myung, J.I., 2021. Data-driven experimental design and model development using Gaussian process with active learning. Cognitive Psychology 125, 101360. https://doi.org/10.1016/j.cogpsych.2020.101360

Del Duca, V., Chirone, R., Coppola, A., Scala, F., Salatino, P., 2022. Application of Multivariate Statistical Analysis for Pyrolysis Process Optimization. Chemical Engineering Transactions 96, 277–282. https://doi.org/10.3303/CET2296047

Guarino, G., Carotenuto, C., Cristofaro, F.D., Papa, S., Morrone, B., Minale, M., 2016. Does the C/N ratio really affect the Bio-methane Yield? A three Years Investigation of Buffalo Manure Digestion. Chemical Engineering Transactions 49, 463–468. https://doi.org/10.3303/CET1649078

Hsieh, W.W., 2000. Nonlinear canonical correlation analysis by neural networks. Neural Networks 13, 1095–1105. https://doi.org/10.1016/S0893-6080(00)00067-8

Moretta, F., Goracci, A., Manenti, F., Bozzano, G., 2022. Data-driven model for feedstock blending optimization of anaerobic co-digestion by BMP maximization. Journal of Cleaner Production 375, 134140. https://doi.org/10.1016/j.jclepro.2022.134140

Triolo, J.M., Sommer, S.G., Møller, H.B., Weisbjerg, M.R., Jiang, X.Y., 2011. A new algorithm to characterize biodegradability of biomass during anaerobic digestion: Influence of lignin concentration on methane production potential. Bioresource Technology 102, 9395–9402. https://doi.org/10.1016/j.biortech.2011.07.026

Wilks, D.S., 2019. Multivariate Analysis of Vector Pairs. Statistical Methods in the Atmospheric Sciences 669–694. https://doi.org/10.1016/b978-0-12-815823-4.00014-6

World Energy Outlook 2022 – Analysis [WWW Document], n.d. . IEA. URL https://www.iea.org/reports/world-energy-outlook-2022 (accessed 3.31.23).