

# Application Of Multivariate Statistical Analysis For Pyrolysis Process Optimization

Vincenzo Del Duca<sup>a,\*</sup>, Roberto Chirone<sup>b</sup>, Antonio Coppola<sup>a</sup>, Fabrizio Scala<sup>a,b</sup>, Piero Salatino<sup>a,b</sup>

<sup>a</sup>STEMS, Consiglio Nazionale delle Ricerche, Piazzale Vincenzo Tecchio 80, 80125 Napoli, Italy

<sup>b</sup>DICMaPI, Università degli Studi di Napoli Federico II, Piazzale Vincenzo Tecchio 80, 80125 Napoli, Italy

[vincenzodelduca@stems.cnr.it](mailto:vincenzodelduca@stems.cnr.it)

The identification of the most efficient biomass valorization paths is vital for reaching the target of Renewable Energy Sources consumption by 2030. In this context, within a National project named ‘Biofeedstock’, the applicability of multivariate statistical analysis, i.e. Canonical Correlation Analysis (CCA), is implemented for the definition of specific correlations describing quantitatively and qualitatively the fast pyrolysis process outputs. The database used for the CCA contains 59 observations and it has been built up using literature data specifically on fluidized bed fast pyrolysis without any catalyst, in the temperature range of 450-550°C. The results show that the CCA correctly describes the process analysed with a discrete degree of confidence. However, it shows two main drawbacks, firstly the dataset constitution, and secondly possibility to individuate only linear correlations between inputs and outputs.

## 1. Introduction

Concerns related to environmental burdens and security of energy supply are stimulating the exploitation of residual biomass for production of biochemical platform and biofuel platform (IEA (2019), *World Energy Outlook 2019*, IEA, Paris <https://www.iea.org/reports/world-energy-outlook-2019>, 2019). In December 2018, the European Commission published the new renewable energy directive 2018/2001/EU, known as RED II (Council, 2018; European Union, 2009). The overall target of Renewable Energy Sources consumption by 2030 was increased to 32% and a minimum share of 14% of renewable energy consumed in road and rail transportation is targeted by 2030. Furthermore, the lower bounds to contribution of advanced biofuels as a share of final energy consumption in the transport sector are set at 0.2 % in 2022, 1 % in 2025 and 3.5 % in 2030 (Littlejohns et al., 2018). Substitution of fossil fuels with biofuels aims at minimizing environmental burdens related to both production and consumption (Littlejohns et al., 2018) as well as at decreasing the net CO<sub>2</sub> emissions (Colling Klein et al., 2018). Biofuels are classified as first-, second- and third-generation biofuels based on the carbon source of biomass feedstock (Ziolkowska, 2020). First-generation biofuels, which are produced directly from food crops, are attractive from a techno-economic perspective, as they show higher conversion efficiency and lower costs compared to the others. However, societal constraints as food vs fuel competition are driving regulations and markets toward exploitation of residual biomass and non-food crops, namely second- and third-generation biofuels i.e. cellulosic/waste biomass and algae (Hirani et al., 2018).

Moreover, engineering supply systems that deliver affordable, high-quality biomass or “biofeedstock” are a challenge for the emerging bioenergy industry (Lamers et al., 2015). Biomass feedstocks are distributed on broad spatial and temporal scales and have widely different physical and chemical properties (Razik et al., 2019). “Biofeedstock”, is a National project funded by the Italian Ministry of University and Research, comprising 12 Italian industrial and academic partners, aimed at the development of smart technology platforms for the residual biomass valorization paths. The basic idea behind the project is the development of extended supply chains based on decentralized biomass harvesting and preprocessing stages for the production of “biofeedstocks”, namely biogenic energy carriers.

“Biofeedstocks” may conform to specification standards to represent tradable commodities. They can be eventually upgraded at centralized processing sites or biorefineries for the generation of end products (biofuels and biochemicals) of commercial interest. In the frame of the Biofeedstock project, viable conversion routes belong to either thermochemical or biochemical pathways. The thermochemical pathways include slow, fast and catalysed pyrolysis, gasification, torrefaction and hydrothermal liquefaction, while the biochemical pathways include anaerobic and aerobic fermentation of organic substrates. The comparative assessment of alternative valorization strategies may be accomplished by assuming relevant objective functions expressing the yield and quality of the biofeedstock as well as the fate of pollutant precursors.

Obviously, the comparison among the different processes is possible if models describing the performance of single process, in terms of quantity and quality of the products, are available. Generally, models adopted for process simulations are based on equations, which describe the chemical-physical phenomena. This approach offers high reliability, however the development of models based on physicochemical relations are expensive in terms of time and money. The alternative could be represented by the application of statistical analysis exploiting dataset edited by scientific literature. The latter approach provides less insights and quality of the process description compared to the other approach, however it could be useful for the constitution of ‘light’ tools for decision-support systems related to biomass valorisation. This work focused the attention on the individuation of specific correlations related to the fast pyrolysis process representing the first attempt for the building up of the decision-support system. The tool adopted for the implementation of the fast pyrolysis model was the application of multivariate statistical analysis, in particular Canonical Correlation Analysis (CCA). The database used for the statistical analysis has been edited by the data presented in the scientific literature.

## 2. Methodology

Canonical Correlation Analysis (CCA), described for the first time by Harold Hotelling in 1936 (Hotelling, 1992)<sup>1</sup>, is widely used to extract the correlated patterns between two sets of variables,  $\mathbf{x}$  and  $\mathbf{y}$ .  $\mathbf{x}$  generally represents the input matrix with size  $(n \times I)$ , where  $n$  is the number of observations and  $I$  the number of specific input variables; conversely,  $\mathbf{y}$  represents the output matrix with size  $(n \times J)$ , where  $J$  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 which are able to manage multiple predictors (inputs) and multiple predictands (outputs). If  $\mathbf{x}$  is the set of predictors and  $\mathbf{y}$  the predictands, then CCA can be used to predict  $\mathbf{y}$  when new observations of  $\mathbf{x}$  become available. The method finds linear combinations of the original variables,

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

and

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

by projecting them onto coefficient vectors  $a_m$  and  $b_m$ , which are 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. 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  $\mathbf{x}$  and  $\mathbf{y}$ . The canonical vectors  $a_m$  and  $b_m$  are the choices that result in the canonical variates having the following properties:

$$\text{Corr}(v_1, w_1) \geq \text{Corr}(v_2, w_2) \geq \dots \geq \text{Corr}(v_M, w_M) \geq 0 \quad (2a)$$

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

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

and

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

<sup>1</sup> The reference refers to the reprinting of the original paper in the book ‘Breakthroughs in Statistics. Springer Series in Statistics’ of 1992.

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:

$$[R_c] = \begin{bmatrix} r_{c_1} & 0 & 0 & \dots & 0 \\ 0 & r_{c_2} & 0 & \dots & 0 \\ 0 & 0 & r_{c_3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & r_{c_M} \end{bmatrix} \quad (3)$$

Equations 2b and 2c state that each canonical variate is uncorrelated with all the other canonical variates except its specific counterpart in the  $m^{\text{th}}$  pair; finally equation 2d states that each of the canonical variates has variance equal to 1.

The basic idea behind forecasting with CCA is straightforward: simple linear regressions are constructed that relate the predict and canonical variates  $w_m$  to the predictor canonical variates  $v_m$ :

$$\hat{w}_m = \hat{\beta}_{0,m} + \hat{\beta}_{1,m}v_m, \quad m = 1, \dots, M \quad (4)$$

where  $\hat{\beta}_{0,m}=0$ , because the CCA is calculated from the centered data  $\mathbf{x}'$  and  $\mathbf{y}'$ , and  $\hat{\beta}_{1,m}=r_{c_m}$ , because the canonical variates are scaled to have unit variance, so 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, specifically only papers regarding pyrolysis tests in fluidized beds without any catalyst have been considered (Chai et al., 2020; Christoforou et al., 2018; Garcia-Perez et al., 2010; Greenhalf et al., 2013; lisa et al., 2016; Jung et al., 2008; Ly et al., 2019, 2020; Mullen et al., 2018; Paasikallio et al., 2014; Williams et al., 2000; Zhang et al., 2009). Furthermore, only tests carried out in the temperature range of 450-550°C have been collected because representative of the optimal range in terms of bio-liquid<sup>2</sup> yield, so obtaining a dataset with 59 observations. The input data  $\mathbf{x}$  have been constituted by combination of ultimate and proximate analysis of the raw biomass, while the output data  $\mathbf{y}$  report bio liquid and bio char yields, and the weight H/C and O/C ratios in the bio liquid. Table 1 summarizes the input and output variables used for CCA.

Table 1: Variables of  $\mathbf{x}$  and  $\mathbf{y}$  used for CCA statistical analysis.

x data	Definition	y data	Definition
M/CM	Moisture to Combustible matter weight ratio of raw biomass on dry basis	w% Bio Liquid	Weight yield of bio liquid
A/CM	Ash to Combustible matter weight ratio of raw biomass on dry basis	w% Bio Char	Weight yield of bio char
H/C	Hydrogen to Carbon weight ratio of raw biomass on dry basis	(O/C) <sub>BL</sub>	Oxygen to Carbon weight ratio in the bio liquid
N/C	Nitrogen to Carbon weight ratio of raw biomass on dry basis	(H/C) <sub>BL</sub>	Hydrogen to Carbon weight ratio in the bio liquid
O/C	Oxygen to Carbon weight ratio of raw biomass on dry basis		

The CCA analysis has been implemented in MatLab™ environment by the utilization of 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, MAPE and  $R^2$ <sup>3</sup> have been calculated in the evaluation of the correlations between  $\mathbf{x}$  and  $\mathbf{y}$  but for simplification, only MAPE has been reported here.

### 3. Results

Table 2 reports the linear correlations and relative MAPE obtained by CCA from  $\mathbf{x}$  and  $\mathbf{y}$  data. In general, good prediction performance have been obtained for all outputs, witnessed by the MAPE values which are lower than 10%, except for (O/C)<sub>BL</sub> which is around 13%. Figure 1 shows a comparison between experimental and predicted values for all outputs considered. Interesting the results show the Bio-liquid production yield is highly dependent on the presence of ash (A/CM). In particular, the larger is the ash content, the lower amount bio-liquid and higher bio-char are produced.

<sup>2</sup> The bio liquid corresponds to the liquid organic phase + the pyrolytic water.

<sup>3</sup> MAE=mean absolute error; MSE=mean square error; SAE= sum absolute error; MAPE=mean absolute percentage error;  $R^2$ =coefficient of determination.

Table 2: x-y correlations from CCA Analysis

	Correlations	MAPE
$w\% \text{ Bio Liquid}$	$= 1.36 \times \frac{M}{CM} - 20.92 \times \frac{A}{CM} - 262.55 \times \frac{H}{C} - 46.74 \times \frac{N}{C} - 26.09 \times \frac{O}{C} + 118.63$	7.66 (5)
$w\% \text{ Bio Char}$	$= 81.30 \times \frac{M}{CM} + 21.35 \times \frac{A}{CM} + 199.60 \times \frac{H}{C} - 38.64 \times \frac{N}{C} + 12.25 \times \frac{O}{C} - 23.05$	9.11 (6)
$\left(\frac{O}{C}\right)_{BL}$	$= -2.02 \times \frac{M}{CM} + 0.60 \times \frac{A}{CM} - 4.77 \times \frac{H}{C} - 4.22 \times \frac{N}{C} - 0.08 \times \frac{O}{C} + 1.53$	13.27 (7)
$\left(\frac{H}{C}\right)_{BL}$	$= -0.27 \times \frac{M}{CM} + 0.07 \times \frac{A}{CM} - 0.42 \times \frac{H}{C} - 0.58 \times \frac{N}{C} + 0.04 \times \frac{O}{C} + 0.16$	7.27 (8)

This result can be explained considering: i) the ash, as a non-pyrolyzable fraction, remains in the solid state, consequently its higher content (typical of agricultural residues) increases the production of the solid phase to the detriment of the others; ii) the ash has catalytic properties which favour the cracking phenomena of the pyrolysis vapours, consequently producing a smaller amount of bio-liquid. Furthermore, even a higher carbon content in the raw biomass has a positive effect for the production of the bio-liquid, indeed lower ratios of H/C, N/C and O/C tend to increase the yield of bio-liquid, while the situation is almost the reverse for the solid phase.

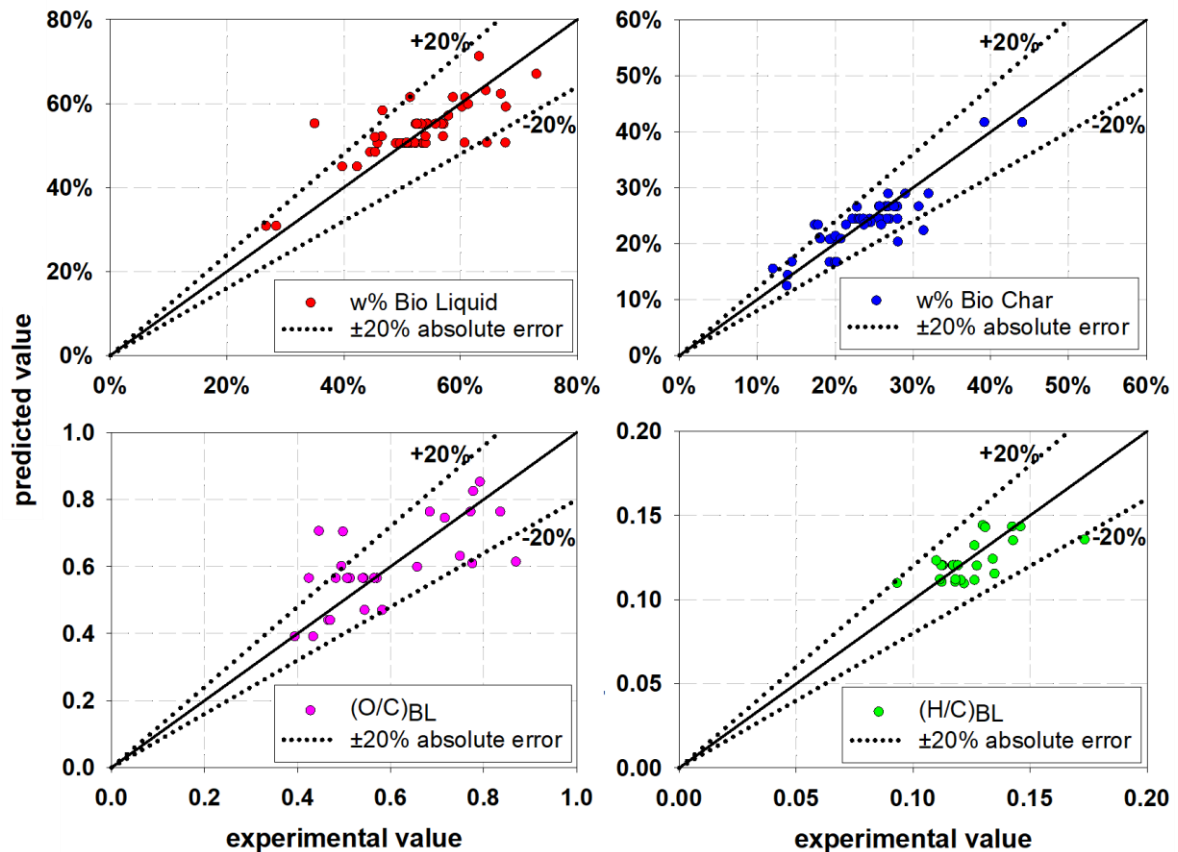


Figure 1: Comparison of the predicted value of the CCA model with experimental values.

Regarding the O/C ratio in the bio-liquid, a significant effect of the M/CM, H/C and N/C ratios can be noted, but the O/C of the raw biomass seems not to be very relevant. This result is perhaps imputable to the nature of the macro-components making up the biomass and their relative decomposition pathways and how humidity can play a role in these pathways. Conversely, H/C in bio-liquid appears to have a roughly equal dependence on all inputs.

#### 4. Conclusions

The present work investigated the applicability of statistical analysis for describing thermochemical processes with simple linear correlations. The CCA method represents the first attempt to create simple tools for evaluating the performance of different processes for the residual biomass valorisation. In particular, CCA was used to quantitatively and qualitatively describe the process outputs. The results show that the predicted values of bio-liquid and bio-char yield, and O/C and H/C ratios are within the 20% of error with respect to their experimental values. The main advantage of using the CCA is its simple implementation. However, the main limitation is that it only applies linear correlations between inputs and outputs; despite this, results can be considered acceptable considering the trade-off between prediction performance and computational time requirement.

Future perspectives are to find useful prediction models looking for the optimal compromise between prediction efficiency and computation speed, in order to develop an efficient decision-support system and a platform for comparative assessment of alternative pathways for the production of bio-based fuels and chemicals from raw biomass residues and their blends. Other techniques, such as machine learning, can be used for more complex prediction based on non-linear correlations. Furthermore, for all techniques based on datamining, their applicability strictly depends on the quality of dataset used. In this sense, techniques based on natural language process for text generation can help the compiling of large datasets.

#### Acknowledgments

This study has been carried out in the frame of the project PON ARS01\_00985: Biofeedstock: Development of Integrated Technological Platforms for Residual Biomass Exploitation, funded by the Italian Ministry for University and Research.

#### References

- Chai, M., He, Y., Nishu, Sun, C., & Liu, R. (2020). Effect of fractional condensers on characteristics, compounds distribution and phenols selection of bio-oil from pine sawdust fast pyrolysis. *Journal of the Energy Institute*, 93(2), 811–821. doi: 10.1016/j.joei.2019.05.001
- Christoforou, E. A., Fokaides, P. A., Banks, S. W., Nowakowski, D., Bridgwater, A. V., Stefanidis, S., Kalogiannis, K. G., Iliopoulou, E. F., & Lappas, A. A. (2018). Comparative Study on Catalytic and Non-Catalytic Pyrolysis of Olive Mill Solid Wastes. *Waste and Biomass Valorization*, 9(2), 301–313. doi: 10.1007/s12649-016-9809-5
- Colling Klein, B., Bonomi, A., & Maciel Filho, R. (2018). Integration of microalgae production with industrial biofuel facilities: A critical review. *Renewable and Sustainable Energy Reviews*, 82, 1376–1392. doi: 10.1016/j.rser.2017.04.063
- Council, O. F. T. H. E. (2018). *DIRECTIVE (EU) 2018/2001 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 December 2018 on the promotion of the use of energy from renewable sources (recast)*. 2018(April 2009).
- European Union. (2009). *DIRECTIVE 2009/28/EC of the European Parliament and of the Council*. *Official Journal of the European Union*, 5(1), L 140/16-L 140/62.
- Garcia-Perez, M., Shen, J., Wang, X. S., & Li, C. Z. (2010). Production and fuel properties of fast pyrolysis oil/bio-diesel blends. *Fuel Processing Technology*, 91(3), 296–305. doi: 10.1016/j.fuproc.2009.10.012
- Greenhalf, C. E., Nowakowski, D. J., Harms, A. B., Titiloye, J. O., & Bridgwater, A. V. (2013). A comparative study of straw, perennial grasses and hardwoods in terms of fast pyrolysis products. *Fuel*, 108, 216–230. doi: 10.1016/j.fuel.2013.01.075
- Hirani, A. H., Javed, N., Asif, M., Basu, S. K., & Kumar, A. (2018). A review on first-and second-generation biofuel productions. In *Biofuels: Greenhouse Gas Mitigation and Global Warming* (pp. 141–154). Springer.
- Hotelling, H. (1992). *Relations Between Two Sets of Variates BT - Breakthroughs in Statistics: Methodology and Distribution* (S. Kotz & N. L. Johnson (eds.); pp. 162–190). New York, NY: Springer New York. doi: 10.1007/978-1-4612-4380-9\_14
- IEA (2019), *World Energy Outlook 2019*, IEA, Paris <https://www.iea.org/reports/world-energy-outlook-2019>. (2019).
- lisa, K., French, R. J., Orton, K. A., Yung, M. M., Johnson, D. K., Ten Dam, J., Watson, M. J., & Nimlos, M. R. (2016). In Situ and ex Situ Catalytic Pyrolysis of Pine in a Bench-Scale Fluidized Bed Reactor System. *Energy and Fuels*, 30(3), 2144–2157. doi: 10.1021/acs.energyfuels.5b02165
- Jung, S., Kang, B., & Kim, J. (2008). *Production of bio-oil from rice straw and bamboo sawdust under various reaction conditions in a fast pyrolysis plant equipped with a fluidized bed and a char separation system*. 82, 240–247. doi: 10.1016/j.jaap.2008.04.001

- Lamers, P., Roni, M. S., Tumuluru, J. S., Jacobson, J. J., Cafferty, K. G., Hansen, J. K., Kenney, K., Teymouri, F., & Bals, B. (2015). Techno-economic analysis of decentralized biomass processing depots. *Bioresource Technology*, *194*, 205–213. doi: 10.1016/j.biortech.2015.07.009
- Littlejohns, J., Rehmann, L., Murdy, R., Oo, A., & Neill, S. (2018). Current state and future prospects for liquid biofuels in Canada. *Biofuel Research Journal*, *5*(1), 759–779.
- Ly, H. V., Choi, J. H., Woo, H. C., Kim, S. S., & Kim, J. (2019). Upgrading bio-oil by catalytic fast pyrolysis of acid-washed *Saccharina japonica* alga in a fluidized-bed reactor. *Renewable Energy*, *133*, 11–22. doi: 10.1016/j.renene.2018.09.103
- Ly, H. V., Park, J. W., Kim, S. S., Hwang, H. T., Kim, J., & Woo, H. C. (2020). Catalytic pyrolysis of bamboo in a bubbling fluidized-bed reactor with two different catalysts: HZSM-5 and red mud for upgrading bio-oil. *Renewable Energy*, *149*, 1434–1445. doi: 10.1016/j.renene.2019.10.141
- Mullen, C. A., Tarves, P. C., Raymundo, L. M., Schultz, E. L., Boateng, A. A., & Trierweiler, J. O. (2018). Fluidized Bed Catalytic Pyrolysis of Eucalyptus over HZSM-5: Effect of Acid Density and Gallium Modification on Catalyst Deactivation. *Energy and Fuels*, *32*(2), 1771–1778. doi: 10.1021/acs.energyfuels.7b02786
- Paasikallio, V., Lindfors, C., Kuoppala, E., Solantausta, Y., Oasmaa, A., Lehto, J., & Lehtonen, J. (2014). Product quality and catalyst deactivation in a four day catalytic fast pyrolysis production run. *Green Chemistry*, *16*(7), 3549–3559. doi: 10.1039/c4gc00571f
- Razik, A. H. A., Khor, C. S., & Elkamel, A. (2019). A model-based approach for biomass-to-bioproducts supply Chain network planning optimization. *Food and Bioproducts Processing*, *118*, 293–305.
- Williams, P. T., & Nugranad, N. (2000). Comparison of products from the pyrolysis and catalytic pyrolysis of rice husks. *Energy*, *25*(6), 493–513. doi: 10.1016/S0360-5442(00)00009-8
- Zhang, H., Xiao, R., Huang, H., & Xiao, G. (2009). Comparison of non-catalytic and catalytic fast pyrolysis of corncob in a fluidized bed reactor. *Bioresource Technology*, *100*(3), 1428–1434. doi: 10.1016/j.biortech.2008.08.031
- Ziolkowska, J. R. (2020). Biofuels technologies: An overview of feedstocks, processes, and technologies. In *Biofuels for a More Sustainable Future* (pp. 1–19). Elsevier.