Technical University of Denmark

The importance of the relationship between lignin and carbohydrate contents when predicting biomethane potentials from lignocellulosic biomass

Thomsen, Sune Tjalfe; Kroff, Pablo; Østergård, Hanne

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

The importance of the relationship between lignin and carbohydrate contents when predicting biomethane potentials from lignocellulosic biomass

13th World Congress on Anaerobic Digestion. Recovering (bio) Resources for the World.

Sune Tjalfe Thomsen*[1], Pablo Kroff[1] and Hanne Østergård[1]

[1]Center for Bioprocess Engineering, Dept. of Chemical and Biochemical Engineering, Technical University of Denmark, DTU Risø Campus, DK-4000 Roskilde

Introduction

Why?!

- We are aiming to find the best way of finding predicted BioMethane Potential (pBMP) from compositional data of a lignocellulosic biomass – without making time-consuming BMP tests.
- We are questioning the use of standard linear regression, or multiple linear regression, of the type $pBMP = \alpha_0 + \sum_i \beta_i x_i$ such as proposed previously [1, 2], these models often only find lignin to be significant to pBMP.
- The previous models described the data adequately, but were not intuitive and did not go well with the theoretic background of anaerobic digestion (e.g. could not show significance of carbohydrates).

What's new?!

- We introduce a new way of statistically describing the relationship of biomass constituents and pBMP. **– The Mixture model –**this was previously used when describing effects of chemical mixtures. In that way we can show the true importance of each of the biomass constituents.
- In mixture models, the variables are proportionate nonnegative amounts of different constituents $0 \le x_i \le 1, i = 1, 2, \ldots, q$. In this case the variables are biomass constituents: Cellulose (X_G), hemicellulose (X_H), carbohydrates ($X_C = X_G + X_H$), lignin (X_L)
- In mixture models the variables sum up to one, $\sum_{i=1}^{q} x_i = 1$. Therefor what is known as residuals when looking at biomass composition is included in the model – in this case everything which is not carbohydrates or lignin is characterized as residuals, $X_R = 1 - (X_G + X_H + X_L)$.
- We test the linear model for mixture designs $Y = \sum_{i=1}^{p} \beta_i x_i$ as well as the quadratic model for mixture designs $\boldsymbol{Y = \sum_{i=1}^{p} \beta_i x_i + \sum\sum_{i<j}^{p} \beta_{ij} x_i x_j}$ as developed by Henry Scheffé and peers from 1953 and onwards [4,5]

Results

How was the data collected?!

- We assessed data from peer reviewed literature, with the prerequisites that the data should include BMP as well content of cellulose, hemicellulose and lignin, and that the biomass in question should be lignocellulosic.
- We collected 49 data sets from literature from a broad spectrum of different lignocellulosic biomasses.
- Furthermore, we prepared 16 data sets for this study, summing to a total of 65 data sets.
- Data and references are not shown – but will be displayed in following paper submission.

Comparison of models:

- Models are compared on the basis of: significance of regression coefficients, R^2-values, P-values, and relative Root Mean Square Error = $rRMSE = \left(\sqrt{\sum_{t=1}^{n}(BMP - pBMP)^2/n}\right) \cdot 100\% \,/\, mean(BMP)$, see table below.
- Notice that the rRMSE – which shows how far the pBMP's are from the measured BMP's– is fairly similar for the different models.
- Also notice, that the R^2 - a measure of how well observed outcomes are predicted by the model – is much higher in the mixture designs presented in this study.

Equation	Source Presented first by:	Applied to data Estimate (St. Dev.)	Significance	R^2	p	rRMSE
pBMP ~ $\beta_G X_G + \beta_H X_H + \beta_L X_L + \beta_R X_R$	This study	β_G = **378** (62) β_H = **354** (94) β_L = **-194**(69) β_R = **313** (45)	*** *** ** ***	**0.9637**	< 2.2e-16	19.72%
pBMP ~ $\beta_C X_C + \beta_L X_L + \beta_R X_R$	This study	β_C = **369** (27) β_L = **-194**(68) β_R = **312** (45)	*** ** ***	**0.9637**	< 2.2e-16	19.73%
pBMP ~ $\alpha_0 + \beta_L X_L$	[1, 2]	β_L = **-545**(80) α_0 = **348** (14)	*** ***	**0.4265**	= 5.0 e-09	19.85%
pBMP ~ $\alpha_0 + \beta_L X_L + \beta_G X_G$	[1]	β_G = **67** (84) β_L = **-517**(87) α_0 = **322** (35)	n. s. *** ***	**0.766**	= 3.2e-08	22.02%
Buswell on carbohydrates: pBMP ~ $\beta_G X_G + \beta_H X_H$	[3]	β_G = **414** β_H = **423**		-	-	29.60%

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 'n. s.' 1

Conclusion

1. Mixture models?!

Using the statistical theory of mixture models instead of standard linear regression provides models with R^2 values over 96%, where the regression coefficients for the different biomass constituents are all significant and where operational signs are intuitive according to AD theory.

2. The best equation to date?!

Based on a large data set from literature and own unpublished data the following equation to predict BMP was developed:

pBMP = 378 · [cellulose] + 354 · [hemicellulose] – 194 · [lignin] + 313 · [residuals]

Since this model is based on a large data set, we claim that this is the best model for predicting pBMP from lignocellulosic biomass to date!

3. Are the interactions between biomass constituents important?

Sine the blending terms quadratic model for mixture designs were not significant we are not able to prove that the interaction between biomass constituents are as important as the sheer amounts.

4. What are the perspectives?!

This way of describing the relationship between biomass composition and BMP can be used to make more precise models of more specific biomasses

Discussion

Why are the previous models both wrong and correct?!

- The inherent mathematical connection between the mixture models and the models with intercept is described by the following calculations:

$$\mathbf{pBMP \sim \beta_C X_C + \beta_L X_L + \beta_R X_R} = \beta_C X_C + \beta_L X_L + \beta_R (1 - X_C - X_L) = (\beta_C - \beta_R) X_C + (\beta_L - \beta_R) X_L + \beta_R = \mathbf{\beta_{C*} X_C + \beta_{L*} X_L + \alpha_0}, \text{ where } \beta_{C*} = \beta_C - \beta_R,\ \beta_{L*} = \beta_L - \beta_R,\ \beta_R = \alpha_0$$

- The problem by the latter equation is that since β_C and β_R are both positive and in the same order of magnitude then $\beta_{C*} = \beta_C - \beta_R$ will be small, **thus it will seem insignificant**. It might even be negative even though that is unintuitive and contrary to AD theory.
- Likewise, $\beta_{L*} = \beta_L - \beta_R$ will be a large negative variable with high significance since β_L already is a negative regressor – **Thus, the role of lignin have been exaggerated!**
- These problems are linked to a concept known as multicollinearity – an effect which is known to interfere with the significance of variables.

What is the consequence of not summing to 1

- The relative nature of all compositional data which have units of w/w% or g/100 g TS, are not taken into account.
- The effects of methane yielding carbohydrates cannot be proved significant.
- Residuals, including methane yielding biomass constituent like lipids and proteins, are not taking into account.
- Instead the effects of carbohydrates and residuals are hidden in the intercept (a_0)
- Since the variables (biomass constituents) have the same order of magnitude, and since they have an underlying connection, multicollinearity is a problem.

Why is our model more intuitive?!

- Since the biomass constituents that in theory should be energy yielding are positive (e.g. cellulose), while lignin which is known to have a inhibitory shielding effect on BMP is negative.
- Since the pBMP is not only determined by lignin content.
- Regression coefficients for contents of cellulose, hemicellulose, lignin, and residuals are all highly significant.

Can we prove that the interactions between lignin and carbohydrates are important?!

- When testing the quadratic model for mixture designs $pBMP = \sum_{i=1}^{p} \beta_i x_i + \sum\sum_{i<j}^{p} \beta_{ij} x_i x_j$ non of the blending terms were significant.
- Thus a linear mixture model is preferred, and we cannot prove that the interactions are important.
- However, a quadratic model for mixture designs might be more relevant on a more homogeneous data set, E.g. with only one lignocellulosic biomass type.

Acknowledgement

The authors would like to acknowledge statistical support from Professor Henrik Spliid, DTU. This work was supported with by Danida Fellowship Centre of the Danish Ministry of Foreign Affairs, as a part of the project "2GBIONRG", DFC journal no. 10-018RISØ
If interested visit http://www.2gbionrg.dk/

*sunt@kt.dtu.dk

Selected references

[1]: 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, Bioresour. Technol. 102, 9395-9402.
[2]: Monlau, F., Sambusiti, C., Barakat, A., Guo, X.M., Latrille, E., Trably, E., Steyer, J.-., Carrere, H., 2012b. Predictive models of biohydrogen and biomethane production based on the compositional and structural features of lignocellulosic materials, Environmental Science and Technology. 46, 12217-12225.
[3]: Symons, G.E., Buswell, A.M., 1933. The methane fermentation of carbohydrates, J. Am. Chem. Soc. 55, 2028-2036.
[4]: Cornell, J.A., 2011. A retrospective view of mixture experiments, Quality Engineering. 23, 315-331.
[5]: Scheffe, H., 1963. The Simplex-Centroid Design for Experiments with Mixtures, Journal of the Royal Statistical Society.Series B (Methodological). 25, 235-263.