Modeling of tensile index using uncertain data sets

Citation for the original published paper (version of record):
Bengtsson, F., Karlström, A., Wik, T. (2020)
Modeling of tensile index using uncertain data sets
Nordic Pulp and Paper Research Journal, 35(2): 231-242
http://dx.doi.org/10.1515/npprj-2019-0089

Paper physics

Fredrik Bengtsson*, Anders Karlström and Torsten Wik

Modeling of tensile index using uncertain data sets

https://doi.org/10.1515/npprj-2019-0089
Received October 22, 2019; accepted March 4, 2020; previously published online April 10, 2020

Abstract: The objective of this investigation is to analyze and model tensile index. Two approaches are used, one based on training and validation data, while the other novel approach tests models using all possible combinations of data points. This approach is focused on small data sets which have here been obtained from nineteen pulp samples at different refining conditions in a full-scale TMP production line with a CD-76 refiner as a primary stage. From each pulp sample twenty handsheet strips for tensile index measurements were performed. Initially, specific energy and the external variables (dilution water feed rates and plate gaps) are used as predictors in a modeling approach based on an adjusted R^2 approach. Thereafter, the resulting models are compared with a combination of specific energy and internal variables (primarily consistencies) obtained from temperature measurements inside the refining zones using a soft sensor concept. It is found that specific energy and internal variables as predictors outperform the external variables when estimating tensile index.

Keywords: linear regression; modeling; tensile index; TMP; uncertain data sets.

Introduction

The laboratory test procedures in pulp and paper industry have been discussed for decades. A consequence of tedious and complex measurements of pulp and handsheet properties often results in very few samples, which makes it difficult to verify data sets statistically. Many robust techniques, such as modified Z-score, box plot, sample kurtosis and the Shapiro-Wilk W test (Barnett and Lewis 1994) can be natural tools when improving measurement quality in laboratory data. However, it can also be crucial to link the variations in the mechanical pulping process variables to the composition of particle shapes and sizes in the pulp (Forgacs 1963), which opens for a number of possibilities when selecting predictors for pulp and handsheet property models.

Hence, there are at least three challenges that must be considered simultaneously when deriving models in this area:

1. Pulp measurement accuracy.
2. Interpreting and processing laboratory data and process information from plant control systems, on-line pulp sampling devices etc. Detecting anomalous process conditions which can give rise to outliers.
3. Selection of predictors for handsheet property modeling including validation of considered models.

To cope with the first challenge a modified detection algorithm based on a generalized Extreme Studentized Deviate procedure can be used (Rosner 1983). This method was primarily used for environmental pollution monitoring to avoid the problem of masking (Gilbert 1987). This methodology was applied to pulp and paper applications by Karlström et al. (2019), where its was utilized to detect discordant outliers and anomalies in laboratory samples. The method, however, is not suitable when the data sets for each pulp sample are too small, being best suited for data sets larger than 40 samples.

For the second challenge there are two types of outliers needed to be considered. Firstly, on each pulp sample numerous measurements are made, some of which may be outliers. Secondly, the sample itself may be an outlier, due to anomalous process conditions or flawed measuring procedure during the time the sample was taken.

Regarding the third challenge, to determine the number of predictors, one can use a general guideline for the minimum number of events per variable (EPV) in multivariate analysis. Harrell et al. (1985) and Freedman and Pee (1989) demonstrated that overfitting was inflated when the ratio of the number of variables to the number of

***Corresponding author: Fredrik Bengtsson,** Department of Electrical Engineering, Chalmers University of Technology, SE 412 96 Göteborg, Sweden, e-mail: fredben@chalmers.se, ORCID: https://orcid.org/0000-0002-6670-7493
Anders Karlström, Torsten Wik, Department of Electrical Engineering, Chalmers University of Technology, SE 412 96 Göteborg, Sweden, e-mails: anderska@chalmers.se, torsten.wik@chalmers.se

observations was greater than 1/4, which corresponds to an EPV $\geq$ 4. Peduzzi et al. (1996) suggested an increase of that number to at least ten events per variable analyzed to maintain the validity of the final model. Draper and Smith (1998) also suggest the use of an EPV of 10, but in industrial applications, this is usually not possible due to tedious laboratory analysis and uncertainties in the measurements, which limits the number of reliable samples (Karlström and Hill 2017a,b,c). The modeling approach in mechanical pulping processes, has been that external variables, such as specific energy (i. e. the ratio between motor load and production), dilution water added to the refiners, plate gaps (disc clearance) etc., should be used for process follow up of pulp and handsheet properties (Strand 1996, Härkönen et al. 2000, Sabourin et al. 2001, Härkönen et al. 2003, Strand and Grace 2014, Nelsson 2016).

However, when using external variables as predictors, the process non-linearities tend to negatively affect the result. To cope with that soft sensors, describing physical phenomena in the refining zone, have been developed during the last decade (Karlström and Eriksson 2014a,b,c,d). The soft sensor's outputs can be seen as estimates of internal variables (such as fiber residence time, consistency profile, forces on bars, distributed defibration, thermodynamic work etc.) which are difficult to measure directly in the process. Typically, such soft sensors are non-linear but have become important for advanced process optimization. Specifically, consistency and fiber residence time have been candidates for such activities for some years, as they provide a link to e. g. tensile index, mean fiber length and Somerville shives (Karlström et al. 2015, 2016a,b, Karlström and Hill 2017a,b,c).

In earlier articles Karlström et al. (2015, 2016a,b) have shown that the use of internal variables as predictors outperform the use of external variables when making polynomial fits of pulp and handsheet properties. The use of consistency and fiber residence time were in focus in those articles. Furthermore, it was shown that adding specific energy, i. e. the ratio between motor load and production as a predictor did not improve the modeling of tensile index (Karlström and Hill 2017a,b,c) when the excitation in the production rate is small. In this study however, the changes in the production rate is much larger then compared with the previous cases.

When specific energy is used as one of many predictors, multicollinearities are often introduced. Karlström and Eriksson (2014a,b,c,d) showed that nonlinearities in the refining zones are considerable and should be treated with care as both consistencies and fiber residence times correlate to some extent with the specific energy but not as strongly as with the external variables dilution water feed rates and plate gap measurements.

In our case we have a small data set, obtained from nineteen pulp samples at different refining conditions in a CD-76 refiner. To resolve the above mentioned challenges we start by designing models using linear regression. These models are then evaluated using two methods; using training and validation data sets and by testing all possible subsets of ten samples. From this evaluation we locate outliers and determine which predictors are best to use.

The main idea in this article, is to derive a linear model for predicting tensile index using available measured (external) and estimated (internal) variables. From this we will be able to determine which variables seem to be the most useful when modelling tensile index. Furthermore we will present methods for designing models when data sets are very small and consist of relatively unreliable data from laboratory experiments. Ultimately, a model for tensile index using specific energy and the consistency is found and presented.

Materials and methods

In this section, the control strategy and pulp sampling procedure is first presented, then considerations related to measurement accuracy in pulp samples are discussed, followed by an overview about the needs to link laboratory data and process information in time-domain. Finally, we give a short introduction to the basic approach to select pulp and handsheet property candidates for process modeling purposes.

Conditions during the trial

The measurements were taken during a five day period, during which the process was controlled using a similar strategy as described in Karlström and Hill (2017a,b,c). That is to say the goal was to increase the production by increasing the amount of work done in the flat zone. To do this the dilution water feed rate in the flat zone was automatically controlled to regulate output consistency, while the dilution water feed rate of the conical zone was manually adjusted to try to keep the values of the consistencies in the two zones close to each other. The plate gaps were then adjusted to increase the work done in the flat zone.

At each sampling interval pulp was collected during a two minute period, to be used to make handsheets for analysis.

Measurement accuracy in pulp samples

When conducting measurements on pulp samples, numerous measurements are taken at a single sampling time to compensate for low measurement accuracy. We will assume that the dynamic variations in the process, during each pulp sampling, can be considered small and that the obtained average of each pulp sample is representative for the process conditions during each sampling interval. By preparing handsheets from each pulp sample this means that possible outliers are most likely related to the handsheets and not to the variations in the process. As an example, consider Figure 1 where three strips are provided for analysis. It is natural that each strip can contain none, one or more shives, which certainly affect the strength (Bajpai 2012). Moreover, the basis weight of the handsheets can vary even though they are coming from the same pulp sample.

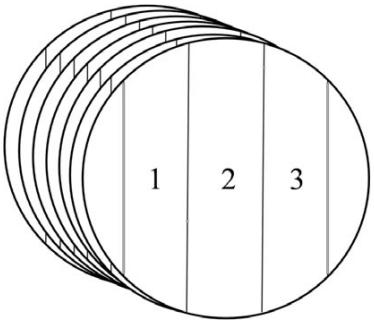


Figure 1: A schematic drawing of three strips obtained from each of the seven handsheets.

In short, mainly three different approaches can be used to calculate the tensile index (τ), namely

$$\tau_{ij} = \begin{cases} \sigma_{ij}/\bar{\mu} & \text{Case A} \\ \frac{\sigma_{ij}}{\frac{1}{l}\sum_{i=1}^{l}\mu_{ij}} & \text{Case B} \\ \sigma_{ij}/\mu_{ij} & \text{Case C} \end{cases} \quad (1)$$

where σ_{ij} is the tensile strength of strip i from handsheet j, μ_{ij} is the corresponding basis weight and l is the number of strips. The denominator in Case A is the average basis weight for all handsheets, i. e. one measure for the complete batch of handsheets. The denominator in case B can be seen as the most logical average basis weight to use for each handsheet. Case C requires information about both tensile strength and basis weight for each strip. In general, Case C is time consuming and not so often used as a standard procedure. Instead, Case A is normally used, which of course can affect the accuracy when it comes to handsheet variations in the forming procedure.

To illustrate the need to treat the tensile index measurements with some care we study the tensile index variation in the set of data used in this paper. The full raw data is provided in Bengtsson et al. (2019).

As can be seen, 19 pulp samples are studied and for each pulp sample 20 strips are used for tensile measurements. The variation in these tensile indices is quite large. This is probably due to variation in the number of shives in the different strips. To compensate for this we generate a set by only utilizing data within two standard deviations from the mean, disregarding the other data points as outliers.

This removes some of the points as outliers, as shown in Figure 2.

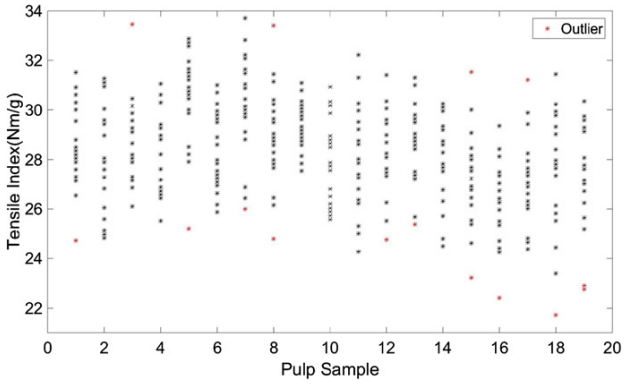


Figure 2: Tensile measurements, showing which measurements will be disregarded (red) as they deviate too far from the mean.

To get a single value of tensile index for each sample, we then took the mean of the remaining measurements.

Laboratory data and process information linked in time domain

The 19 samples analyzed above are taken from a full-scale TMP production line where the primary stage is a CD76-refiner. This type of refiner consists of two serial refining zones, called flat zone (FZ) and conical zone (CD). In both zones, radial sensor arrays with eight sensors have been mounted to measure the entire temperature profiles, see Figure 3. In total, 16 temperatures are measured, which makes it possible to follow the process from a completely different perspective compared to traditional sensor installations.

The temperature measurements can be seen as internal variables that are measured together with traditional process variables, such as production rate, dilution water flows, plate gaps and motor load (external variables). The temperature profile varies considerably when changes are made in process conditions and in the refining segment pattern (Karlström and Hill 2017a,b,c).

Figure 3: Two sensor arrays for temperature measurements mounted on the stator side (FZ-to the right and CD-to the left) in a CD76-TMP refiner. The chips are introduced in the central part of FZ.

Both the internal and external variables are used in the extended entropy model (Karlström and Eriksson 2014a,b,c,d), which can be used for estimation of e. g. the consistency profile and the fiber residence times in the FZ and CD zones (Karlström and Hill 2017a,b,c).

Production was calculated from input chips per unit time, without taking their density variation into account. This limitation of course also affects specific energy as such.

The test in this study was performed during five days where all process variables are segmented around the periods for pulp sampling according to Figure 4. The time of each test point was well-documented (Karlström et al. 2018) and the test program was focused on the 19 test points indicated in Figure 4. The pulp samples were taken from the blow-line valve over a period of 3 minutes each.

The process sampling rate was 1 second and each pulp sampling was performed during a 3 minutes period. Therefore, average process conditions based on 180 samples were used in the analysis. Thereafter, the set of process data was then treated in the same way as the tensile index data; i. e. outliers were removed, and a single measurement was derived for each test point by taking the mean of the remaining data points. The full unmodified data is presented in Bengtsson et al. (2019).

Even though a reduced set of tensile index measurements are used it is hard to see any correlation between the process variables shown in Figure 4 and the tensile index measurements in Figure 2 only by visual inspection.

Predictors for handsheet property modeling

To estimate tensile index we will be using linear models in the form of

$$\hat{f}(x) = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k + b, \tag{2}$$

where $[\theta_1, \ldots, \theta_k]$ represents the parameter vector, $[x_1, \ldots, x_k]$ the predictor values and k the number of predictors.

The challenge in this paper is to find out which predictors are most suitable to use when modeling tensile index for the pulp samples shown in Figure 2.

The coefficient of determination (R^2) is useful for comparing different models. In our case we use the adjusted R^2 to set a penalty for the number of predictors (independent variables) in the model, according to

$$adj.R^2 = 1 - \frac{n-1}{n-d-1}\frac{\sum(f-\hat{f})^2}{\sum(f-\bar{f})^2}, \tag{3}$$

where $\sum(f-\hat{f})^2$ represents the sum of the squared residuals from the regression and $\sum(f-\bar{f})^2$ the sum of the squared differences from the mean of the dependent variable, while n is the number of observations and d is the number of predictors (Draper and Smith 1998).

When designing models with multiple predictors there is always the risk of multicollinearities, that is to say a linear interdependancy between the different predictors. This can be due to both poorly designed experiments, where various predictors are not exited separately, or due to a true physical link between the predictors as is often found in serially linked processes. To analyze and detect multicollinearities the Variance Inflation Factors (VIF) will be used in this paper (Belsley et al. 1980). VIF quantifies how much the variance is inflated, that is

$$VIF_k = \frac{1}{1-R_k^2}, \tag{4}$$

where R_k^2 is the R^2 value obtained by regressing the kth predictor on the remaining predictors. A $VIF_k = 1$ means that there is no linear correlation between the kth predictor and the other remaining predictor variables, i. e. the variances in the estimated coefficients in Equation (2) are not inflated. If $VIF > 4$, a general rule is that further analysis should be performed, while $VIF > 10$ indicates serious multicollinearities and a need to find a modified set of predictors. However, in serially linked processes like the one studied in this article, some collinearities occur naturally and further analysis should be considered from a more holistic perspective.

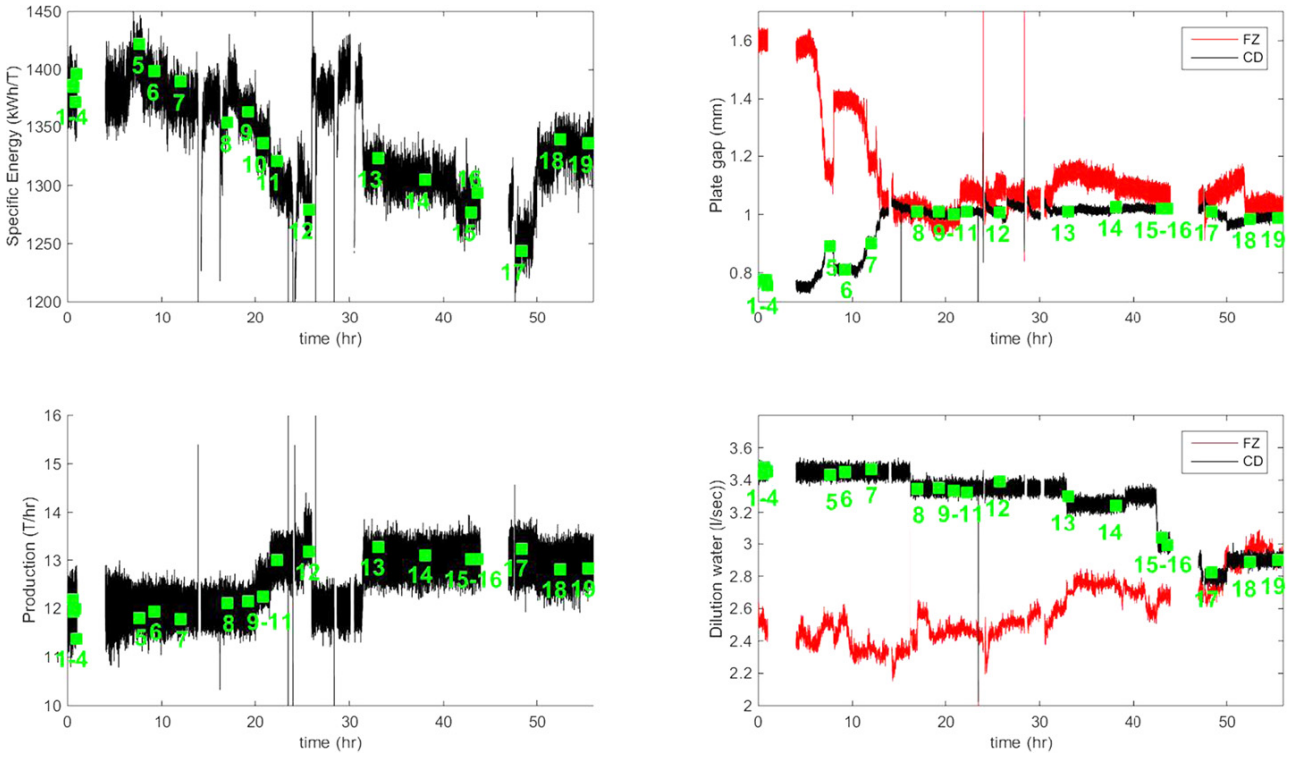


Figure 4: Specific energy, dilution water feed rates, plate gaps and production during the 5 days of testing. The numbers included (1–19) refers to the pulp sampling periods.

Validation of the models considered

Here, we propose a method for analysis and validation of models that is specifically tailored to cases with small data sets. Use roughly half the samples to design a model (e. g. 10 samples for a set of 19). Evaluate the resulting model by calculating the adjusted R^2. We then repeat this for all possible combinations of samples. This allows us not only to get a general impression of model quality but also evaluate the quality of each measurement sample. For example, if a sample only appears in poor models, it may be indicative that the sample itself is an outlier.

For small test series including less than 20 pulp samples, it is viable to test all combinations. The possible combinations can be interpreted as the number of ways there are to pick r unordered outcomes from n possibilities, i. e. n choose r:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}. \tag{5}$$

To reduce the number of possible combinations one may use a vector v, which consists of measurements at different operating conditions, if these measurements are reliable. This reduces the amount of combination while still ensuring that all operating conditions are considered. Doing this will reduce the number of combinations to

$$\binom{n-n_v}{k} = \frac{(n-n_v)!}{k!(n-n_v-k)!}, \tag{6}$$

where n_v is the number of elements in v and k is the number of samples used beyond those in v.

Once models have been designed the result from this test can be used to analyze the model structure.

Results and discussion

In this paper, the internal and external variables as predictors are in focus. The internal variables are estimated from process measurements and consist of the consistencies and fiber residence times in the refining zones, i. e. the FZ- and the CD-zones. The external variables are dilution water feed rates and plate gaps which are measured directly. Here we will use production or specific energy as predictors in conjunction with either internal or external variables. Firstly, we examine the presence of collinearities using the VIF (see Table 1).

From Table 1 both consistencies and dilution water feed rates seem viable for designing models, with at most

Table 1: Variance inflated factors for different combinations of potential predictors (each row is one combination, blank spaces denote that the predictor was not included in the test). Internal variables C=Consistency, RT=Fiber residence time, and external variables such as Sp.E=Specific energy, Dil.W=Dilution water, Gap=Distance between refining segments and Prod=Production.

Variance inflation factors (VIF)					
Potential predictors					
Sp.E	Prod.	Dil.WFZ	Dil.WCD	GapFZ	GapCD
8.5		5.2	10.2	34.3	44.1
2.3		3.9	6.2		
3.3				20.8	28.7
	4.6	3.4	4.5	19.4	26.8
	1.8	3.2	4.1		
	3.2			17.4	24.7
Potential predictors					
Sp.E	Prod.	C.FZ	C.CD	RT.FZ	RT.CD
5.6		2.6	9.1	9	4.3
2.5		1.8	2.9		
4.6				3.2	2.4
	194.1	2.8	8	75.4	45.1
	6.3	1.8	7.4		
	150.7			64.8	32.7

moderate collinearities. Moreover, from a physical perspective it is sound to include the dilution water feed rates and consistencies as predictors. The reason is that they are related to changes in fiber distribution in the refining zones (Fernando et al. 2013), which certainly affect the final pulp and handsheet properties even if the specific energy is not changed.

Residence times are strongly correlated with production but less so for specific energy, where the VIF is reasonable. This is somewhat surprising as we would have expected a higher correlation between specific energy and residence time.

Consistencies, residence times and dilution water feed rate seem to have comparable VIF, regardless if they are used as predictors together with specific energy or production, with the aforementioned exception of residence time and production which has a very poor VIF. We will henceforth use specific energy and not production as a predictor, in part due to the fact that the VIF was high for production and fiber residence times, and in part as further testing gave that using specific energy yielded better results also when using consistencies and dilution water feed rates as predictors.

Hence, following the motivation above four cases of predictor combinations are considered:

(a) specific energy;
(b) specific energy and consistencies;
(c) specific energy and fiber residence times;
(d) specific energy and dilution waters.

When analyzing the linear fits of tensile index measurements. It can be argued that Table 1 suggests that specific energy, consistencies and fiber residence time is a viable alternative. However with 5 predictors and only 19 samples, this would give us less then 4 samples per predictor, which is a limit Harrell et al. (1985) and Freedman and Pee (1989) suggested to avoid overfitting.

To assess the presence of outliers and to analyze which predictors are best to use, two different methods will be used. In the first, all combinations of the 10 different pulp samples will be examined to see which combinations yield a high adjusted R^2 as described in Section . In the other method, the data set will be divided into training and validation data to compare different combinations of predictors.

Using all combinations of 10 different pulp samples

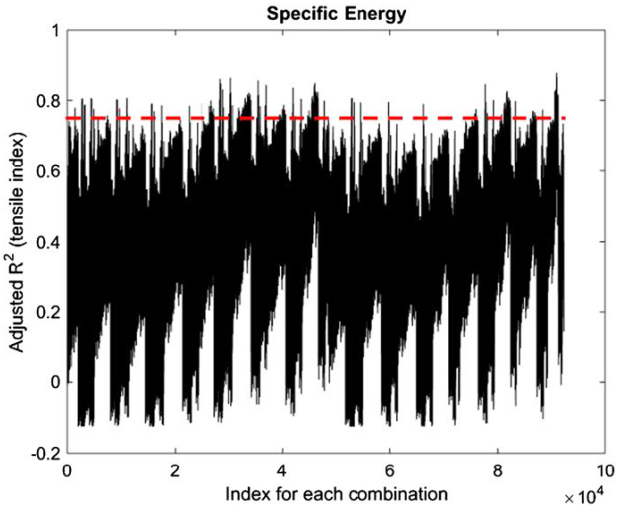
When examining all possible combinations of ten samples from the set of 19 samples we have 92378 combinations to study. We start with case (a), i. e. using specific energy as a sole predictor (see Figure 5).

To get an indication on how important a pulp sample is for modeling, a counter r of how many times each pulp sample occurs in a combination that results in an adjusted $R^2 > 0.75$ is introduced. This counter is represented by the vector r, where the ith element in r denotes how many times sample i occurred in a combination having an adjusted $R^2 > 0.75$.

If we also use consistencies together with specific energy the adjusted R^2 is increased considerably as is shown in Figure 6. Even with a more stringent criterion of adjusted $R^2 > 0.9$ we have more than 12000 combinations fulfilling the criterion. This indicates that considerable improvements to the model are made by utilizing the consistencies. Note that some samples occur much less often in good models then others. This is something we will look into further when analyzing outliers.

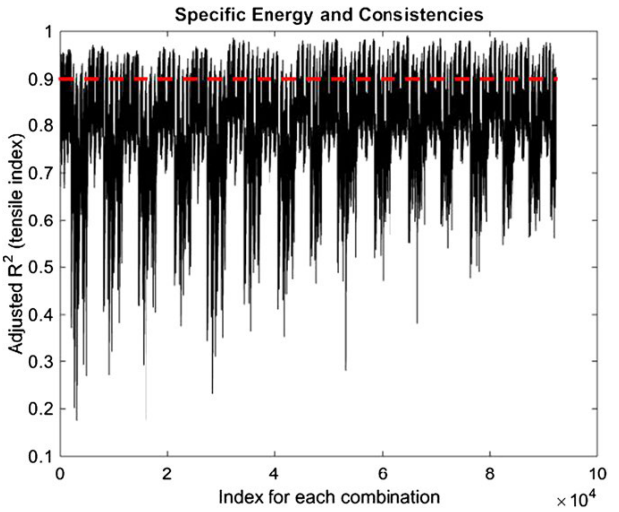
If we instead use fiber residence times together with consistencies, i. e. Case (c) (see Figure 7) we see that these do not yield nearly as good results with only around 480 combinations with an adjusted R^2 over 0.9.

Finally, we test specific energy together with the dilution water feed rate to each refining zone, i. e. Case (d). As can be seen in Figure 8, the number of combinations fulfilling an adjusted $R^2 > 0.9$ is quite low.



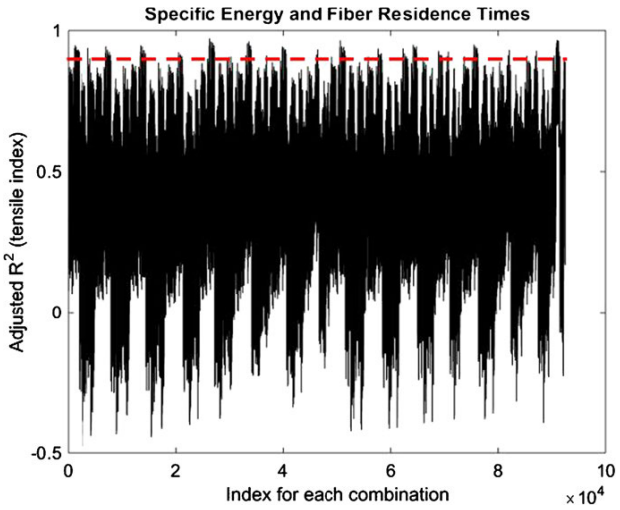
n = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]
r = [679 92 568 243 823 173 440 624 582 692 493 47 284 702 869 869 889 657 534]

Figure 5: Adjusted R^2 for Case (a) versus indices in all possible combinations (92378) using 19 pulp samples. r is the vector containing the repeated indices in all combinations.



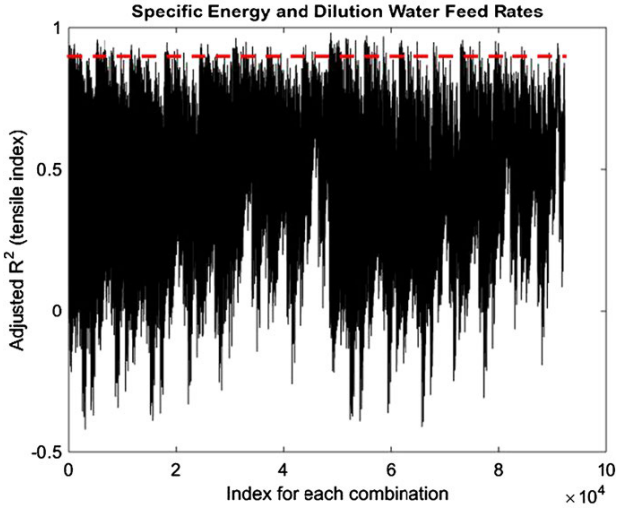
n = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]
r = [6180 7035 7331 6704 10588 6934 6718 670 7236 6091 7469 7051 7592 7247 8025 8990 165 7079 7045]

Figure 6: Adjusted R^2 versus indices for Case (b) in all possible combinations using 19 pulp samples. r is the vector containing the repeated indices in all combinations.



n = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]
r = [268 265 283 307 474 29 205 234 80 210 330 218 264 390 392 272 268 159 162]

Figure 7: Adjusted R^2 versus indices in all possible combinations using 19 pulp samples for Case (c).



n = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]
r = [141 190 219 164 269 200 273 134 242 211 34 33 33 26 45 278 245 181 202]

Figure 8: Adjusted R^2 versus indices in all possible combinations using 19 pulp samples for case (d).

One reason for this poor result is the lack of knowledge about the steam balance using only the dilution water feed rates as predictors. In other words, when using the dilution water, the backward- and forward flowing steam is not taken into account in the same way as it is when using the consistencies as predictors as discussed in Karlström and Eriksson (2014a,b,c,d). This is maybe the most important argument why internal variables should be used though it requires access to the non-linear model.

In conclusion, using this method the best predictors to use seem to be specific energy along with consistencies. This result matches well with the fact that consistency has

been shown to have considerable impact on pulp quality, as demonstrated in Hill et al. (1979) and Johansson et al. (1980).

Using validation data

We will now examine which predictors yield the best result using the other method, where we split the data into two sets: a training set and a validation set. The training set is used to design the model, while the validation set is used to evaluate the model. The evaluation is done by calculating the squared difference between the predicted tensile strength from the model, and the actual tensile strength as found in the validation data set.

In our case we used a training set of data from 12 samples and a validation set from 7 samples, with the samples distributed randomly between the sets. To compensate for the small amount of data in each set we repeated this procedure 10000 times, randomizing new training and validation sets for each iteration. This was repeated for all the cases of predictor choice, with the results presented in Table 2.

Table 2: The mean squared prediction error (MSE) on validation data when using different combinations of variables.

	MSE
Specific energy	0.72
Specific energy and consistencies	0.32
Specific energy and dilution water feed rates	0.9
Specific energy and fiber residence times	0.91

This method confirms that using the consistencies and specific energy yields the best result. However it is not clear that all three of these independent variables contribute to the model quality. To investigate this we test different combinations where we only use two of the predictors, see (Table 3). From Table 3 it seems that the consistency in the conical zone does not significantly improve model quality. This could be due to a few reasons, for example as the zones are serially linked the consistencies between the zones are not entirely independent, especially as no specific effort was made to excite the zones separately when performing the experiment. However, as noted previously the VIF between these predictors is quite low which indicates that they are not particularly correlated.

Further exploring this by calculating the Pearson's correlation coefficient (Randolph and Myers 2013) between the consistency of the conical and flat zone yields a correlation coefficient of 0.65, which indicates a moderate positive correlation between the variables. However it is not so high that there is a reason to believe that the consistency in the two zones was not excited separately (indeed the correlation coefficient between the consistency of the conical zone and specific energy is in absolute terms somewhat larger at −0.77).

Moreover during the experimental procedure special care was taken to increase the amount of work done in the flat zone, which may have contributed to the consistency of the flat zone having a more pronounced impact on the tensile index.

Another reason for the fact that the consistency of the conical zone was not beneficial for model quality can instead be that there was much larger variation of the consistency measurements in the conical zone than in the flat zone during each sampling instance, which may indicate that these measurements were less reliable and therefore less useful for prediction.

However, one cannot entirely exclude the possibility that the consistency in the flat zone has a greater impact on the tensile index then that of the consistency of the conical zone.

Table 3: The mean squared prediction error (MSE) on validation data when using different variables. FZ=Flat zone and CD=Conical zone.

	MSE
Specific energy and FZ consistency	0.31
Specific energy and CD consistency	0.66
FZ and CD consistencies	1.16
Specific energy, FZ and CD consistencies	0.32

As using the consistency in the conical zone does not yield a significant improvement we will disregard it. Hence moving forwards we will use specific energy and the consistency of the flat zone as our only predictors as they appear sufficient to derive good models from our data.

Detecting outliers

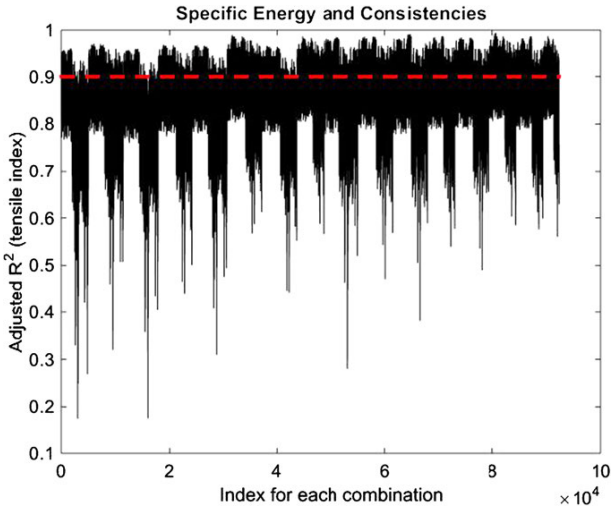
Now we have selected a model structure where specific energy and the consistency of the flat zone are the predictors. However we have not examined the possibility that some of the tensile measurements may be outliers. Looking at

Figure 6 it is clear that two samples do not appear in many good models, namely sample 8 and 17 so these warrant further investigation.

Established outlier detection methods such as DFFITs and Cooks distance (Rawlings et al. 2001) both highlight sample 17 as an outlier. However they do not indicate that sample 8 is an outlier.

We can see from Figure 4 that a step is taken in specific energy during the beginning of the 240 second sampling duration for sample 8. The assumption of our measurement methods is that process conditions are relatively constant during the 240 second sampling duration as we are measuring tensile index from handsheet made from pulp taken during the entire sampling period. Moreover one cannot disregard the risk of mistakes being made when documenting the time of each pulp sample. If we shift our data back 240 seconds so we use measurements taken just before the step and evaluate the results as shown in Figure 9 the results improve considerably, with sample 8 no longer appearing in only a few good models. Moreover, if we examine the mean square error of a validation data set in the same way as when determining which predictors to be used, we find that this change to sample 8's specific energy measurement improves the model, and reduces the mean prediction error on validation data from 0.31 to 0.26. This seems to indicate that the time the pulp sample was taken was not correctly recorded.

Hence, from these investigations we conclude that both sample 8 and sample 17 should be regarded as outliers and therefore be disregarded.



n = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]
r = [13238 14407 15418 13608 21641 14890 14696 15718 14959 12255 15577 14627 15629 15254 16475 18498 465 14608 14667]

Figure 9: Adjusted R^2 versus indices in all possible combinations using 19 pulp samples for case (a) with sample 8 using values of specific energy before the step. r is the vector containing the repeated indices in all combinations.

Resulting model

Designing a model on a reduced data set with sample 8 and 17 removed results in the following model

$$\text{Tensile Index} = -51 + 0.032x_1 + 0.9x_2$$

where x_1 is specific energy and x_2 is the consistency in the flat zone. This model was trained on the entire data set, which yielded an adjusted R^2 of 0.89.

While using the consistency of the conical zone does not improve model quality, it may be useful to include it to gauge the effect the consistency of the conical zone has on tensile index. This as the consistency of the conical zone can be controlled separately from the consistency of the flat zone. This makes it a potentially useful control input, especially in cases where one wishes to control more then one pulp property. If we include the consistency of the conical zone we get the following model:

$$\text{Tensile Index} = -62.69 + 0.036x_1 + 0.83x_2 + 0.18x_3,$$

where x_1 is specific energy, x_2 is the consistency in the flat zone and x_3 is the consistency in the conical zone.

If we compare our model with those derived for tensile index from measurements of a CTMP production line (Karlström and Hill 2017a,b) using a similar experimental procedure, we can see that also for this case it was found that using specific energy and internal variables yielded the best result. While there were some differences in operating condition, this seems to indicate a general trend that internal variables such as consistencies are more useful when modelling tensile index then external variables.

Furthermore, Johansson et al. (1980) who examined pulp properties for TMP refiners and found that plate gaps, production and consistencies were sufficient to explain 91 % of changes in pulp quality during there experiments. So there seems to be a general trend of consistency being found to be useful to model a variety of pulp and paper properties.

Process stability using tensile index models

As stated in Karlström et al. (2018) the internal variables (specifically temperature profile and consistencies) can constitute a new concept for CD refiner control. It was shown that refining zone conditions can vary consider-

ably when running the refiners in specific energy control. It was also claimed that the internal variables can be a backbone in future control concepts of CD refiners. This was strengthened by analyzing the control efforts from three different perspectives to 1) reach desired operating conditions by manipulating the maximum temperatures and outlet consistencies out from the flat and conical zones followed by 2) optimize the production (if requested) and finally 3) implement the complete set of control algorithms in automatic mode to maintain the process at the desired production level to stabilize process. The weakness in the study performed in Karlström et al. (2018) was that no pulp and handsheet data were analyzed in more detail. Instead, all pulp and handsheet properties were derived from earlier CTMP experiments performed in Karlström and Hill (2017a,b,c) to indicate stability issues when running the process in control mode. This of course motivates this paper, and to finalize the research in terms of the model procedure for selected TMP properties, we want to reflect upon the dynamics obtained when using the derived models.

Suppose that we use the same methodology as outlined above when estimating the tensile index. We have already concluded that there are large deviations in tensile index measurements. However despite this we managed to design models with an adjusted R^2 of around 0.9, which can be considered acceptable.

Now we have a model for tensile index estimation based on specific energy and consistency in the refining zones of a CD76-refiner. By applying it on process data we can get an indication of the expected variations in tensile index as seen in Figure 10. This opens up for a number of ways to achieve disturbance rejection and tune the process.

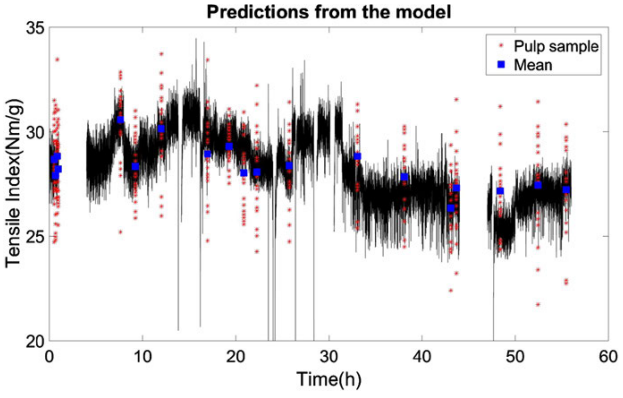


Figure 10: Estimated tensile index for the period studied in Figure 4. Measured tensile index and median measured tensile index included as references.

In this paper the focus has been on specific energy as a predictor. When running refiners in normal modes, without too large changes in the operating points, an alternative is to use the fiber residence times in each refining zone instead of the specific energy as predictors. The reason is that specific energy is related to the total motor load and an assumed production.

By measuring the temperature profile in each refining zones we could provide predictors from both zones, providing additional predictors which could be used for modelling. However, it is possible to divide the work into more than two different zones, potentially providing means to use additional predictors by soft sensor measurements. Moreover, it makes it possible to analyze the impact from different plate patterns etc. When it comes to disturbances related to variable production, it is also possible to capture such changes by analyzing the inner sensors in the sensor array (FZ). For a deeper discussion on this see Karlström and Hill (2017a,b,c).

Concluding remarks

The main purpose of this paper is to show how to derive models for tensile index estimations using specific energy and internal variables in terms of consistencies to the flat and conical zones (FZ and CD) in a CD76-refiner. It is shown that the internal variables outperform the external variables as predictors even though the variance inflation factors are similar. This is similar to what was found for a CTMP refiner in Karlström and Hill (2017b,c). The conical zone may not be necessary, as similarly good models could be derived when using only the specific energy and the consistency in the flat zone. This is not entirely satisfying as the impact of the consistency in the conical zone on the tensile index may be important for future implementation of automatic controllers. Further experimentation where the consistency of the conical zone is excited separately from the other inputs would allow us to better explore what impact, the consistency of the conical zone has on the tensile index. Another possible expansion to consider is to examine the use of other possible predictors such as for example temperature measurements from within the refiner.

It is furthermore worth noting that the model was derived only on one set of measurements, which do not cover all operating conditions. Thus, further investigations on the impact of the other inputs are also warranted.

We also argue that outliers can be detected by using the methodology outlined in this paper. By testing differ-

ent combinations of samples we found that outliers in the data could be detected. For this data set this proved to be particularly effective, finding an outlier that conventional outlier detection methods missed.

This opens for a natural outlier detection and forms a first step toward an on-line implementation of the tensile index model. It is believed that stability can be obtained faster and control be more precise in future applications utilizing this model as it most likely will be possible to optimize the process in terms of temperature control, consistency control, energy control as well as pulp and handsheet property control.

The modelling procedure used here can probably also be used for other pulp and handsheet properties. This as they face many the same issues as when modelling tensile index such as limited and unreliable data. The reason why tensile index is in focus is that this variable is hard to measure online, but important for final property follow-up.

Acknowledgments: The authors gratefully acknowledge the funding of the Norges forskningsråd (NFR). Special thanks go tothe Norske Skog Skogn mill for running trials and providing the excellent process data used in this study.

Funding: The study was funded by the Norges forskningsråd (NFR), Grant No. 256449.

Conflict of interest: The authors declare no conflicts of interest.

References

Bajpai, P. Biotechnology for Pulp and Paper Processing. Springer, New York, USA, 2012.

Barnett, V., Lewis, T. Outliers in Statistical Data. 3rd edition. J. Wiley & Sons, Chichester, UK, 1994. .

Belsley, D.A., Kuh, E., Welsch, R.E. Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, Hoboken, USA, 1980.

Bengtsson, F., Karlström, A., Hill, J., Johansson, L. (2019) Raw data for tensile index estimations from a CD72-refiner. Technical report. Chalmers University of Technology. Available at https://research.chalmers.se/publication/510615.

Draper, N.R., Smith, H. Applied Regression Analysis. vol. 326. John Wiley & Sons, New York, USA, 1998.

Fernando, D., Gorski, D., Sabourin, M., Daniel, G. (2013) Characterization of fiber development in high-and low-consistency refining of primary mechanical pulp. Holzforschung 67:735–745.

Forgacs, O. (1963) The characterization of mechanical pulps. Pulp Pap. Mag. Can. 64:89–118.

Freedman, L.S., Pee, D. (1989) Return to a note on screening regression equations. Am. Stat. 43:279–282.

Gilbert, R.O. Statistical Methods for Environmental Pollution Monitoring. John Wiley & Sons, New York, USA, 1987.

Härkönen, E., Huusari, E., Ravila, P. (2000) Residence time of fibre in a single disc refiner. Pulp Pap. Can. 101:330–335.

Härkönen, E., Kortelainen, J., Virtanen, J., Vuorio, P. (2003) Fiber development in TMP main line. In: Int. Mech. Pulping Conf., Quebec, Canada. pp. 171–178.

Harrell, J.F., Lee, K.L., Matchar, D.B., Reichert, T.A. (1985) Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat. Rep. 69:1071–1077.

Hill, J., Westin, H., Bergström, R. (1979) Monitoring pulp quality for process control. In: Int. Mech. Pulping Conf., Toronto, Canada. pp. 111–125.

Johansson, B.L., Karlsson, H., Jung, E. (1980) Experiences with computer control, based on optical sensors for pulp quality, of a two-stage TMP-plant. In: Process Control Conference, Halifax, Nova Scotia. pp. 145–152.

Karlström, A., Eriksson, K. (2014a) Fiber energy efficiency Part I: Extended entropy model. Nord. Pulp Pap. Res. J. 29:322–331.

Karlström, A., Eriksson, K. (2014b) Fiber energy efficiency Part II: Forces acting on the refiner bars. Nord. Pulp Pap. Res. J. 29:332–343.

Karlström, A., Eriksson, K. (2014c) Fiber energy efficiency Part III: Modeling of bar-to-fiber interaction. Nord. Pulp Pap. Res. J. 29:401–408.

Karlström, A., Eriksson, K. (2014d) Fiber energy efficiency Part IV: Multi-scale modeling of refining processes. Nord. Pulp Pap. Res. J. 29:409–417.

Karlström, A., Hill, J. (2017a) CTMP process optimization Part I: Internal and external variables impact on refiner conditions. Nord. Pulp Pap. Res. J. 32:35–44.

Karlström, A., Hill, J. (2017b) CTMP process optimization Part II: Reliability in pulp and handsheet Measurements. Nord. Pulp Pap. Res. J. 32:253–265.

Karlström, A., Hill, J. (2017c) CTMP process optimization Part III: On the prediction of Scott-bond, z-strength and tensile index. Nord. Pulp Pap. Res. J. 32:266–279.

Karlström, A., Hill, J., Ferritsius, O., Ferritsius, R. (2016a) Pulp property development Part III: Fiber residence time and consistency profile impact on specific energy and pulp properties. Nord. Pulp Pap. Res. J. 31:300–307.

Karlström, A., Hill, J., Ferritsius, R., Ferritsius, O. (2015) Pulp property development Part I: Interlacing undersampled pulp properties and TMP process data using piece-wise linear functions. Nord. Pulp Pap. Res. J. 30:599–608.

Karlström, A., Hill, J., Ferritsius, R., Ferritsius, O. (2016b) Pulp property development Part II: Process nonlinearities and their influence on pulp property development. Nord. Pulp Pap. Res. J. 31:287–299.

Karlström, A., Hill, J., Johansson, L. (2018) An overview of some efforts to understand CD-refiners. In: Int. Mech. Pulping Conf., Trondheim, Norway.

Karlström, A., Johansson, L., Hill, J. (2019) On the modeling of tensile index from larger data sets. Nord. Pulp Pap. Res. J. 34:289–303.

Nelsson, E. (2016) Improved energy efficiency in mill scale production of mechanical pulp by increased wood softening and refining intensity. Ph.D. thesis, Mid Sweden University.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R. (1996) A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49:1373–1379.

Randolph, K.A., Myers, L.L. Basic Statistics in Multivariate Analysis. Oxford University Press, Oxford, UK, 2013.

Rawlings, J.O., Pantula, S.G., Dickey, D.A. Applied Regression Analysis: A Research Tool. Springer Science & Business Media, New York, USA, 2001.

Rosner, B. (1983) Percentage points for a generalized ESD many-outlier procedure. Technometrics 25:165–172.

Sabourin, M., Wiseman, N., Vaughn, J., et al. (2001) Refining theory considerations for assessing pulp properties in the commercial manufacture of TMP. In: 55th Appita Annual Conference. Appita Inc.. pp. 195–204.

Strand, B. (1996) Model-based control of high-consistency refining. Tappi J. 10:140–146.

Strand, B., Grace, B. (2014) Implementation of advanced supervisory control within a TMP refiner quality control system. In: Int. Mech. Pulping Conf., Helsinki, Finland.