Data Analysis in Moving Windows for Optimizing Barley Net Blotch Prediction

Outi Ruusunen[1], Marja Jalli[2], Lauri Jauhiainen[2], Mika Ruusunen[1], and Kauko Leiviskä[1]
[1]University of Oulu, Control Engineering, Environmental and Chemical Engineering Research Unit, Oulu, Finland
[2]Natural Resources Institute Finland, Jokioinen, Finland
Email: outi.ruusunen@oulu.fi

***Abstract*—In modern agriculture, the pesticides and the need to decrease their use is under discussion. Optimization methods and modelling tools are important research areas in this context. In this paper, data analysis, feature generation and selection in moving windows have been utilized for the evaluation of net blotch risk in barley. Two different datasets: The open data from the Finnish Meteorological Institute and the historical observation of the net blotch severity in different fields in Finland are combined with feature generation techniques. T-test is then applied to select the most statistically suitable features for prediction the net blotch risk from weather measurements. Analysis proceeds in moving data windows to indicate the most informative time period to predict the risk of net blotch during the growing season. Results show that the selection of the proper time instance and the length of data window may enhance strongly the potential performance of prediction methods for risk analysis on plant disease.**

***Index Terms*—advanced data analysis, moving window, barley net blotch, t-test, feature generation, IPM**

I. Introduction

The modern agriculture and the higher efficiency of crop production are in the focus of global research. Worldwide discussion is related to innovations, which help to fulfil the demands of growing population. The challenges in crop farming include increased risks caused by insects and plant diseases besides the global warming and the natural disasters. Forecasting of different insects and plant disease risks is one tool that has paid attention. For example, Donatelli *et al.* [1] have published the state of the art review and discussed the challenges in insect or plant disease modelling. Jones *et al.* [2] have reviewed the agricultural systems modelling and briefly its history and in [3], discussed the current state of agricultural systems science, and the capabilities and limitations of agricultural systems models. El Jarroudi *et al.* [4] have studied the meteorological conditions in the development of Septoria leaf blotch disease forecasting in winter wheat. One application about the situation awareness in environmental monitoring and disease outbreak in agriculture have presented and discussed in the article on Stocker *et al.* [5].

Barley is globally important cultivated cereal and in 2017, it was grown on 47 million hectares [6]. Plant diseases are affecting the yield and can decrease it up to 20% of the annual average barley yield [7]. One common fungal disease in barley is net blotch, which is caused by the ascomycete *Pyrenophora teres* Drechsler. It is reported that P. teres correlates positively with temperature, but negatively with relative humidity and leaf wetness [8].

Barley net blotch is commonly prevented by pesticides. European Union has codified the Integrated Pest Management (IPM) directive to control the use of pesticides and the IPM directive guides to justified and well-documented pesticide use. The overuse of chemicals is harmful for environment and an unnecessary cost for a farmer, thus the tools for pest forecasting and optimizing the spraying needs a thorough study. Pesticide residues are a burning topic in global food safety discussion [9].

This study focuses on barley net blotch, which is the most important yield reducing plant pathogen in Finnish barley [10]. Data analysis and feature generation methods are used to enrich the information content of existing data from past decades. The aim is then to predict the net blotch density by utilizing the weather measurements and the long-term net blotch observation data. The existing datasets have been utilized for a new purpose and no extra measurement campaigns have been arranged. The main idea is to study, if it is possible to group the yearly weather data according to the density of net blotch in successive moving time windows to find an optimal starting point and window size for the prediction of the plant disease risk.

As the plant disease severity depends on the environmental conditions, it is important for any forecasting method that the meteorological data includes enough information for risk evaluation. In this context, the starting time of the forecast and the length of the time window are of great interest in developing any decision support systems for plant disease forecasting. The earlier study and the principle of weather measurement utilization in net blotch forecasting is presented in [11].

This study applies time series analysis in successive, moving time windows as for example in Nikula *et al.* [12]. Remarkably, this principle aims to identify the most informative time windows from the beginning of the growing seasons that enable predictive classification to the high or low severity of net blotch, solely based on

Manuscript received May 9, 2020; revised October 28, 2020.

doi: 10.18178/joaat.7.2.38-42

weather data. The analysis in each time window utilizes feature generation from weather data together with a simple statistical t-test. This way, a systematic analysis for the optimal starting point and window length is conducted along the studied time series.

The rest of the paper presents the analyzed data, the applied analysis procedure in detail and the findings from the data. Finally, discussion based on the results is given.

II. Materials and Methods

A. Data

In this study, two different datasets - the weather measurements and field observations of net blotch density from the test fields are combined. The weather data is downloaded from the open database of the Finnish Meteorological Institute (FMI) and the net blotch observations are from the official variety trials database of Natural Resources Institute Finland (Luke). The yearly observation period is 1991 – 2017 and the test fields were located in the Central and Southern Finland in four different cropping zones (I-IV). The net blotch density is scaled to categories 0, 1 and 2, where the category 0 means the maximum net blotch severity value 0.5%, 1 means 0.6-5% net blotch severity and the category 2 means over 5.1% net blotch severity in the test field. In this research, the category 0 and 1 net blotch data were utilized and the amounts of used years (datasets) per cropping zone are listed in Table I.

TABLE I. The Amount of Data for Analysis

	Data amount; category 0	Data amount; category 1
Cropping zone I	4	4
Cropping zone II	4	4
Cropping zone III	4	4
Cropping zone IV	3	4
Cropping zones I- IV	15	16

The studied variables in the weather data were:

- The place of observation,
- The date of observation,
- The rainfall per day [mm], R,
- The average temperature per day, T_{av} [℃],
- The daily minimum temperature, T_{min} [℃] and
- The daily maximum temperature, T_{max} [℃].

The FMI datasets were grouped depending on the observation place (cropping zone I - IV) and the net blotch severity (categories 0-1). The grouping principle is presented in Table II.

TABLE II. FMI Data Grouping Principle

	Net blotch appearance	Net blotch category
Cropping zone I	Yes No	1 0
Cropping zone II	Yes No	1 0
Cropping zone III	Yes No	1 0
Cropping zone IV	Yes No	1 0

B. Data Analysis and Feature Generation

The data analysis and the feature generation methods are utilized for enriching the information content of data and to find the most informative time windows of each growing season datasets. The beginning of the growing season varies yearly because of the different weather conditions. To get the datasets comparable, the beginning of the growing season was firstly identified and set to time step k=0. Here, the growing season was expected to start when the daily mean temperature value remained over plus five degrees of Celsius at five consecutive days. The utilisation principle for the data sets is in Fig. 1.

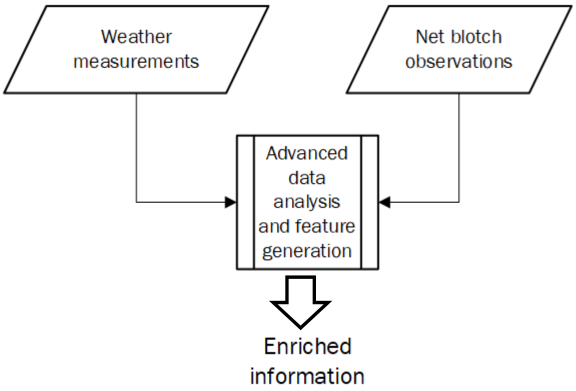

Figure 1. The data usage for the analysis.

In feature generation, new variables were computed applying common mathematical operations to the original variables. The operations were summation, subtraction, multiplication, division, involution, logarithm, square root and the different combinations of those. This results in transformed variables, features. More about the feature generation technique is presented for example in Blum and Langley [13] and Garcia-Torres *et al.* [14]. The technique, which is utilized here, is presented first in Ruusunen [15] with details and the tested features in Ruusunen [15, appendix 1].

In this case, the suitable features were those, with which the weather dataset indexed with 1 could be separated statistically from the category 0 datasets. All features listed in [15] were tested and the feature validation was performed with t-test (Eq. (1)). That makes 110 feature candidates, which were tested further with 16 variable combinations at each time step. The analysis of four different cropping zones was performed in Matlab® environment with the Matlab® function *ttest2* and with 70% confidence interval. To test the generalization of used analysis, the cropping zone I-IV datasets were fused together and then the described analyzing steps were performed with 80% confidence intervals. The amount of data increases by fusing the category I - IV datasets (Table I), for this reason 80% confidence value was applied for feature selection in this case. Two-sample t-test was performed with the assumption that the pair of data vectors consists of independent random samples with unknown variance in the applied short time windows.

Two-sample t-test statistic is described here for the classification criterion, t, as:

$$t = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{s_x{}^2}{n}+\frac{s_y{}^2}{m}}}, \qquad (1)$$

where $\overline{x}$ and $\overline{y}$ are the sample means of the two studied data vectors, s_x and s_y are their sample standard deviations respectively, and n and m are the sample sizes. In this case, where two data samples are assumed to be from the populations with unequal variances, the test statistic t under the null hypothesis has an approximate Student's t distribution with a number of degrees of freedom given by Satterthwaite's approximation. This arrangement can be called also Welch's t-test. [16]

Fig. 2 presents an example of the feature generation and evaluation for single feature. Same procedure was repeated with every feature in the observation window.

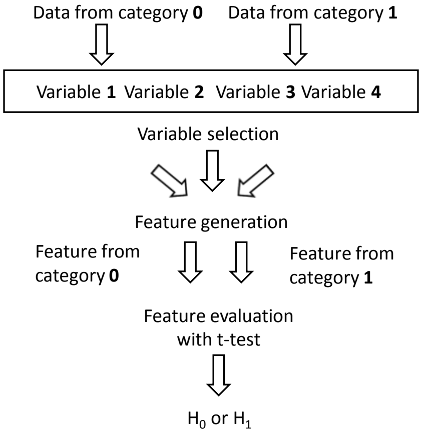
Data from category **0**
Data from category **1**
Variable **1** Variable **2** Variable **3** Variable **4**
Variable selection
Feature generation
Feature from category **0**
Feature from category **1**
Feature evaluation with t-test
H_0 or H_1

Figure 2. Feature generation and validation for a single generated feature. Mathematically transformed (feature generation) category 1 data is compared to category 0 data with same variable transformation applying t-test.

The most informative time period that maximizes the number of daily alternative hypotheses (H_1) is studied by the moving time window technique. The analysis was repeated at time steps

$$k, k+1, k+2, \ldots, k+n \qquad (2)$$

where k is the beginning of the growing season and n is 50 days. According to Fig. 3, the analysis proceeds in time windows, starting from the beginning of the growing season and proceeding at steps of one day for 50 days (Rounds). Three time windows were tested: 7, 14 and 21 days.

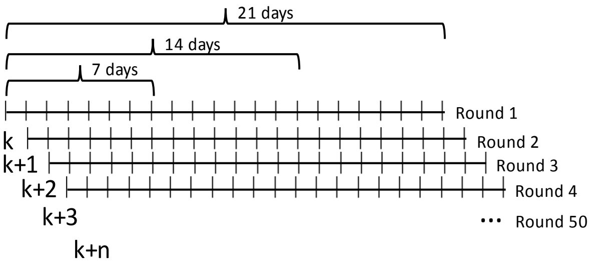

Figure 3. The applied moving window procedure.

All the generated features were tested and the data from net blotch categories 1 was compared to the reference data category 0 (no net blotch occurrence) with t-test (Eq. (1)) and the following hypotheses:

H_0 = The daily feature values have equal means and equal but unknown variances in tested datasets,

H_1 = The daily feature values have unequal means.

The number of days when feature values from category 1 statistically differ from reference data is computed (Fig. 3). The feature generation and classification is demonstrated with four years datasets from both categories. The t-test analysis is performed on a daily basis. Then, the number of realised alternative hypotheses H_1 is summed and presented as the number of separable days. It means that the maximum number of separable days (days with potential net blotch) for example at the time window of 7 days is also 7.

III. Results and Discussion

The analysis results are presented in Fig. 4 for each cropping zone and window size. In Fig. 4-Fig. 6 the results of different, time windows are marked as 7 days with (-*-), 14 days with (-o-) and 21 days with (-x-). There, the results with the best separating features according to t-test are shown as the number of separable days at Y-axis. The X-axis is time (0-50 days). It has to be noted that the maximum Y-value varies according to the window size. According to the data analysis, the best separating feature varied between window size and cropping zones remarkably.

The results (Fig 4a-4d.) suggests that the optimal starting point for forecasts differs in the south (cropping zone I) from northern Finland (cropping zone IV). This can be seen as varying maximum peak instances in separable days, namely maximum values reaching in 1 - 35 days at cropping zones I and II, whereas at cropping zones III and IV the maximum values are reached at 40 - 50 days.

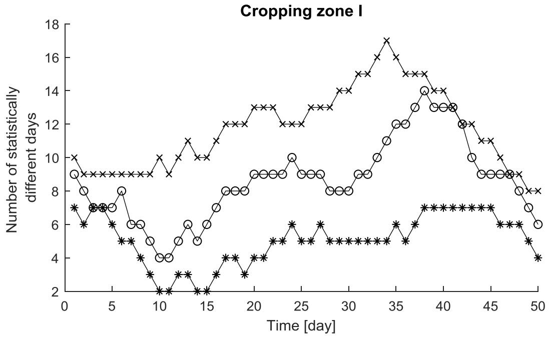

Figure 4a. Data analysis results for cropping zone I.

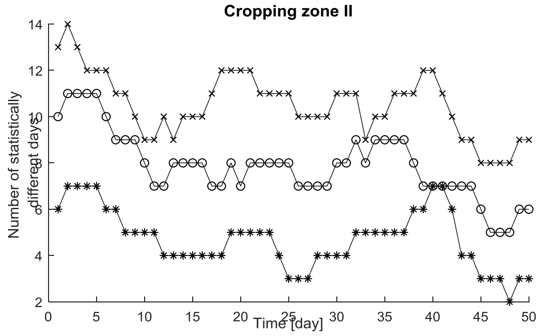

Figure 4b. Data analysis results for cropping zone II.

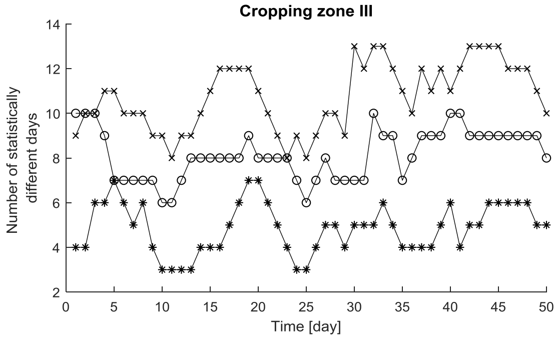

Figure 4c. Data analysis results for cropping zone III.

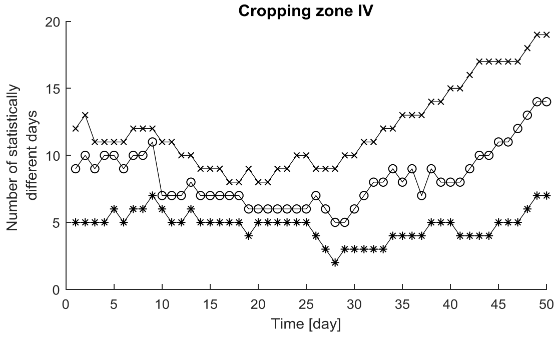

Figure 4d. Data analysis results for each cropping zone IV.

Fig. 4 a-d show data analysis results for each cropping zone. Time windows of 7 (-*-) 14 (-o-) and 21 (-x-) days. Y-axis: number of statistically different days between net blotch groups 0 and 1 according to t-test at 70% confidence interval X-axis: time (0-50 days).

The datasets from four different cropping zones have been combined together resulting data with 15 years weather measurements in category 0 and 16 years weather data in category 1. The data of cropping zones I-IV have been analysed as described above. With the t-test, the most informative window size and the analysis starting point have been evaluated. The results are presented in Fig. 5 and in Fig. 6. There, window sizes of 21 and 14 days seem to produce similar curves of separable days, peaking at 20 days.

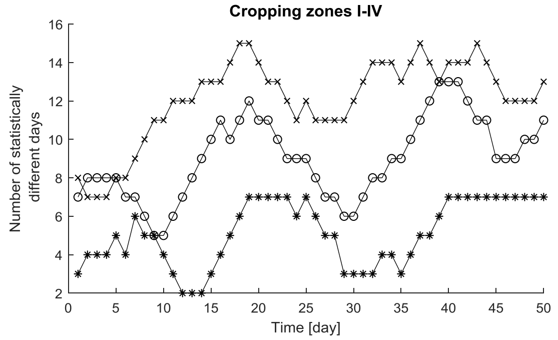

Figure 5. Analysis result for all data sets combined (cropping zones I-IV).

To evaluate the suitability and robustness of analysis with different window sizes, the analysis results were normalised by dividing the number of separable days with the size of a time window. In Fig. 6, the normalised results for the analysis of all data sets are presented. Based on these results, variance related to amount of separable days seem to be highest at the shortest time window of 7 days. On the other hand, the shortest time window provides a full separation with respect to number of days (7 out of 7) between rounds 16 and 24. In this case, these time instances would be the optimal starting points for prediction. When compared to Fig. 4 with different cropping zones, the optimal starting point and window size varies between observation fields. For example, (Fig. 5 and Fig. 6) the feature with the following formulation resulted in the highest number of separable days at Rounds 40 and window size of 14 days: *(log (R)+log(T_{max}) log(T_{av})).*

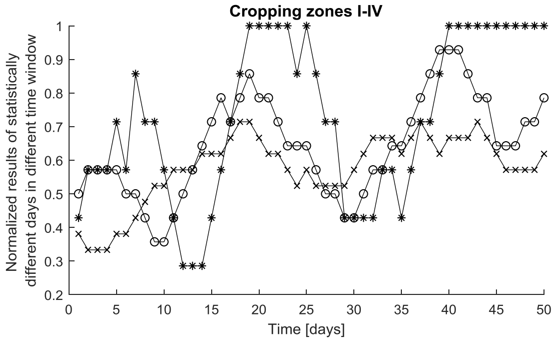

Figure 6. The normalised results for the analysis of cropping zone I-IV data sets. Time windows of 7 (-*-) 14 (-o-) and 21 (-x-) days. Y-axis: normalized results of statistically different days between net blotch groups 0 and 1 according to t-test at significance 0.2. X-axis: Round number (0-50 days).

The optimum time period, or actually the most informative data set according to the above described analysis and t-test is the *k*=19…25 days or *k*=39…50 days with the window size seven days. From statistical point of view, the seven days forms a small amount of information and can lead to overfitting in modelling. For that reason, these results need careful validation.

In Fig. 7, an example of net blotch prediction based on the analysis results is presented for cropping zones I-IV. There, cumulative summed values of a single generated feature *(log (R)+log(T_{max}) log(T_{av}))* are plotted starting at points *k* = 40 and *k* = 30 with time window of 14 days. It can be seen that occurrence of net blotch (dashed lines) differ somewhat from the reference data sets (solid lines) in Fig. 7a (*k* = 40 days. In Fig. 7b (*k* = 30 days), the prediction starting point is not the optimum, and the different datasets (category 0 and category 1, Table II) cannot be separated with presented feature.

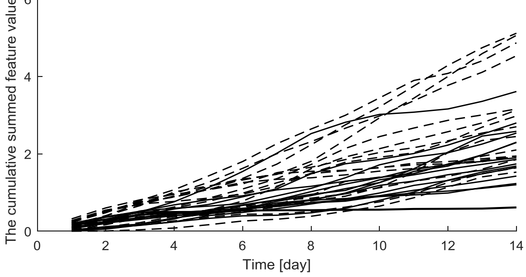

Figure 7a. Example of net blotch prediction, with optimal starting point k = 40 days and window size 14 days.

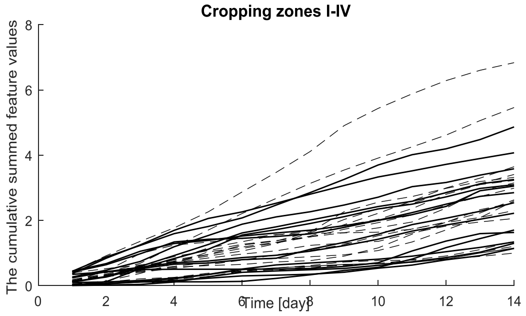

Figure 7b. Example of net blotch prediction, with randomly chosen starting point k = 30 days and window size 14 days.

Fig. 7a-b show example of net blotch prediction based on the analysis results for cropping zones I-IV. Category 0 data: solid lines, category 1 data: dashed lines. Both results are achieved with the feature *(log (R)+log(T_{max}) log(T_{av}))*, but the starting point is 30 days in upper figure and 40 in lower figure.

IV. Conclusion

In this paper, the moving window method is used to find the most informative time period of each studied datasets. The information content of the existing weather data was enriched by advanced data-analysis, thus avoiding extra measurement campaigns. The feature generation methods were utilized to combine two existing datasets for estimation of net blotch density. T-test was then applied to select the most statistically suitable features for prediction. Finally, the analysis is performed in moving data windows to find the most informative time period for each growing season.

According to the results, the selection of the proper starting time and length of data window may enhance strongly the performance of net blotch prediction. The data-based modelling or as here, the risk estimation based on measurements and observations rests on the reliable and representative data. For that reason, the proper data pre-processing and analysing are in the key role here. This research increase the knowledge about the optimal time period selection in the case of barley net blotch density and these results will be utilized in further research of predictive net blotch risk estimation during growing season.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Outi Ruusunen conducted the research, and wrote original draft preparation, and data analysis; Marja Jalli wrote and edited the paper, Lauri Jauhiainen collected and pre-processed net blotch density data, wrote and edited the paper; Mika Ruusunen wrote and edited the paper; Kauko Leiviskä supervised and edited the paper; all authors had approved the final version.

Acknowledgement

The results were achieved during the MaDaKas project, financially supported by Ministry of Agriculture and Forestry of Finland. The authors are acknowledging The Ministry for financial support.

References

[1] M. Donatelli, R. D. Magarey, S. Bregaglio, L. Willocquet, J. P. M. Whish, and S. Savary, “Modelling the impacts of pests and diseases on agricultural systems,” *Agricultural Systems*, vol. 155, pp. 213-224, 2017.

[2] J. W. Jones, *et al.*, “Brief history of agricultural systems modeling,” *Agricultural Systems*, vol. 155, pp. 240-254, 2017.

[3] J. W. Jones, *et al.*, “Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science,” *Agricultural Systems*, vol. 155, pp. 269-288, 2017.

[4] M. E. Jarroudi, *et al.*, “Improving fungal disease forecasts in winter wheat: A critical role of intra-day variations of meteorological conditions in the development of Septoria leaf blotch,” *Field Crops Research*, vol. 213, pp. 12-20, 2017.

[5] M. Stocker, *et al.*, “Representing situational knowledge for disease outbreaks in agriculture,” *Journal of Agricultural Informatics*, vol. 7, no. 2, pp. 29-39, 2016.

[6] FAO. (2018). FAOSTAT. [Online]. Available: http://www.fao.org/faostat/en/#data

[7] G. M. Murray and J. P. Brennan, “Estimating disease losses to the Australian barley industry,” *Australasian Plant Pathology*, vol. 39, pp. 85-96, 2010.

[8] A. R. Martin and K. S. Clough, “Relationship of airborne load of Pyrenophora teres and weather variables to net blotch development on barley,” *Can. J. Plant Pathol.*, vol. 6, pp. 105-110, 1984.

[9] G. Kaushik, S. Satya, and S. N. Naik, “Food processing a tool to pesticide residue dissipation - A review,” *Food Research International*, vol. 42, pp. 26-40, 2009.

[10] M. Jalli, P. Laitinen, and S. Latvala, “The emergence of cereal fungal diseases and the incidence of leaf spot diseases in Finland,” *Agricultural and Food Science*, vol. 20, no. 1, pp. 62-73, 2011.

[11] O. Mäyrä, M. Ruusunen, M. Jalli, L. Jauhiainen, and K. Leiviskä, “Plant disease outbreak – Prediction by advanced data analysis,” *SNE Simulation Notes Europe*, vol. 28, pp. 113-115, 2018.

[12] R. P. Nikula, M. Ruusunen, and K. Leiviskä, “Data-driven framework for boiler performance monitoring,” *Applied Energy*, vol. 183, pp. 1374-1388, 2016.

[13] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245-271, 1997.

[14] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, and J. M. Moreno-Vega, “High-dimensional feature selection via feature grouping: A variable neighborhood search approach,” *Information Sciences*, vol. 326, pp. 102-118, 2016.

[15] M. Ruusunen, “Signal correlations in biomass combustion - An information theoretic analysis,” PhD thesis, University of Oulu, 2013.

[16] Matlab Help Documentation, T-test, © 1994-2017, The MathWorks, Inc., referred in 20th June, 2019.

Outi Ruusunen (née Mäyrä) is a post-graduate student (M. Sc. (tech)) in Control Engineering research group in Environmental and Chemical Engineering Unit at University of Oulu. Her research interests include data analysis both in industrial and non-industrial applications and process modelling. She is active with the teaching development and takes studies on education in addition to PhD studies. She has been the responsible project leader in Modelling and data analysis as a tool in plant protection-project since 2018.