LANDTECHNIK 75(2), 2020, 51–59

DOI:10.15150/lt.2020.3232

LANDTECHNIK
AGRICULTURAL ENGINEERING

Investigation of the efficiency of work processes within a farm using open data

Thoralf Stein

Research has long been concerned with the investigation and optimization of agricultural processes. Still, there is potential to make processes more resource-efficient. An important step in exploiting this potential is to examine whether differences in efficiency within a company are due to natural reasons such as soil properties or other reasons such as vehicle control. In this work a possible procedure for such an investigation is presented. For this purpose, recorded machine data from the project "BiDaLAP" is combined with various open databases. With that, efficiency parameters can be quantified. A large part of the natural causes for efficiency differences can be identified by the databases. It can also be shown which operational parameters still have optimization potential and how big the efficiency differences really are.

Keywords
Work efficiency, Big Data, Open Data, Support Vector Regression

The efficiency of an agricultural work process depends on various parameters. Many scientific papers on this topic deal with different approaches to examine productivity and efficiency. The results of such examinations depend heavily on the definition of the term "efficiency", which is selected or examined in the corresponding work. It can be examined with regard to resource efficiency, the carbon footprint (Lal 2004), cost efficiency or yield per hectare (Kudaligama and Yanagida 2000). It can also concern the energy efficiency (Toll 2013) as well as the degree of sustainability of individual companies (Pacini et al. 2003). The problem, however, is that most of the considerations require clear local and time limits. These are usually at individual companies or regions, because properties such as height profile, soil properties or field geometry can be just as different as the weather conditions when carrying out individual work processes. It is therefore rare or only possible to transfer the test results to a limited extent.

The aim of this work is to minimize the above-mentioned influences using a correction procedure and to make the efficiency values more comparable. It should be possible, for example, to compare individual work processes in such a way that only the operator-dependent efficiency differences are visible and quantifiable. It shows how big the differences can already be within a company and how well a correction procedure is suitable for adjusting the efficiency.

The process was developed using process data from the project "Innovative use of big data in agricultural processes" (BiDaLAP). The BiDaLAP project aims to develop an electronic infrastructure consisting of a platform architecture and mobile data loggers. The development will be available in the future as an operational and strategic decision support system.

Furthermore, various open databases were included in the correction procedure in order to be able to quantify as many influences as possible. For example, data from the Deutscher Wetterdienst (DWD), the Deutsches Zentrum für Luft und Raumfahrt (DLR) and Bundesanstalt für Geowissen-

received 17 June 2019 | accepted 25 March 2020| published 4 May 2020

schaften und Rohstoffe (BGR) were included. All databases used are described in more detail below. Furthermore, the correction procedure is explained and the results are presented afterwards.

Material and method

During the BiDaLAP project, a test farm was equipped with mobile data loggers that record the GPS position at 1 Hz intervals. The crop years 2017 and 2018 of the farm in Saxony were thus recorded. As an example, this article analyses and compares all work that was carried out with a Väderstad precision seed drill. After clearing the GPS data, 48 operations with 2 different tractors on 40 different fields are available for analysis. The area output is chosen as a measure of the efficiency for this process. The quantity of the seed applied nor the fuel consumption of the machines can be extracted from the data. The quality of the work is also unknown. The distribution of the area output is shown in Figure 1.

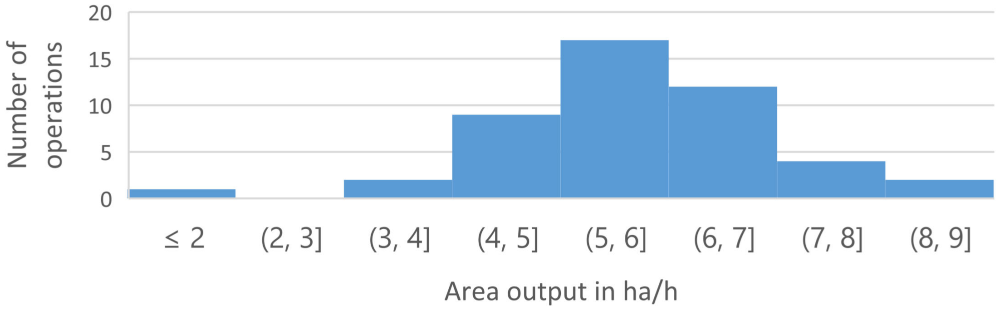

Figure 1: Distribution of the area output of the work processes with the seed drill

It can be clearly seen that the work efficiency of the individual processes deviates significantly. It is important to analyse whether this is due to the environmental characteristics of the company or if the machine operator has an influence on it. The following openly accessible databases are used:

- Deutscher Wetterdienst (DWD): The german weather service offers a variety of data that includes weather stations and also interpolated square grids with an edge length of 1 km, which are calculated based on the data from the stations. The soil moisture and the amount of rainfall are required for the model (Stein and Henschel 2019).
- Deutsches Zentrum für Luft und Raumfahrt (DLR): The german aerospace center offers data from the TanDEM-X mission. With the radar satellites of the TanDEM-X mission, a highly precise, digital 3D image of the earth is recorded. Precise elevation data are collected in a 12 m grid for the entire earth and converted into a uniform map material. For the model, the climbing resistance can be determined using the height data.
- Bundesanstalt für Geowissenschaften und Rohstoffe (BGR): The BGR is the Federal Institute for Geosciences and Natural Resources and offers comprehensive soil mapping of various accuracies or establishes a connection to the state offices. The type of soil with the associated sand, clay and silt ratio is required for the model (Düwel et al. 2007).

The only non-open data that are obtained directly from the farmer are the GPS tracks as well as field data and borders. The field data could also be taken from the geoportal of the respective federal state. The combination of these open databases and the in-house GPS tracks make up the approach of this work. Next, as many environmental parameters as possible are extracted from the databases that can have an impact on the area performance. Some values were not taken directly from the databases, but were subsequently calculated in order to optimally reflect the properties of the environment (Table 1).

Table 1: Environmental parameters with the corresponding correlation coefficients for area output, green positive correlation, red negative correlation

Parameter	Unit	Correlation coefficient	Parameter	Unit	Correlation coefficient
intensity of the height profile	m	-0.07	sand ratio	%	0.20
soil temperature	°C	0.01	year		-0.04
soil moisture	%	0.01	month		0.04
daily rainfall	mm	0.25	track length	m	0.52
clay ratio	%	-0.19	complexity of the field geometry		-0.14
slit ratio	%	-0.19	machined area	ha	0.46

The operationalisation of certain parameters was carried out as described below:

- Intensity of the height profile:
  The intensity of the height profile was calculated using DLR's altitude data. It can be well represented by the standard deviation slope. A higher standard deviation implies a higher change in altitude during the operation. Therefore, the slope along the GPS track is determined and its standard deviation is calculated.
- Track length:
  The track length refers to the distance travelled during work in the field between two turns. For automatic determination, the GPS data was divided by an algorithm into working points and turning points and the distance between two turns was calculated. The mean track length was chosen as a comparative value.
- Complexity of the field geometry: The differences in the field geometry are shown in Figure 2. The more angled a field is, the longer the machining time. In order to be able to drive over all parts of the field completely, additional driving manoeuvres must be carried out. This requires additional working time and thus the area output is reduced. For quantification, the number of angle changes of the field geometry is chosen, which is greater for complex geometries than for simple

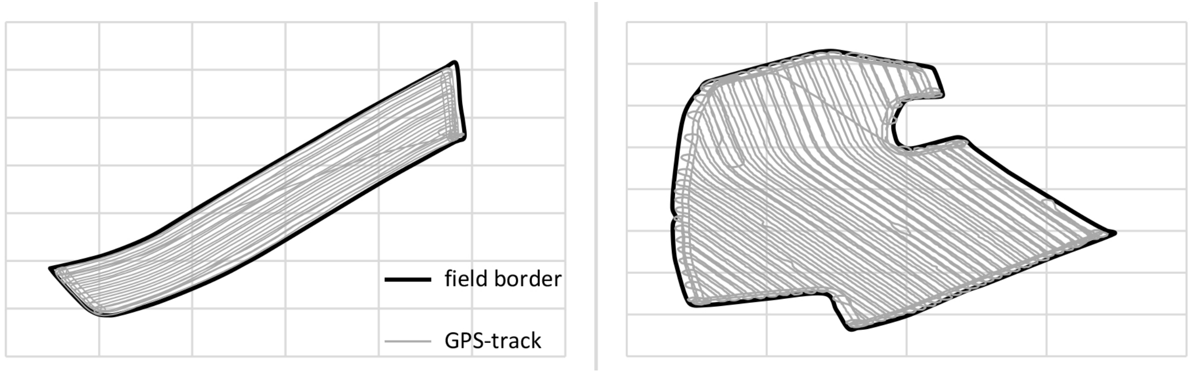

Figure 2: Display of two different field geometries with GPS track data, left simple geometry (value=4), right complex geometry (value=11)

First, the individual variables were compared with their linear correlation coefficient to the area output. This represents the dependence of two variables on each other and has a value range from -1 to 1. 0 means no dependence of the variables, -1/1 strong linear dependence. Furthermore, the p-values of each variable must be calculated in order to be able to make a statement about its significance for the area output (Ross and Heinisch 2006).

It can be seen from the correlation coefficients in Table 1 that the influence on the area output of the parameters differs. The coefficients vary from practically no correlation between soil temperature and area output to a clearly recognisable correlation between track length and area output.

In addition to these natural influences on the area output, there are also the influences that can be traced back to the control of the machine. These are again shown with the corresponding correlation coefficient in Table 2. It can be seen that the coefficients are significantly higher than for the natural parameters. This results from the fact, that these variables have a direct effect on the working time and thus on the area output. The listed parameters were calculated entirely from the GPS lanes and the corresponding time stamp. The p-values are calculated by means of a t-test for all parameters. This resulted in a p-value of less than 0.01 for almost all listed parameters, which indicates a high significance (Bortz and Döring 2006). Only the p-values for the duration of the stops per hour and the complexity of the field geometry are higher than 0.05. This suggests a less obvious statistical correlation between these variables and the area output.

Table 2: Influencing parameters that depend on the operator and their correlation coefficient to area output

Parameter	Unit	Correlation coefficient	Parameter	Unit	Correlation coefficient
stop count per hour		-0.685	actual work width	m	0.638
mean length of stops	s	-0.029	work speed	m/s	0.828
turning time	s	-0.667			

When considering the area output of the individual processes, the question arises how the two parameter groups can be used to describe the area output. Furthermore, it must be possible to distinguish between them again afterwards. A possibility is a regression model. A regression model reflects the functional relationship between several independent variables and a dependent variable (Dodge and Jurečková 2000). In this study, a linear relationship was first investigated (equation 1):

$$Y = a_0 + a_1 \cdot X_1 + \cdots + a_n \cdot X_n \qquad \text{(Eq. 1)}$$

The parameters a_0 to a_n re the so-called regression parameters, which describe the influence of a variable in the model. The area output is selected as the dependent variable and the environmental and operator parameters listed up to this point are selected as independent variables. The independent variables can be used to form a target function that estimates the area output. After the first analysis, it was shown that the mean square error became too large for both a linear and a nonlinear approach. The mean square error is a central quality criterion in statistics. It represents the dispersion of the model around the expected value. In this case, the error was 1.782 ha^2/h^2 for the linear approach and 1.340 ha^2/h^2 for the non-linear approach. These approaches were therefore not applicable.

A Support Vector Regression (SVR) was implemented as an alternative regression approach. This enables the solution of the regression problem in higher-dimensional space with the help of so-called kernel functions. Dolan 2019, Steinwart and Christmann 2008 as well as Cherkassky and Dhar 2010 explain the exact mode of operation of the Support Vector Regression, which is why it will not be discussed in detail here. As a measure of the accuracy of the regression, the mean square error is also chosen here, which reflects the deviation between the recorded area output and the regression model. The coefficient of determination can be used as a further quality feature. This coefficient reflects how well the individual independent variables are suited to explain the variance of the dependent variable (Bortz and Döring 2006). A disadvantage of the SVR is that the model parameters cannot be represented as simply as in formula 1. These are hidden in the support vectors and a representation at this point would not be helpful.

In order to be able to quantify the influence of the parameters, the SVR Support Vector Machine must be trained using the existing work processes. Then all working processes are averaged to the same environmental parameters and a "corrected" area output is determined with the regression model. It can be shown how immense the dependence of the area output on the operator parameters is and which ones have the utmost influence.

Finally, it is shown how large the potential savings of time and costs can be if the corresponding parameter is optimized. This is done using a sample calculation with a 10 ha field, which corresponds to the average field size of the farm under consideration. The field work calculator (KTBL 2018) from the Kuratorium für Technik und Bauwesen in der Landwirtschaft (KTBL) as well as the experience records for inter-farm machine work (Uppenkamp 2018) are used for cost calculation.

Results

First of all, it will be compared how well the regression model represents the actual area output with the available parameters. Figure 3 shows the results for each record with the corresponding model value. It can be seen that there is a usable precision of the model. This is also confirmed by the mean square error of 0.219 ha^2/h^2. The average percentage deviation is 0.4% and the standard deviation 5%. Furthermore, the coefficient of determination is 81%, which is an acceptable value for such a application. After training, the model consists of 43 support vectors, in which the weighting of 20 parameters each is stored. The higher accuracy of the model is due to the better handling of nonlinear and nonmonotonic data. Also, SVM's have the ability, to generalize regression problems very well

(Steinwart and Christmann 2008). It can be concluded that the regression model using a Support Vector Machine is suitable for modelling the process with a precision seeder.

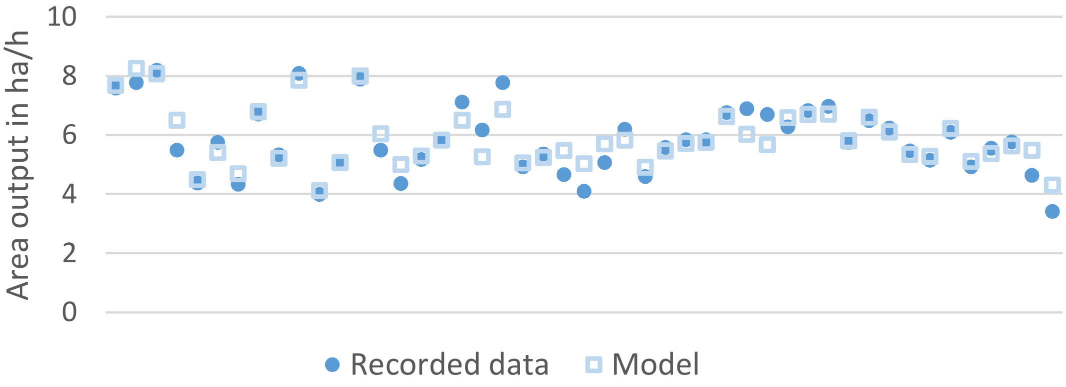

Figure 3: Comparison between recorded area output and regression model

The next step is to adjust the environmental parameters. For this purpose, the averaged parameters of all processes are assigned to each recorded process. The area output is then calculated using the regression model and the adjusted environmental parameters and the original operator parameters. The recorded and the corrected area output are shown in Figure 4 as box diagrams. These graphs clearly show the distribution. The box is bounded by the upper and lower quartile, the total extent represents minimum and maximum of the considered size. It can be seen that the spread of the corrected area output has become considerably smaller in contrast to the recorded one. The original area output ranges between 3.5 and 8.1, the corrected area output only between 4.5 and 7.0. It reflects the deviation of the area output, which is caused solely by the operator. This also means that the environmental parameters have a considerable influence on the area output and must always be taken into account when examining the efficiency of work processes. For the sake of completeness, the model result is also shown without adjusted environmental parameters. It can be seen that the optimum has not yet been achieved here, since the model apparently does not fully reflect the wide spread of the observations.

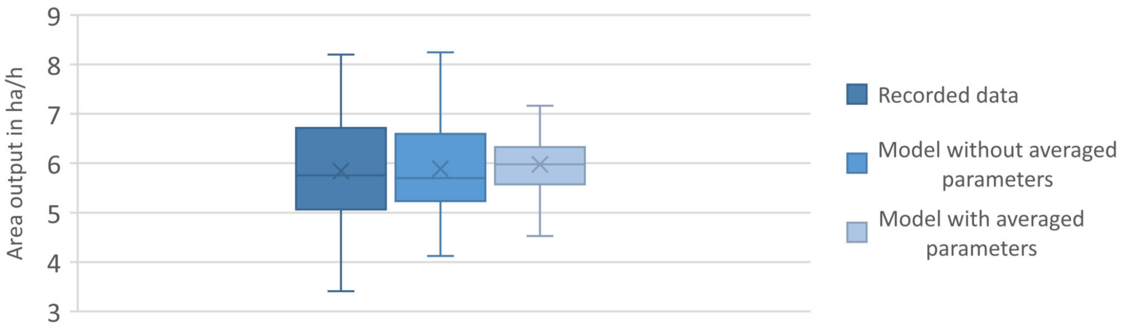

Figure 4: Box diagram of the area performance of the recorded data, as well as the calculated area output of the SVM without and with averaged environmental parameters

In order to determine the influence of the individual operator parameters, the next step is to parameterize the model with the original environmental data and to change only one operator parameter at a time to the best possible value. In order not to overestimate the optimization potential, the 90% quantile is assumed here. It is shown how large the increase in area output could be. The given values are the average increases for all observations. The initial values are the actually observed parameters and the increase refers to the adjustment of the respective optimization parameter. The individual parameters show different effects on the area output. These can already be seen from the correlation coefficients in Table 2, it is clarified again in Table 3. The lowest optimisation potential according to Table 2 is in the stops per hour, followed by the average turning time and the actual working width. Obviously, the work speed offers the most potential. The calculated cost savings in Table 3 result from the reduced working time and the associated labour cost savings and the average tractor costs per hour without fuel.

Table 3: Change in efficiency and savings for a precision seeder if one of the specified parameters equals the 90% quartile; calculations for a 10 ha field

	Stops per hour	Turning time	Actual work width	Work speed
Increased area output	0.21	0.15	0.11	0.32
Time savings	9.7 min	7.1 min	5.2 min	14.5 min
Cost savings	5.50 €	3.98 €	2.94 €	7.41 €

The wage costs are assumed to be 17 €/h for a tractor driver according to the usual rates of experience. Depending on the size, the tractor costs range between 15 and 23 € per hour. If the savings are added up, there is a potential of half an hour, as well as wage and tractor costs of just about 20 €, which corresponds to about 2% of the total costs of the process. This is an acceptable figure for savings that can only be achieved by optimised driving behaviour.

Table 3 shows average effects of the parameters shown on the area output. These effects describe the changes of a dependent variable (the area output) by an independent variable. All other independent variables are kept constant. A comparison of these effects for all independent parameters listed in tables 1 and 2 with the corresponding correlation coefficients is a further indication of the validity of the model (Table 4). The effects were determined using the 90% quantile of the corresponding value. It can be seen that the trends of correlation coefficients and the change in area output are very similar, only the intensity does not match in part. This fact is especially true for the duration of stops and the complexity of the field geometry. This could be due to the relatively high p-values of these independent parameters.

Table 4: Comparison of the correlation coefficients and average effects of the individual independent parameters; marginal effects were calculated using the 90% quantile

Parameter	Correlation coefficient	Changes in	Parameter	Correlation coefficient	Changes in
stop count per hour	-0.685	-0.204	clay ratio	-0.190	-0.008
mean length of stops	-0.029	-0.064	slit ratio	-0.190	0.013
turning time	-0.667	-0.081	sand ratio	0.200	0.115
actual work width	0.638	0.127	year	-0.040	-0.093
work speed	0.828	0.323	month	0.040	0.149
intensity of the height profile	-0.070	-0.042	track length	0.520	0.232
soil temperature	0.010	0.054	complexity of the field geometry	-0.140	0.084
soil moisture	0.010	0.107	machined area	0.460	0.252
daily rainfall	0.250	0.193			

When investigating efficiency changes, it should be noted that variables such as work speed can also have an enormous influence on the quality of work (Hannusch et al. 1984). Quality of work can also be a measure of the efficiency of a process, which is definition-dependent. It also depends on the nature of the process and the work objective. When sowing it is the even distribution and optimum quantity of seed possible measures of work quality. When ploughing, besides the area output, the total labour costs and the desired soil appearance are important measures of work quality. As explained in the section on materials and methods, this has not been dealt with in the context of the work, as there is no data on resources such as fuel and material spread or collected. If these data are also available in future studies, the definition of efficiency will need to be expanded. The target figure could then be, for example, the process costs per hectare. An optimisation function could be used to determine the ideal trade-off between work quality and time and costs.

Conclusions

The procedure has shown that a correction of efficiency values is possible with the help of public databases and Support Vector Regression. Individual parameters can also be examined for their optimization and savings potential. However, other parameters, such as fuel consumption and the quality of field work, which are also important efficiency variables, should be considered in subsequent investigations.

In future, it will be necessary to demonstrate how well the procedure is suited to comparing work processes of the same type on different farms. Furthermore, it must be investigated which processes are suitable for the analysis and where the weaknesses of the approach are located. More extensive data sets with additional heterogeneities could lead to an increased number of support vectors and could improve the model and reduce the error sizes. Previous work has provided corresponding experience in this respect (Stein and Meyer 2018). A use of the method as a sub-function for a decision support system is also being considered in the course of the project BiDaLAP.

References

Bortz, J.; Döring, N. (2006): Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler, mit 87 Tabellen, Heidelberg, Springer-Medizin-Verlag

Cherkassky, V.; Dhar, S. (2010): Simple Method for Interpretation of High-Dimensional Nonlinear SVM Classification Models. In: Ed. Stahlbock, R.; Crone, S.F.; Abou-Nasr, M.; Arabnia, H.R.; Kourentzes, N.; Lenca, P.; Lippe, W.M.; Weiss, G.M, Proceedings of The 2010 International Conference on Data Mining, DMIN 2010, July 12-15, 2010, Las Vegas, Nevada, USA. CSREA Press

Dodge, Y.; Jurečková, J. (2000): Adaptive Regression. New York, Springer New York

Dolan, H. (2019): A Practical Guide to Interpreting and Visualising Support Vector Machines. https://towardsdatascience.com/a-practical-guide-to-interpreting-and-visualising-support-vector-machines-97d2a5b0564e, accessed on 28 May 2019

Düwel, O.; Siebner, C.S.; Utermann, J.; Krone, F. (2007): Bodenarten der Böden Deutschlands. https://www.bgr.bund.de/DE/Themen/Boden/Produkte/Schriften/Downloads/Bodenarten_Bericht.pdf?__blob=publicationFile, accessed on 30 Oct 2018

Hannusch, L.; Lubadel, O.; Frießleben, R.; Koschitzke, E.; Jeske, A. (1984): Einfluß der Arbeitsgeschwindigkeit und der Bodenfreiheit von Mineraldüngerstreuern und Pflanzenschutzmaschinen auf den Ertrag. In: Agrartechnik. Landtechnische Zeitschrift der DDR. Hg. Kammer der Technik, Berlin, VES Verlag Technik, S. 479–484

KTBL (2018): KTBL-Feldarbeitsrechner. https://daten.ktbl.de/feldarbeit/entry.html#0, accessed on 5 Sept 2018

Kudaligama, V.P.; Yanagida, J.F. (2000): A Comparison of Intercountry Agricultural Production Functions: A Frontier Function Approach. Journal of Economic Development 25, pp. 57–74

Lal, R. (2004): Carbon emission from farm operations. Environment international 30(7), pp. 981–990, https://doi.org/10.1016/j.envint.2004.03.005

Pacini, C.; Wossink, A.; Giesen, G.; Vazzana, C.; Huirne, R. (2003): Evaluation of sustainability of organic, integrated and conventional farming systems: a farm and field-scale analysis. Agriculture, Ecosystems & Environment 95(1), pp. 273–288, https://doi.org/10.1016/S0167-8809(02)00091-9

Ross, S.M.; Heinisch, C. (2006): Statistik für Ingenieure und Naturwissenschaftler. München, Elsevier Spektrum Akad. Verlag

Stein, T.; Henschel, T. (2019): Potenziale von Open Data für die Effizienzsteigerung von mobilen Arbeitsmaschinen. In: 39. GIL-Jahrestagung: Digitalisierung für landwirtschaftliche Betriebe in kleinstrukturierten Regionen – ein Widerspruch in sich? 18.–19. Februar 2019 in Wien, Österreich. Hg. Meyer-Aurich, A., Gandorfer, M., Barta, N., Gronauer, A., Kantelhardt, J. & Floto, H., Bonn, Gesellschaft für Informatik e.V., S. 251–256

Stein, T.; Meyer, H.J. (2018): Automatic machine and implement identification of an agri-cultural process using machine learning to optimize farm management information systems. In: 6th International Conference on Machine Control and Guidance, 01.–02.10.2018, Berlin, Bornimer Agrartechnische Berichte, Heft 101, S. 19–26

Steinwart, I.; Christmann, A. (2008): Support vector machines. New York, Springer Verlag

Toll, C.v. (2013): Energieorientierte Analyse der Landmaschinentechnik. Untersuchung zur maschinenrelevanten Energiebilanzierung in der Getreideproduktion mit Erfassung von CO_2-Einsparpotenzialen. Dissertation, Technische Universität Berlin

Uppenkamp, N. (2018): Erfahrungssätze für Maschinenring-Arbeiten unter Landwirten. https://www.landwirtschaftskammer.de/landwirtschaft/beratung/pdf/erfahrungssaetze-rh.pdf, accessed on 5 June 2019

Author

Thoralf Stein M.Sc. vehicle engineering, is a research assistant at the Technischen Universität Berlin, Straße des 17. Juni 144, 10623 Berlin. E-Mail thoralf.stein@tu-berlin.de

Acknowledgements

The project BiDaLAP is fully funded by the Bundesministerium für Landwirtschaft und Ernährung.