

Sonic Log Prediction Based on Extreme Gradient Boosting (XGBoost) Machine Learning Algorithm by Using Well Log Data

*Abdul Hakam*¹, *Widya Utama*^{2*}, *Sherly Ardhya Garini*³, *Omar Abdul Jabar*², *Alif Nurdien Fitrah Insani*², and *Yudi Rosandi*⁴

¹Department of Civil Engineering, Universitas Andalas, Padang, Indonesia

²Department of Geophysics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

³Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

⁴Department of Geophysics, Universitas Padjadjaran, Sumedang, Indonesia

Abstract. Sonic log is an important aspect that provides a detailed description of the subsurface properties associated with oil and gas reservoirs. The problem that frequently occurs is the unavailability of sonic log data for various reasons needs to be given an effective solution. The alternative approach proposed in this research is sonic log prediction based on Extreme Gradient Boosting (XGBoost) machine learning algorithm, using available log data to build a reliable sonic log prediction model. In this research, the predicted DT log type is the Differential Time Shear Slowness (DTSM) log, which is the velocity of shear waves propagating in a formation. Log features used for training include gamma ray (GR), density (RHOB), porosity (NPHI), resistivity (RS and RD) logs with DTSM log as the prediction target. To optimise the performance and generalisation of the XGBoost algorithm in predicting log DTSM, hyperparameter tuning was applied using grid search technique to obtain optimal parameters for the prediction model. Based on the experimental results, this research found that hyperparameter tuning using grid search technique improved the accuracy of sonic log (DTSM) model prediction based on XGBoost algorithm, as proven by the decrease of RMSE and MAPE values to 19.699 and 7.713%. The results also pointed out the need for methods other than listwise deletion to handle missing values as an alternative to improving model accuracy. This research highlighted the need for continuous improvement in data processing methods and algorithm optimization

* Corresponding author: widya@geofisika.its.ac.id

to advance the application of machine learning in geophysical exploration.

1 Introduction

Sonic log along with other petrophysical logs such as GR, NPHI, RHOB, etc., are used to evaluate lithology, reservoir properties, hydrocarbon carrier properties, and other important parameters in hydrocarbon resources. Sonic log also contain important information about the formation, which is necessary for oil and gas exploration and production activities [1]. The application of machine learning in the geophysical domain has experienced a significant improvement in recent years. Several researches have showed the potential of machine learning algorithms in accurately predicting log values in various geoscience and petroleum engineering applications.

Artificial neural networks have also been shown to predict sonic logs compressional and shear wave along a wellbore from a producing well [1]. The application of artificial neural network models to predict missing log data [2], the research and experiment conducted in that study aimed to predict petrophysical properties of the Mishrif Formation in the Nasiriya Oilfield using artificial neural networks (ANN). The ANN models were trained using well logs data from surrounding wells, and the missing logs data of the targeted Ns-X Well were generated. The results showed good correlation between the original and predicted logs, with correlation coefficients ranging from 0.575 to 0.914. However, the Inductive Log Depth (ILD) log prediction had the lowest correlation of 0.575. It was also observed that lithological variation within the formation could affect the accuracy of the ANN models. The study concluded that ANN could provide accurate predictions in less heterogeneous formations and recommended further research to validate the predicted logs with cased log holes. Besides that application of machine learning with artificial neural networks for shear log prediction in the Volve field, Norwegian North Sea [3]. The research focused on using machine learning with artificial neural networks (ANN) to predict shear logs in the Volve field in the Norwegian North Sea. The study used a total of 50,885 observation data points from six wells for training and testing the ANN model. The feature selection process showed a statistically significant relationship between each input log and the shear log. The ANN model demonstrated promising results with a Coefficient of Determination (R^2) between 0.84 and 0.97 and a Root Mean Square Error (RMSE) between 26.68 and 119.21. This research provides a tested machine learning (ML) approach for synthesizing shear logs on a full field scale data set, which can be valuable for reservoir characterization applications and geomechanical parameters calculation [2].

One of the powerful machine learning algorithms known for its ability to efficiently handle missing data and weak structure prediction models to build more accurate models is the Extreme Gradient Boosting (XGBoost) algorithm [4]. XGBoost is a boosting algorithm based on the evolution of the Gradient Boosting Decision Tree (GBDT) algorithm, which has achieved remarkable results in practical applications due to its accuracy, high speed, and unique information processing scheme [5]. The study by Chen et al the found that XGBoost has outstanding prediction capabilities, especially for prediction with tabular data, and its ability to handle missing values is considered a significant advantage. In conclusion, XGBoost is better at handling problems with missing values and large training data than commonly used methods, such as random forests and networks [6]. Based on this, this study will predict the log Differential Time Shear Slowness (DTSM) value using Extreme Gradient Boosting (XGBoost) machine learning algorithm with hyperparameter tuning using grid search technique. The results of this research are expected to be the basis for the application and development of machine learning in the field of Geophysics.

2 Research Method

The data used is open access data of Northeast Australia region owned by ConocoPhillips which can be accessed at the following URL link <https://www.occam.com.au/poseidondata>. The data used totalled 7 wells, i.e. Kronos 1, Pharos 1, Poseidon 1, Poseidon 2, Poseidon North 1, Proteus 1, and Torosa 1. This process started with the compilation of well log data such as gamma ray (GR), density (RHOB), porosity (NPHI), and resistivity (RS and RD). Then, removing rows of data where there are missing values through the listwise deletion technique [7]. After that, a feature scaling step was applied to normalise the values in the dataset, thus ensuring each feature contributes equally to the classification model. Next, the dataset was divided into two parts: 80% was used as training data and 20% as test data. In this research, the XGBoost algorithm is used to train the model, XGBoost is known for efficiency and effectiveness in prediction applications. To optimise the model, grid search technique is applied. This technique systematically searches through combinations of predefined hyperparameters, using cross validation to evaluate each combination and determine the one that gives the best performance. Figure 1 shows workflow of this research which provide step-by-step data processing, model development, and result of the research.

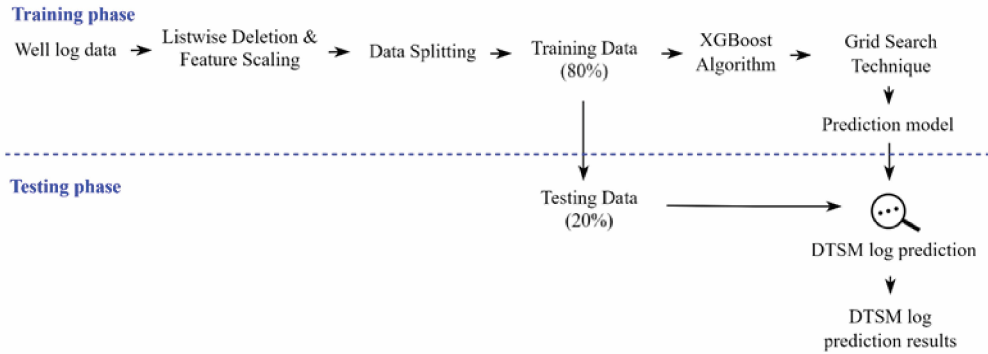


Fig. 1. Research workflow.

After the model was trained, testing phase was carried out using the pre-separated test data, which was also used to assess how reliably the model predicted the log DTSM based on the previous features in training. The trained prediction model then evaluated the performance using Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) evaluation metrics, using the following equations:

$$MAPE = \sum_{i=1}^n \left| \frac{f_i - a_i}{a_i} \right| \times 100\% \quad (1)$$

Where n is the number of data, f_i is the predicted value at the i -th data, a_i is the actual data at the i -th data. Through the calculation of MAPE, the error value can be interpreted easily. Then, RMSE is used as an additional error calculation so that the error calculation can be validated by other calculations. This is formulated as follows:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

The variable n is the number of data, $\hat{y}_i$ is the predicted value, and y_i is the true /actual value. The RMSE value obtained is the average error value generated by the predictive model [8].

3 Research Results

Based on the construction of the DTSM log prediction model using the XGBoost algorithm involving hyperparameter tuning using grid search technique, parameter optimization was obtained as shown in Table 1.

Table 1. Default parameters and hyperparameters used in the research.

No	Parameter	Default Parameters	Hyperparameters
1	n_estimators	100	75
2	max_depth	6	9
3	learning_rate	0.3	0.1
4	reg_lambda	1	1
5	Gamma	-0.214	None
6	colsample_bytree	0.1	0.8

In the default parameters, the XGBoost model is configured with 100 'n_estimators', which is the number of trees used in boosting. The maximum depth of each tree ('max_depth') is set at 6, and the 'learning_rate', which controls how fast the model learns, is 0.3. The parameter 'reg_lambda' as a regularization parameter on the model weights, is set at 1. 'Gamma', which determines the minimum loss reduction required to make further divisions at a tree node, is set at -0.214, which may indicate the use of default values. 'colsample_bytree' which specifies the proportion of features used for each tree is 0.1, indicating only a fraction of the features are used, possibly to avoid overfitting.

In machine learning algorithms, hyperparameters have a significant influence on model performance and generalisation [9]. To achieve optimal model performance, it is crucial to choose the right hyperparameters [10], because hyperparameters manage various aspects of the model class and learning procedure, and their optimisation is an important but difficult part of the machine learning pipeline [11], [12]. Hyperparameters significantly affect the performance of machine learning algorithms, and their adjustment can result in significant model optimisation improvements [13]. Moreover, it has been proven that the influence of hyperparameters on model performance can be found in various fields, such as image classification [14], stock price forecasting [15], and deep learning [16].

Table 2. Model Evaluation on default parameters and hyperparameters.

No	Parameters	RMSE	MAPE (%)
1	Default parameters	19.699	7.713
2	Hyperparameters	19.028	7.425

Table 2 are the results of a comparative evaluation of two predictive model configurations: one using default parameters and the other using a set of hyperparameters. The effectiveness of each model configuration is evaluated using two metrics: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), both of which are expressed in percentage. For the model running with the default parameters, the RMSE was 19.699%, which indicates the standard deviation of the prediction error, essentially measuring how far the predicted value differs from the actual value on average. For this configuration, the MAPE was 7.713%, which indicates the accuracy of the prediction in terms of the percentage error compared to the actual value. Both metrics show a slight improvement when the model is adjusted with hyperparameters. The RMSE dropped to 19.028%, indicating that the average prediction was closer to the actual value than the model with the

default settings. Similarly, the MAPE dropped to 7.425%, indicating slightly more accurate predictions in percentage terms than the model with the default parameter settings, it can be seen in Figure 2.

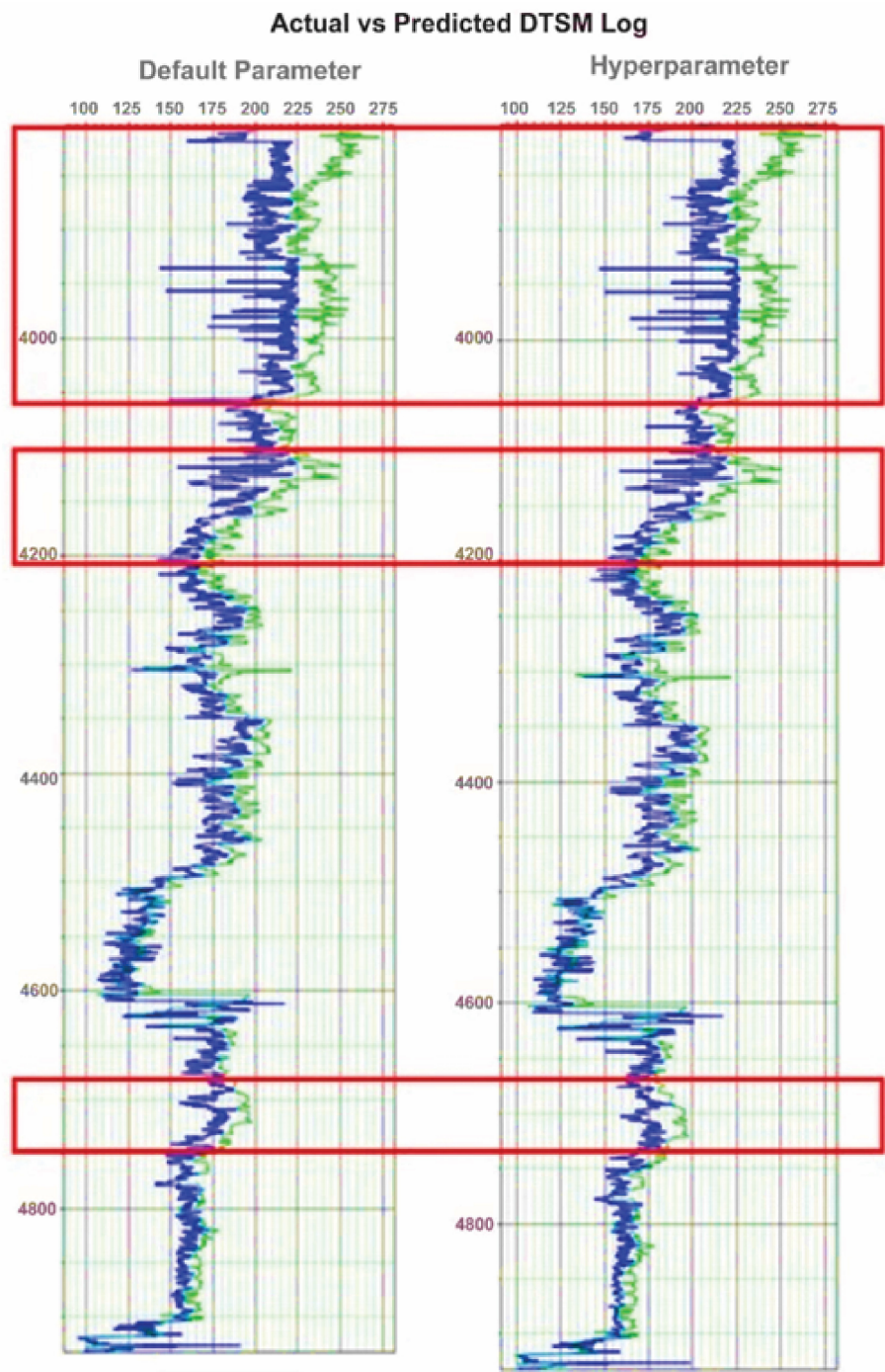


Fig. 2. Predicted and actual DTSM logs (green line represents actual DTSM logs and blue line represents predicted results using XGBoost algorithm).

4 Discussion

Based on the experimental results, it can be seen in Table 2, that the set of tuned hyperparameters: 'n_estimator' was reduced to 75, which was an attempt to reduce computation time or avoid overfitting. 'max_depth' was increased to 9, which allows the model to capture more complex interactions in the data, but also increases the risk of overfitting. 'learning_rate' was reduced to 0.1, which may make the model slower in adjusting the weights but may result in a more generalised model. 'reg_lambda' remains at 1, indicating that the degree of regularization has not changed. 'Gamma' is changed to 'None', which could mean that this parameter is not set or return to the XGBoost default value. 'colsample_bytree' increased significantly to 0.8, indicating that now most of the features are used to create each tree, this may improve model performance if the features are relevant and useful. Overall, the changes in hyperparameters indicate an attempt to adjust the model to better fit the specific dataset at hand, balancing between bias and variance and potential overfitting. The experimental results in Figure 2 show an improvement when the model is adjusted with hyperparameters using the grid search technique. The evaluation matrix value changes, the RMSE value and MAPE value decrease to 7.425% and 19.028%, which indicates that the average prediction is closer to the actual value compared to the model using the default parameter settings that can be said that the prediction performance has improved to be more accurate. Overall, this is in line with some previous research, that understanding and optimizing the parameters in the application of machine learning algorithms is crucial to achieve the best performance in various applications [13], [17], [18].

In Figure 2, which shows the results of the DTSM log prediction, , qualitatively at some depths there are still differences with the actual data. Such as at depths of 4552 - 4710 meters, 4725 - 4740 meters, 4755 - 4775 meters, and 4960 - 4975 meters. This is due to the lack of quantity and quality of data so that the accuracy of the algorithm is not well enough to predict DTSM logs at a particular depth. The quality of the data used has decreased due to the use of the listwise deletion technique which is one of the processes in the preprocessing phase for handling missing values. Listwise deletion is a method to handle missing data by deleting all rows of records that have missing values. Although this approach is simple, straightforward, and is used in almost 97% of studies [7], but listwise deletion can lead to a reduced sample size, potentially causing a loss of statistical power and biased parameter estimates [19]. According to some studies this method may not be the most efficient method to deal with missing data, as it discards valuable information and can produce mathematically inconsistent metrics [19], [20]. In this research, the large amount of missing data means that a lot of data has to be deleted making the dataset smaller, which can significantly affect the results of the DTSM log prediction model. This smaller dataset constrains the ability of the model to learn effectively, especially if the original dataset is not large. In addition, listwise deletion may introduce bias into the model, especially if the missing data is not random. Therefore, the model may not be able to accurately represent the formation at some wellbore depths, resulting in less accurate predictions. It is important to carefully consider and find alternative methods, e.g., imputation techniques for optimization to handle these cases of missing values. Overall, there is a significant potential contribution of machine learning to the sonic log prediction (DTSM) problem, which is represented by the experimental results that show good performance with a MAPE value of 7.713% and an RMSE of 19.699.

5 Conclusion

The results of this research show significant impact of hyperparameter tuning using the grid search technique on the performance of machine learning models in sonic log prediction (DTSM) is evident. The parameter tuning has resulted in better accuracy, as shown by lower

RMSE and MAPE values of 7.713% and RMSE value of 19.699. Limitations in data quality and quantity affect the training of machine learning algorithms, which is possibly one of the reasons for the inaccuracy of DTSM log prediction results, especially at particular depths in the well bore. In addition, the results highlight the need of alternative missing values handling methods to improve model accuracy. The use of listwise deletion techniques to handle missing data, showed drawbacks such as reduced sample size and potential bias, which adversely affected the limitations in data quality and quantity. This affects the training of machine learning algorithms, which is possibly one of the reasons for the inaccuracy of DTSM log prediction results, especially at certain depths in the well bore. Overall, the results of this research not only showed the potential and significant contribution of machine learning in DTSM log prediction, but also highlighted the need of continuous methodological advancements in data processing and algorithm optimization for future geophysical applications.

Acknowledgement

The authors express their deepest gratitude for the financial support from the Indonesian Institute for Community Service Research and Collaborative Research (Scheme A) – Andalas University, with contract number T/30/UN.16.17/PT.01.03/IS-RKI Scheme A (Partner)/2022, May 24, 2022. The authors also thank all reviewers and editors who have provided criticism and suggestions for the development of this article so that it can finally be published.

References

1. D. Onalo, S. Adedigba, O. Oloruntobi, F. Khan, L. A. James, and S. Butt, “Data-driven model for shear wave transit time prediction for formation evaluation,” *J. Pet. Explor. Prod. Technol.*, vol. 10, no. 4, pp. 1429–1447, 2020, doi: 10.1007/s13202-020-00843-2.
2. A. K. A. Mohammed and M. K. Dhaidan, “Prediction of Well Logs Data and Estimation of Petrophysical Parameters of Mishrif Formation, Nasiriya Field, South of Iraq Using Artificial Neural Network (ANN),” *Iraqi J. Sci.*, vol. 64, no. 1, pp. 253–268, 2023, doi: 10.24996/ijs.2023.64.1.24.
3. A. Al Ghaithi and M. Prasad, “Machine learning with artificial neural networks for shear log predictions in the volve field norwegian north sea,” *SEG Tech. Progr. Expand. Abstr.*, vol. 2020-Octob, pp. 450–454, 2020, doi: 10.1190/segam2020-3427540.1.
4. N. Hou et al., “Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost,” *J. Transl. Med.*, vol. 18, no. 1, pp. 1–14, 2020, doi: 10.1186/s12967-020-02620-5.
5. C. X. Lv, S. Y. An, B. J. Qiao, and W. Wu, “Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model,” *BMC Infect. Dis.*, vol. 21, no. 1, pp. 1–13, 2021, doi: 10.1186/s12879-021-06503-y.
6. K. Kawano et al., “Prediction of mortality risk of health checkup participants using machine learning-based models: the J-SHC study,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–8, 2022, doi: 10.1038/s41598-022-18276-8.
7. Y. Dong and C. Y. J. Peng, “Principled missing data methods for researchers (Expectation Maximaization explained),” *Springerplus*, vol. 2, no. 1, pp. 1–17, 2013.
8. A. Kumar Dubey, A. Kumar, V. García-Díaz, A. Kumar Sharma, and K. Kanhaiya, “Study and analysis of SARIMA and LSTM in forecasting time series data,” *Sustain. Energy Technol. Assessments*, vol. 47, no. August 2020, p. 101474, 2021, doi:

- 10.1016/j.seta.2021.101474.
9. A. Li et al., “A generalized framework for population based training,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1791–1799, 2019, doi: 10.1145/3292500.3330649.
 10. A. Raghu, J. Lorraine, S. Kornblith, M. McDermott, and D. Duvenaud, “Meta-Learning to Improve Pre-Training,” *Adv. Neural Inf. Process. Syst.*, vol. 28, no. NeurIPS, pp. 23231–23244, 2021.
 11. S. Seo and K. Obermayer, “Dynamic hyperparameter scaling method for LVQ algorithms,” *IEEE Int. Conf. Neural Networks - Conf. Proc.*, pp. 3196–3203, 2006, doi: 10.1109/ijcnn.2006.247304.
 12. M. Khodak et al., “Federated Hyperparameter Tuning: Challenges, Baselines, and Connections to Weight-Sharing,” *Adv. Neural Inf. Process. Syst.*, vol. 23, no. NeurIPS, pp. 19184–19197, 2021.
 13. J. N. Van Rijn and F. Hutter, “Hyperparameter importance across datasets,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2367–2376, 2018, doi: 10.1145/3219819.3220058.
 14. K. Uyar, Ş. Taşdemir, and I. A. Özkan, “The analysis and optimization of CNN Hyperparameters with fuzzy tree model for image classification,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 30, no. 3, pp. 961–977, 2022, doi: 10.3906/ELK-2107-130.
 15. K. E. Hoque and H. Aljamaan, “Impact of hyperparameter tuning on machine learning models in stock price forecasting,” *IEEE Access*, vol. 9, pp. 163815–163830, 2021, doi: 10.1109/ACCESS.2021.3134138.
 16. D. Jia, R. Wang, C. Xu, and Z. Yu, “QIM: Quantifying hyperparameter importance for deep learning,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9966 LNCS, pp. 180–188, 2016, doi: 10.1007/978-3-319-47099-3_15.
 17. Y. Mai, Z. Sheng, H. Shi, and Q. Liao, “Using Improved XGBoost Algorithm to Obtain Modified Atmospheric Refractive Index,” *Int. J. Antennas Propag.*, vol. 2021, 2021, doi: 10.1155/2021/5506599.
 18. Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu, and M. Tu, “Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances,” *J. Pet. Sci. Eng.*, vol. 160, no. October 2017, pp. 182–193, 2018, doi: 10.1016/j.petrol.2017.10.028.
 19. J. L. Peugh and C. K. Enders, “Missing data in educational research: A review of reporting practices and suggestions for improvement,” *Rev. Educ. Res.*, vol. 74, no. 4, pp. 525–556, 2004, doi: 10.3102/00346543074004525.
 20. V. Arel-Bundock and K. J. Pelc, “When Can Multiple Imputation Improve Regression Estimates?,” *Polit. Anal.*, vol. 26, no. 2, pp. 240–245, 2018, doi: 10.1017/pan.2017.43.