JSLHU
JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY

ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2025**, 22, 047-051

PHÂN LOẠI DỮ LIỆU MẤT CÂN BẰNG VỚI RỪNG NGẪU NHIÊN KẾT HỢP PHÂN CỤM WARD

Võ Thị Ngọc Hà[1], Nguyễn Thanh Sơn[2], Đặng Đăng Khoa[2], Lê Phương Long[2], Phan Thị Thu Ngân[3*]
[1]Trường Đại học Quang Trung, tỉnh Gia Lai, Việt Nam
[2]Trường Đại học Lạc Hồng, tỉnh Đồng Nai, Việt Nam
[3]Trường Đại học Quốc tế Hồng Bàng, Thành phố Hồ Chí Minh, Việt Nam
* Tác giả liên hệ: *p.thungan87@gmail.com*

THÔNG TIN BÀI BÁO

Ngày nhận:	20/4/2025
Ngày hoàn thiện:	29/5/2025
Ngày chấp nhận:	29/5/2025
Ngày đăng:	15/9/2025

TỪ KHÓA

Imbalanced Data;
Random Forest Algorithm;
Balanced Random Forest;
Customer Churn;
Classification Technique.

TÓM TẮT

Nghiên cứu này giới thiệu thuật toán Modified Balanced Random Forest (MBRF) nhằm cải thiện hiệu suất phân loại trên tập dữ liệu mất cân bằng. Phương pháp đề xuất nâng cao Rừng Ngẫu Nhiên Cân Bằng bằng cách áp dụng chiến lược lấy mẫu giảm dựa trên kỹ thuật phân cụm trong từng vòng lặp khởi tạo dữ liệu. Bốn phương pháp phân cụm được đánh giá gồm K Means, Clustering, Agglomerative Clustering, and Ward Hierarchical Clustering Trong đó, kỹ thuật Ward Hierarchical Clustering cho kết quả tối ưu nhất. Kết quả thực nghiệm cho thấy phương pháp đề xuất vượt trội hơn so với Random Forest (RF) and Balanced Random Forest (BRF) truyền thống, với tỷ lệ phát hiện dương tính đúng đạt 93,42%, tỷ lệ phát hiện âm tính đúng đạt 93,60% và độ chính xác theo diện tích dưới đường cong ROC đạt 93,51%, đồng thời rút ngắn thời gian xử lý. Những kết quả này khẳng định hiệu quả của phương pháp đề xuất trong bài toán phân loại dữ liệu mất cân bằng.

IMBALANCED DATA CLASSIFICATION USING RANDOM FOREST WITH WARD CLUSTERING

Vo Thi Ngoc Ha[1], Nguyen Thanh Son[2], Dang Dang Khoa[2], Le Phuong Long[2], Phan Thi Thu Ngan[3*]
[1]Quang Trung University, Gia Lai Province, Vietnam
[2]Lac Hong University, Dong Nai Province, Vietnam
[3]Hong Bang International University, Ho Chi Minh City, Vietnam
*Corresponding Author: *p.thhungan87@gmail.com*

ARTICLE INFO

Received:	Apr 20th, 2025
Revised:	May 29th, 2025
Accepted:	May 29th, 2025
Published:	Sep 15th, 2025

KEYWORDS

Imbalanced Data;
Random Forest; Balanced Random Forest;
Classification Technique.

ABSTRACT

This study introduces a Modified Balanced Random Forest algorithm to improve classification performance on imbalanced datasets. The proposed method enhances the Balanced Random Forest by applying a clustering based under sampling strategy during each bootstrap iteration. Four clustering methods were evaluated including K Means, Spectral Clustering, Agglomerative Clustering, and Ward Hierarchical Clustering. Among these, the Ward Hierarchical Clustering technique achieved the best performance. Experimental results show that the proposed method outperforms standard Random Forest and Balanced Random Forest, reaching a true positive rate of 93.42 percent, a true negative rate of 93.60 percent, and an area under the curve accuracy of 93.51 percent, while also reducing processing time. These results confirm the effectiveness of the proposed approach for imbalanced data classification.

Doi: https://doi.org/10.61591/jslhu.22.726
Available online at: https://js.lhu.edu.vn/index.php/lachong

1. INTRODUCTION

Random Forest is a widely used machine learning algorithm recognized for its strong classification performance when compared to other traditional classification techniques [1]. Its ease of implementation and robust predictive capabilities have made it a popular choice across diverse application domains [2], [3]. Empirical studies have demonstrated that Random Forest outperforms various other algorithms such as K-Nearest Neighbors (KNN), Naïve Bayes, C4.5, AdaBoost, and Artificial Neural Networks (ANN) [4]. However, one of the major challenges in applying Random Forest is its reduced effectiveness when dealing with imbalanced datasets [5]. The presence of significant class distribution disparity can hinder the learning process, as standard classification algorithms often assume uniform data distribution and equal misclassification costs [6], [7].

To address this limitation, various methods have been proposed in the literature. For instance, Wu et al. [8] explored the use of Random Forest for classifying insurance data characterized by class imbalance, incorporating an undersampling approach based on the KNN algorithm to enhance learning efficiency. Similarly, Khalilia et al. [9] employed random subsampling to manage class imbalance in a medical dataset for disease risk prediction. Their study showed that balancing the training data improved Random Forest's performance, surpassing that of Support Vector Machine (SVM) classifiers. Other researchers have explored hybrid techniques, such as the combination of undersampling with the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in a weighted Random Forest setting [10]. Additionally, the integration of RUSBoost with Information Gain has been applied as a preprocessing step for churn prediction on imbalanced datasets [11].

However, most existing approaches perform data balancing as a preprocessing step, thereby limiting the ability to observe its direct impact on the internal training dynamics of the Random Forest algorithm. To address this, the Balanced Random Forest (BRF) algorithm integrates the undersampling process directly into the tree-building phase of the ensemble, treating imbalance within the learning model itself [12]. BRF achieves this by applying undersampling to the majority class at each iteration of tree construction. Despite its advantages, the random undersampling strategy in BRF may discard valuable majority class samples, potentially degrading model performance.

In response to these limitations, this study proposes a Modified Balanced Random Forest (MBRF) algorithm that seeks to improve classification performance and computational efficiency. The MBRF enhances the BRF framework by replacing random undersampling with a more structured approach based on clustering techniques. Specifically, the training data is segmented into clusters, and samples are drawn in proportion to the number of minority class instances. To determine the most effective clustering strategy, we evaluate four clustering algorithms: K-Means, Spectral Clustering, Agglomerative Clustering, and Ward Hierarchical Clustering [13] – [18]. These algorithms define the number of clusters as a parameter, which in the MBRF framework is set according to the number of minority class instances. This integration aims to preserve informative samples while maintaining class balance, ultimately leading to a more accurate and efficient classification process.

2. METHOD

In this study, we propose the Modified Balanced Random Forest (MBRF) algorithm (Fig. 1) to enhance the predictive performance of traditional Random Forest (RF) and Balanced Random Forest (BRF) algorithms. Our approach introduces a key modification to the BRF process by incorporating clustering algorithms into the undersampling phase. The objective is to address the limitations of conventional undersampling, particularly the loss of potentially informative majority-class instances.

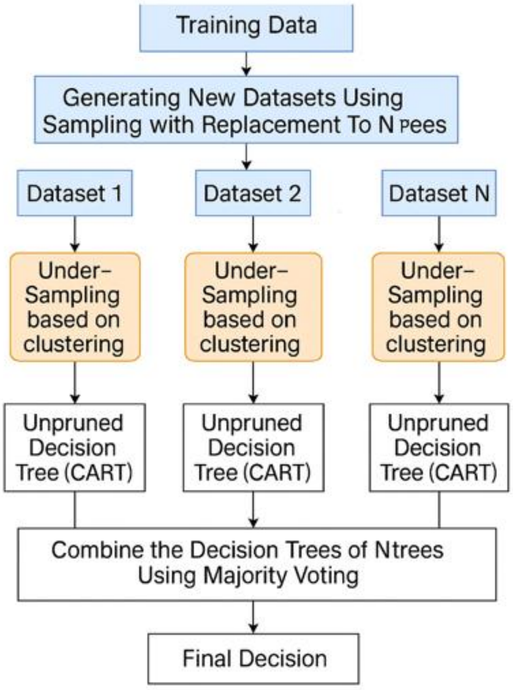
Training Data
Generating New Datasets Using Sampling with Replacement To N Pees
Dataset 1
Dataset 2
Dataset N
Under-Sampling based on clustering
Under-Sampling based on clustering
Under-Sampling based on clustering
Unpruned Decision Tree (CART)
Unpruned Decision Tree (CART)
Unpruned Decision Tree (CART)
Combine the Decision Trees of Ntrees Using Majority Voting
Final Decision

Figure 1. *MBRF flowchart*

The method begins with the training dataset *D*, which is partitioned into *N* subsets corresponding to the number of decision trees (D1, D2, …, DN). Each subset Di contains features Xi and class labels Yi. For each Di, the BRF process is applied by splitting the data into two bootstrap samples: Di+, Di−, representing the minority ("churn") and majority ("non-churn") classes, respectively. To ensure balanced training for each tree, an equal number of instances is drawn from both classes. Unlike the conventional BRF, in our proposed MBRF, the majority class is not randomly undersampled. Instead, we apply a clustering technique to group similar instances, and the resulting cluster centroids are selected to represent the majority class. The number of clusters is set to match the number of minority class instances, thereby preserving class balance while minimizing information loss.

The BRF algorithm is known for its hybrid approach that combines undersampling with ensemble learning strategies [19]. However, the random undersampling used in BRF may discard valuable information from the majority class, which can negatively impact model performance [11]. Our proposed clustering-based undersampling method mitigates this issue by ensuring that representative samples from the entire feature space are retained. This allows each tree to be trained on data that better reflects the overall data distribution, improving the model's generalization capability.

Furthermore, to minimize correlation between individual trees, a random sampling process is used when selecting bootstrap samples for each tree. The total number of bootstrap samples corresponds to the number of trees in the forest, ensuring ensemble diversity. The overall framework and architecture of the proposed MBRF approach are illustrated in Fig. 2.

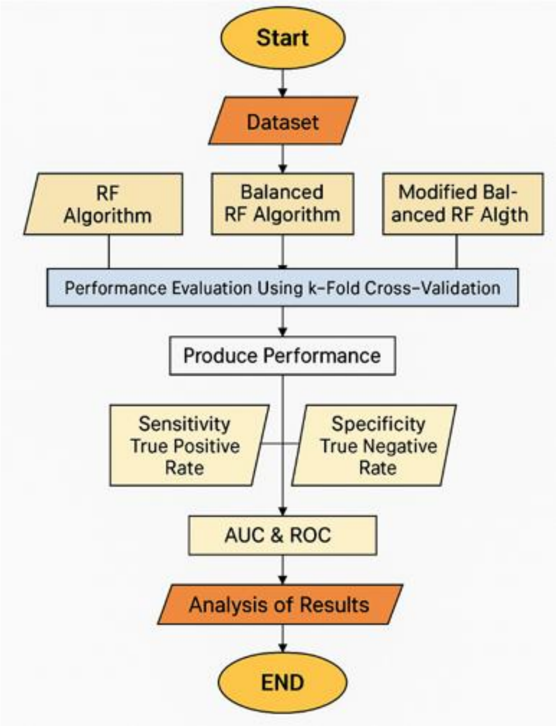
Start
Dataset
RF Algorithm
Balanced RF Algorithm
Modified Balanced RF Algth
Performance Evaluation Using k-Fold Cross-Validation
Produce Performance
Sensitivity True Positive Rate
Specificity True Negative Rate
AUC & ROC
Analysis of Results
END

Figure 2. *Research Design Flowchart*

2.1 Data set

This study employs a customer churn dataset provided by PT Telkom Indonesia [23].

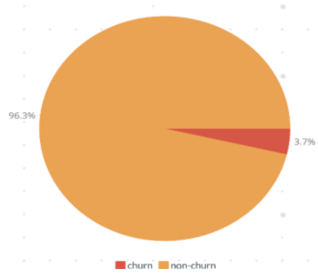

Figure 3. *Distribution of churn and non-churn*

The dataset comprises a total of 200,387 records, of which 192,863 correspond to non-churn instances and 7,524 represent churn cases. This results in a churn rate of approximately 3.75%, indicating a significant class imbalance. The dataset contains 52 attributes describing various customer features (Fig. 3).

2.2 Evaluation Measure

To evaluate the performance of the proposed classification models, we utilized a combination of sensitivity, specificity, the Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC). These metrics are commonly employed in previous studies to assess classifiers in the context of churn prediction is to identify churners, sensitivity is a particularly critical metric [24].

In addition to these, we incorporated the G-means metric, which combines the geometric mean of sensitivity and specificity. This metric is especially useful for imbalanced datasets as it provides a more balanced view of performance across both classes, in contrast to overall accuracy, which tends to be biased towards the majority class [23]. High accuracy may be misleading if the model predominantly favors the majority class while failing to detect minority class instances [25]. G-means mitigates this issue by penalizing models that fail to balance performance across classes [26].

Furthermore, the ROC curve was used to visualize the trade-off between true positive and false positive rates across different thresholds [27]. The AUC provides a scalar value summarizing the model's overall performance; values closer to 1.0 indicate better predictive capability [23].

Prior to evaluating the models, we computed the fundamental classification statistics: true positives (TP), true negatives (TN), false positives (FP), cases where a non-churn instance is incorrectly predicted as churn and false negatives (FN), cases where a churn instance is misclassified as non-churn [28], [29].

3. RESULTS AND DISCUSSION

This study introduces a MBRF approach as a strategy to address the challenges posed by imbalanced datasets. To assess the effectiveness of the proposed method, we conducted comparative experiments against the conventional RF and BRF algorithms. To determine the optimal model configuration, we tested ten different values for the number of decision trees: {5, 10, 15, 20, 25, 50, 60, 70, 90, 100}. Among these, a total of 10 trees yielded the best performance. However, the differences in predictive performance across the range of tested tree counts were not statistically significant. This observation is consistent with previous findings suggesting that RF are relatively insensitive to the number of trees in terms of accuracy. Nonetheless, increasing the number of trees does have implications for computational efficiency, as more trees require longer training and prediction times. To mitigate overfitting and ensure robust evaluation, we employed 10-fold cross-validation. This technique divides the dataset into 10 parts, uses nine parts for training, and one for testing, iteratively cycling through all combinations. This procedure has been widely accepted as a reliable model validation approach [30]. To evaluate the impact of different clustering techniques on the MBRF's performance, we examined four clustering methods under identical conditions (10 trees and

10-fold cross-validation) applied to the majority class during tree construction. Ward's Hierarchical Clustering demonstrated superior performance compared to the other clustering methods. Notably, it outperformed Agglomerative Clustering, although the difference was not statistically significant. Both methods showed scalability and efficiency in handling large datasets and large numbers of clusters. In contrast, K-Means Clustering, despite its ability to handle large datasets, is limited by the need to predefine the number of clusters. Spectral Clustering, on the other hand, yielded the lowest performance, likely due to its inefficiency when applied to large datasets with numerous clusters. Based on these findings, Ward's Hierarchical Clustering was selected for use in the MBRF model.

Table 1. *Experiment results*

Algorithms	***K-Means***	***Spectral Clustering***	***Agglomerative Clustering***	***Ward Hierarchical Clustering***
Sensitifity	89.54%	87.43%	91.72%	**93.42%**
Specificity	90.40%	87.94%	90.94%	**93.60%**
G-Means accuracy	89.96%	87.65%	91.33%	**93.49%**

Table 2 presents a comparative analysis of RF, BRF, and MBRF. To ensure fair evaluation, all models were trained and tested using 10-fold cross-validation. The MBRF achieved the best performance across all evaluated metrics. Specifically, the G-means score reached 0.9349 (93.49%), and the AUC attained 0.9351 (93.51%), both outperforming RF and BRF. Since churn prediction focuses on correctly identifying customers likely to churn (i.e., the positive class), a high sensitivity value is critical. Both BRF and MBRF demonstrated substantially higher sensitivity compared to RF, reflecting their ability to address class imbalance during tree construction. However, the MBRF also exhibited improved running time efficiency, suggesting that the modifications introduced not only enhanced predictive accuracy but also computational performance.

Table 2. *Experiment results on each method*

Evaluation	***RF***	***BRF***	***MBRF***
Sensitifity	57.20%	75.93%	93.42%
Specificity	99.12%	99.12%	93.60%
G-Means	75.16%	86.75%	93.49%
AUC	78.16%	87.52%	93.51%
Running time	435.5 sec	80.5 sec	57.8 sec

The comparative analysis of ROC curves, as illustrated in Fig. 4, further supports these findings. The MBRF curve approaches the ideal point of high True Positive Rate (TPR) and low False Positive Rate (FPR), signifying robust performance in both sensitivity and specificity. BRF, while showing better performance than RF, trails behind MBRF, reaffirming the advantages of data balancing. In contrast, RF's ROC curve lies closer to the diagonal, indicating weaker discrimination capability.

These results highlight the value of incorporating clustering-based undersampling in ensemble models such as the Random Forest. The MBRF approach effectively addresses the shortcomings of traditional undersampling by ensuring representative sampling and reducing loss of information from the majority class. As such, MBRF demonstrates significant improvements in both predictive accuracy and runtime efficiency, making it a strong candidate for churn prediction tasks and other applications involving highly imbalanced data distributions.

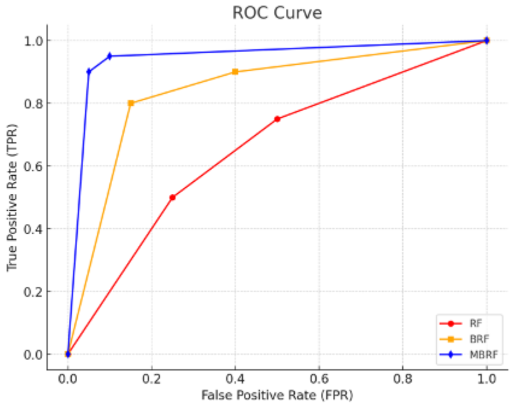

Figure 4. *Comparison of ROC between RF, BRF, and MBRF*

4. CONCLUSION

A novel approach, termed the Modified Balanced Random Forest (MBRF), has been developed to address classification challenges. This method achieves a balanced trade-off between precision in classifying majority and minority classes, yielding comparable sensitivity TP rate and specificity TN rate. Additionally, MBRF demonstrates enhanced computational efficiency by reducing processing time. The Random Forest parameters employed in this study exhibited low sensitivity, resulting in consistent model outcomes across iterations; however, these parameters influenced runtime due to increased tree-building duration. Notably, MBRF is less effective for small datasets, highlighting a limitation. Future research is recommended to refine the MBRF approach and address its constraints, particularly for small-sample applications.

5. REFERENCES

[1] S. Singh and P. Gupta, "Comparative study ID3, cart and C4 . 5 Decision tree algorithm: a survey," *Int. J. Adv. Inf. Sci. Technol.*, **2014**

[2] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 200

[3] H. Aydadenta and Adiwijaya, "A Clustering Approach for

Feature Selection in Microarray Data Classification Using Random Forest," *J. Inf. Process. Syst.*, vol. 14, no. 5, pp. 1167–1175, **2018**.

[4] G. Esteves and J. Mendes-Moreira, "Churn perdiction in the telecom business," in *2016 11th International Conference on Digital Information Management, ICDIM 2016*, **2016**

[5] A. Sonak and R. A. Patankar, "A Survey on Methods to Handle Imbalance Dataset," *Int. J. Comput. Sci Mob. Comput.*, vol. 4, no. 11, pp. 338–343, **2015**

[6] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, pp. 176-203, **2015**

[7] S. Du, F. Zhang, and X. Zhang, "Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach," *ISPRS J. Photogramm. Remote Sens.*, **2015**

[8] Z. Wu, W. Lin, Z. Zhang, A. Wen, and L. Lin, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," in *Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017*, **2017**

[9] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Mak.*, **2011**

[10] V. Effendy and Z. K. a. Baizal, "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest," *2014 2nd Int. Conf. Inf. Commun. Technol.*, **2014**

[11] E. Dwiyanti, Adiwijaya, and A. Ardiyanti, "Handling imbalanced data in churn prediction using RUSBoost and feature selection (Case study: PT. Telekomunikasi Indonesia regional 7)," in *Advances in Intelligent Systems and Computing*, **2017**

[12] Ł. Kobyliński and A. Przepiórkowski, "Definition extraction with balanced random forests," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **2008**

[13] S. Ghosh and S. Kumar, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, **2017**

[14] S. Venkateswara and V. Swamy, "A Survey: Spectral Clustering Applications and its Enhancements," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 185–189, **2015**.

[15] A. Y. Shelestov, "Using the agglomerative method of hierarchical clustering as a data mining tool in capital market," *Int. J. "Information Theor. Appl.*, vol. 15, no. 1, pp. 382–386, **2018**.

[16] K. Sasirekha and P. Baby, "Agglomerative Hierarchical Clustering Algorithm-A Review," *Int. J. Sci. Res. Publ.*, **2013**.

[17] W. Tian, Y. Zheng, R. Yang, S. Ji, and J. Wang, "A Survey on Clustering based Meteorological Data Mining," *Int. J. Grid Distrib. Comput.*, vol. 7, no. 6, pp. 229–240, **2014**.

[18] A. Chowdhary, "Community Detection: Hierarchical clustering Algorithms," *Int. J. Creat. Res. Thoughts*, vol. 5, no. 4, pp. 2320–2882, **2017**.

[19] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *Univ. California, Berkeley*, **2016**.

[20] D. Ramyachitra and P. Manikandan, "Imbalanced Dataset Classification and Solutions: a Review," *Int. J. Comput. Bus. Res.*, vol. 5, no. 4, **2018**

[21] S. Sardari, M. Eftekhari, and F. Afsari, "Hesitant fuzzy decision tree approach for highly imbalanced data classification," *Appl. Soft Comput. J.*, **2017**

[22] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Glob. J. Technol. Optim.*, **2018**.

[23] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27-38, **2013**.

[24] C. G. Weng and J. Poon, "A new evaluation measure for imbalanced datasets," *Proceedings of the 7th Australasian Data Mining Conference.*, vol. 87, no. 6, pp. 27-32, **2008**.

[25] J. S. Akosa, "Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data," *SAS Glob. Forum*, **2017**.

[26] Y. Zhang and D. Wang, "A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets," *Abstr. Appl. Anal.*, **2013**.

[27] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, **2006**.

[28] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, **2015**

[29] A. K. Santra and C. J. Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering," *IJCSI Int. J. Comput. Sci. Issues,* **2017**

[30] J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen, "Estimating the prediction performance of spatial models via spatial k-fold cross validation," *Int. J. Geogr. Inf. Sci.*, **2017**