MEDWIN PUBLISHERS
Committed to Create Value for Researchers

International Journal of Forensic Sciences
ISSN: 2573-1734

Accurate Prediction of Death Time via Integrating Microbial Community Dynamics

Sugich Prandini JA[1,#], Lu M[2,#], Jiang H[3] and An L[1,2,4*]

[1]Department of Biosystems Engineering, University of Arizona, USA

[2]Interdisciplinary Program in Statistics and Data Science, University of Arizona, USA

[3]Department of Statistics and Data Science, Northwestern University, USA

[4]Department of Epidemiology and Biostatistics, University of Arizona, USA

Research Article

Volume 9 Issue 1

Received Date: November 22, 2023

Published Date: February 12, 2024

DOI: 10.23880/ijfsc-16000353

***Corresponding author:** Lingling An, Department of Biosystems Engineering, University of Arizona, Shantz building 403, 1177 4th ST, Tucson, AZ, USA, Tel: (520)6211248; Email: anling@arizona.edu

[#]equally contributed to this article

Abstract

This study addresses the challenge of accurately estimating Postmortem Interval (PMI), the time since death, employing a data-driven approach. PMI determination is crucial in forensic investigations, and traditional methods often lack precision. We focus on utilizing a data mining approach Regularized Random Forest with cross-validation to enhance PMI prediction accuracy. Unlike conventional methods, our approach incorporates external information about the deceased, recognizing the impact of contextual factors on PMI estimation. Recent advancements have seen statistical methods leveraging dynamic changes in microbial communities to predict PMI. However, accuracy has been hindered by various sources of noise. To overcome this limitation, we propose a novel data mining approach, integrating cross-validation techniques and external information to refine PMI predictions. Through an empirical demonstration, we establish that our approach surpasses existing procedures in terms of accuracy, as validated against published datasets. This research contributes to the advancement of PMI estimation methodologies, emphasizing the importance of incorporating comprehensive data mining techniques and contextual information for more precise forensic applications.

Keywords: Metagenomics; Forensic Science; Postmortem Interval; Random Forest; Data Mining

Abbreviations: PMI: Post Mortem Interval; OTUs: Operational Taxonomic Units; RF: Random Forest; MSE: Mean Squared Error; MAD: Median Absolute Deviation; MAPE: Mean Absolute Percentage Error; RRMSE: Root Mean Squared Relative Error; SMAPE: Symmetric Mean Absolute Percentage Error; DTV: Total Variation Distance; AVGRE: Average Relative Error; MAXRE: Maximum Relative Error; RRF: Regularized Random Forest.

Introduction

Forensic science is the application of science to criminal and civil laws. Estimation of postmortem interval (PMI) is an important aspect of forensic sciences. However, the technique is inaccurate and incomplete [1], especially when physical evidence is used to estimate PMI [1]. Over past decades, entomological techniques have been widely

used in estimating the postmortem interval [2]. Common entomological and biological techniques extract DNA from insects such as flies to estimate the life stage of the insect and therefore the age of the corpse. A molecular method for estimating PMI was based on Black soldier flies who lay eggs at predictable times in the dry or post decay stages of decomposition of a corpse [3,4]. In 2002, another method was developed to estimate PMI by characterizing the chemistry associated with the decomposition of human remains to identify time-dependent biomarkers of decomposition [5]. However this method is limited because the correct body temperature must be known when the chemical data are collected. In 2003, quantification of mRNA degradation was found to be a possible estimator of PMI [6] but only give a rough interval.

Recent evidence suggests that bacterial changes occurring during decomposition could act as a 'microbial clock,' providing an additional method to estimate postmortem interval (PMI) and its significance in the forensic field [7-12]. In 2013 [7] tracked microbial changes of decomposing mice in a 48-day study. The microbial communities were sampled from 5 mice at various body sites and at eight time points for each body site. The statistical model, Random Forest (RF), is used to predict PMI based on microbial compositions at various operational taxonomic units (OTUs). Two years later, they conducted another experiment containing three soil types including desert, forest and grassland, around the University of Colorado [13]. They also recorded soil pH under the body at the abdominal, cecum, skin of torso, skin of head body sites. Mice decomposed for 71 days. Five mice were sampled from each soil type at each body site at 8 time points for each body site. They discovered diverse bacterial and fungal groups aiding nitrogen cycling and a consistent network of decomposers with predictable emergence patterns. Later, a comprehensive study was conducted on building Random Forest regression models for prediction of PMI by testing models using different sample types, gene markers, and taxonomic levels [14]. Their results demonstrate that the Random Forest method is very promising in PMI estimation using microbial information. Another study [15] explored the fundamental principles guiding the succession of microbial communities during the decomposition of pig carcasses and the underlying soil using the Random Forest model.

Most PMI studies using microbial data focus on exposed cadavers, but burial scenarios are common in forensics. Researchers used high-throughput sequencing to characterize microbial communities in burial rat cadavers [16] and predicted PMI with random forest models. Their results suggest postmortem microbial data as a potential tool for accurate PMI estimation in buried cadavers.

In this paper, we aim to improve the accuracy of PMI estimation based on microbial changes. We propose to use Regularized Random Forest (RRF) [17], an advanced machine learning technique, to improve the prediction accuracy for postmortem interval. Cross validation is then used to evaluate the model prediction. Furthermore, we also take a full advantage of other information, such as pH value of soil under the body and the body score which records the decomposition stages of a mouse. In the original papers this type of information was not considered in PMI prediction.

Methods

For the Illumina 16S rRNA sequence data, default settings in QIIME2 were used to pre-process the data to barcode 16S rRNA sequences and error-correcting codes were used to demultiplex and reduce the possibility of sample mis-assignment [18]. With a large number of OTUs or taxa, the objective of feature selection was to find an optimal subset of variables or features without losing the predictive information of the response variable which is PMI in our study [19]. Random Forest is a method which can measure feature importance, and can handle both quantitative and qualitative response variable. If Y is a discrete variable that contains C classes, RF will do classification; if Y is a continuous variable, RF will do regression to find the best model to predict Y. However, if there are a large number of predictor variables, many variables will be removed at each iteration and the valuable features with small importance scores could be removed [20]. The Regularized Random Forest is a new method based on RF to select features and do prediction. We propose to use RRF to predict PMI.

The algorithm of Regularized Random Forest is very similar to Random Forest. The main difference between RRF and RF is the Gini information gain value [20]. At node γ, let $\hat{p}_c^{\gamma}$ denotes the proportion of class-c observations at node γ, then the Gini index, Gini(γ) is defined as the following

$$Gini(\gamma) = \sum_{c=1}^{C} \hat{p}_c^{\gamma}\left(1-\hat{p}_c^{\gamma}\right)$$

Let X_i denote the ith predictor variable and Y the response variable. The Gini information gain value of X_i for splitting the node γ in RRF is as follows [20]

$$Gini_R(X_i,\gamma) = \begin{cases} \delta\, Gain(X_i,\gamma) & if\ i \notin F \\ Gain(X_i,\gamma) & if\ i \in F \end{cases}$$

Where

$$Gain(X_i,\gamma) = Gini(Xi,\gamma) - \omega_L\, Gini(X_i,\gamma^L) - \omega_R\, Gini(X_i,\gamma^R)$$

Where γ^L and γ^R are the left and right nodes of γ, ω_L and

ω_R are the proportions of instances assigned to the left and right child nodes, $\delta \in (0,1]$ is the penalty coefficient. A smaller δ turns out a larger penalty. F is the set of features used for splitting in previous nodes and is an empty set at the root node in the first tree. The importance score for variable X_i in the RRF can be calculated as

$$Imp_i = \frac{1}{ntree} \sum_{\gamma \in S_{Xi}} Gain_R(X_i, \gamma)$$

Where S_{X_i} is the set of nodes split by X_i in the RRF and ntree is the number of trees.

Results

Case Study 1 (Metcalf et al. 2016 Data)

The data are from the study of "Microbial community assembly and metabolic function during mammalian corpse decomposition" where mice decomposed for 71 days [13]. Five mice were sampled from each soil type at 5 body sites and 8 time points per location (dessert, forest, or grassland). All analyses were performed using the microbial relative abundances of the OTUs at the family level. After basic cleaning of the data (e.g., exclude the samples without true PMI values and/or samples with very low total number of sequence reads, i.e., <500), 15 datasets were generated for each of the twelve combinations of soil types and body sites. As soil pH value may be an important contributor to microbial community composition during decomposition [10] we added it as a predictor in our RRF model for PMI estimation. Similar to the original studies, we performed 10-fold cross validation and repeated it 10 times.

The accuracy of each model (RRF vs. RF) was evaluated by the average of absolute error, which is the mean of the absolute difference between the true value and the estimated value of PMI. Comparing with Random Forest approach, the Regularized Random Forest improves accuracy in general (Figure 1). Comparing RRF with RF, the average absolute error (i.e., the mean absolute difference between predicted time and true time) is reduced from 5.74 days to 4.98 days, i.e., about 13% reduction, for the desert samples; from 5.74 to 5.41 days (about 5% reduction) for the grassland samples; and from 4.89 days to 2.61 days (about 46.6% reduction) for the forest samples.

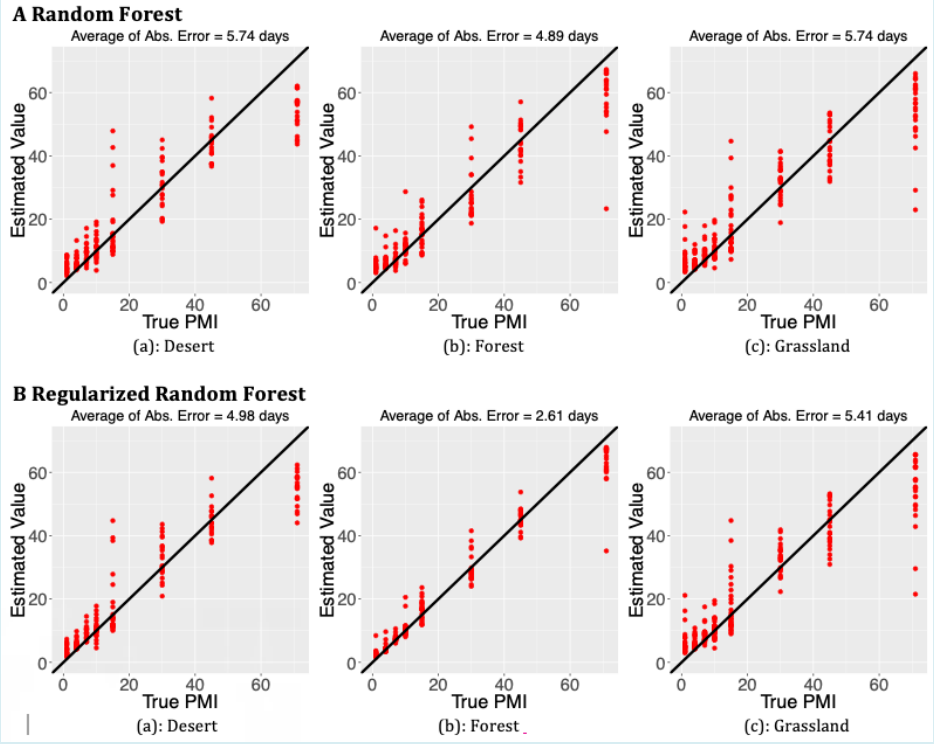

Figure 1: Figure 1. Scatter plot of estimated PMI vs. True PMI; the top plots (A) correspond to the estimated results using Random Forest (RF) and the bottom (B) for Regularized Random Forest (RRF) model. The 45 degree line is also displayed in each plot to show a comparison between the true and the predicted values. The average of absolute error between the predicted number of dates and true number for each situation is shown at the top of each plot.

Nine more metrics were calculated and summarized to compare the performance of RF and RRF (Figure 2), including Mean Squared Error (MSE), Median Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), Root Mean Squared Relative Error (RRMSE), Symmetric Mean Absolute Percentage Error (SMAPE), Total Variation Distance (DTV), Average relative Error (AVGRE), Maximum Relative Error (MAXRE), and Average of absolute error (MeanDiff) [21]. The lower the value of the metrics, the better the performance the method has. All performance measurements show that the RRF is more accurate at predicting PMI than RF, particularly, for the forest soil type (the middle case in each subplot). The details of the metrics can be found in the Appendix.

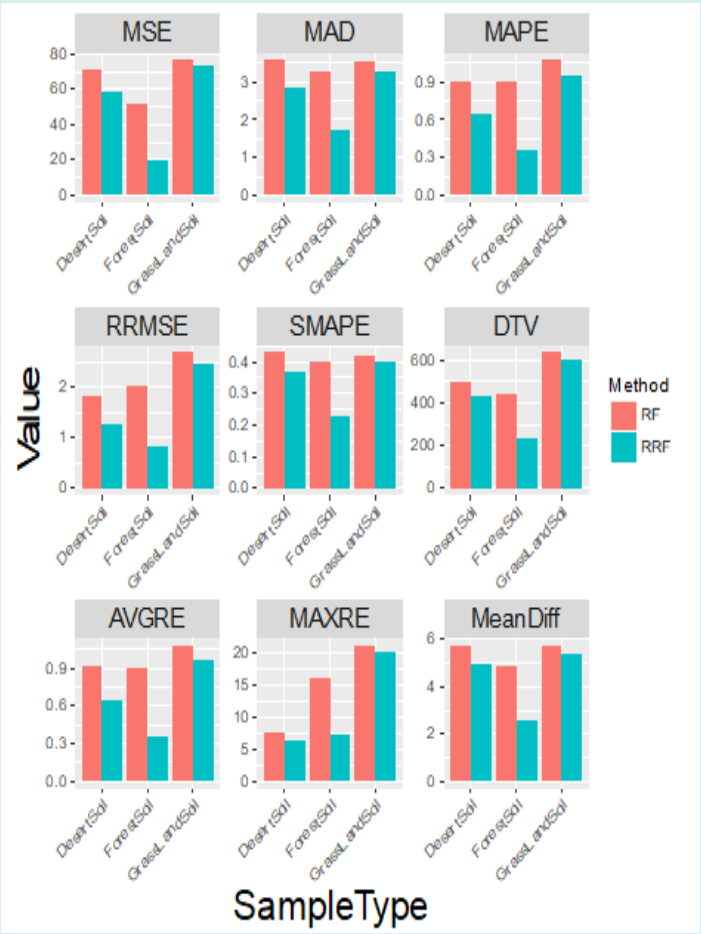

Figure 2: Measurement performance for the comparison of RF and RRF under three different soil types, Desert Soil, Forest Soil and Grass Land Soil. Nine metrics are calculated to compare the predicted values and true values.

Case Study 2 (Metcalf et al. 2013 Data)

This study uses the mouse data from "A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system" [7]. The microbial communities were sampled from the abdominal cavity, skin of head, skin of body, soil under the corpses and soil without corpses. Samples from soil without corpses were treated as the control group. Sample data was collected at eight time points: 0, 3, 6, 9, 13, 20, 34 and 48 days [7]. All analyses were performed using OTUs at the family level. In the original study, Metcalf et al. removed the samples at the last time point, 48 days, and used the RF model with leave-one-out cross validation method to predict PMI. To be comparable, we also removed samples at 48 days. As there was rich prior information was recorded for this study, including soil pH, distended belly, and body score, we included them into our RRF model as predictors.

The distended belly is a three-level categorical variable which measures the body decay [7]. The body score contains two variables, key of head and key of torso. We repeated the RRF model with leave-one-out cross validation 100 times,

and then took the average of these 100 predicted values of each sample. Figure 3 shows the prediction results using the average absolute error. Comparing RRF with RF, the average absolute error was reduced from 6.02 to 2.03 days for abdominal cavity, from 5.85 to 1.96 days for skin of torso, from 9.74 to 1.99 days for skin of head, from 5.85 to 1.82 days for soil under corpse. In another word, the accuracy was improved about 66.3%~ 79.6%. The results from regularized random forest were more accurate than the results from the random forest.

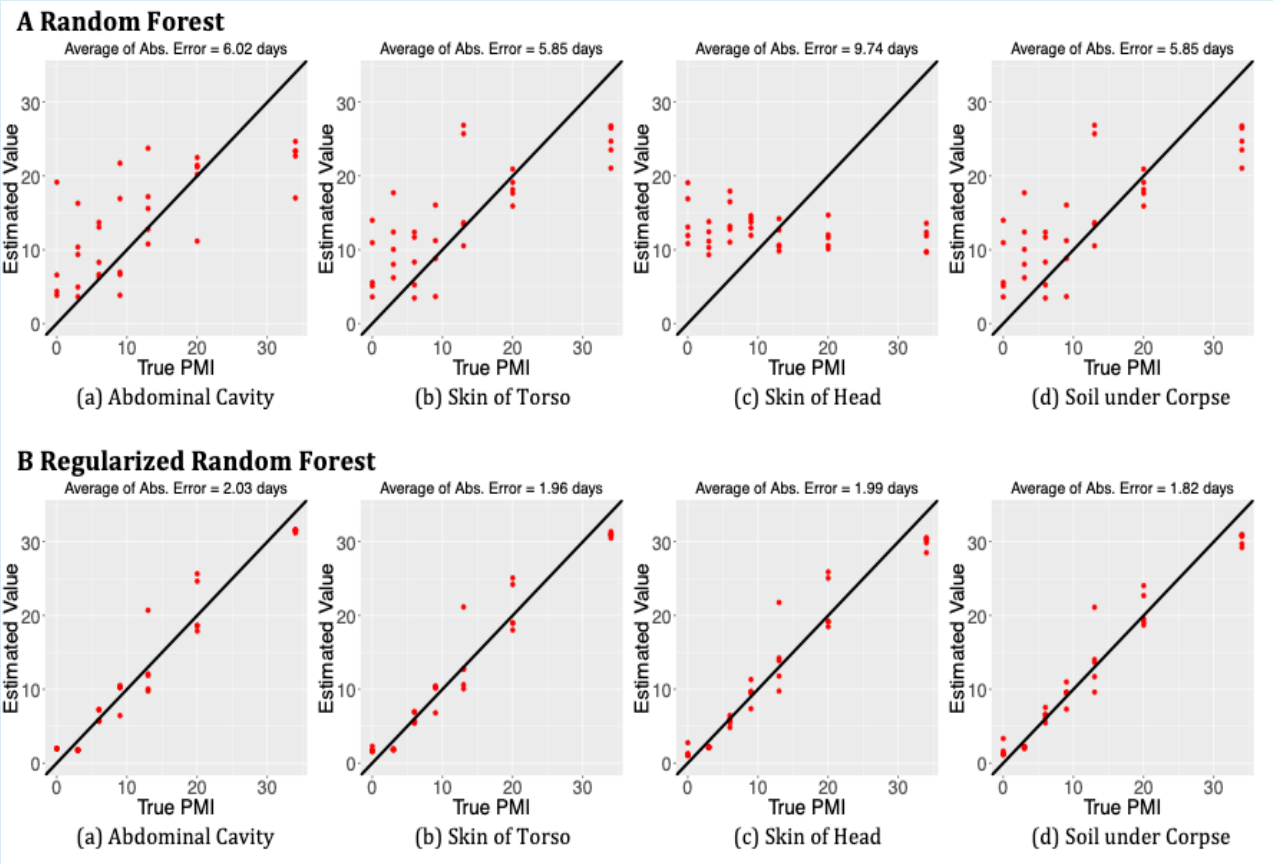

Figure 3: Estimation of PMI using Random Forest (top) and Regularization Random Forest (bottom). The black line (45 degree line) represents a perfect estimation. The x-axis represents the true PMI from the experiments. The average absolute error between the predicted number of days and true PMI for each situation is shown at the top of each panel.

From the nine measurement results as shown in Figure 4, it is obvious that prediction using RRF was much more accurate than that using RF. Our method improved the prediction results by a sizable margin. Figure 5 shows the 95% prediction intervals for PMI by RRF method. We repeated regularized random forest with leave-one-out 100 times, then we took the 2.5% quantile as the lower boundary and the 97.5% quantile as the upper boundary. The error bars in Figure 5 are the 95% confidence intervals. Almost all 95% prediction intervals from sample ID 11 to 30 contain the true value of PMI, which indicates our model has better estimation for inner points (not the start and end points). It also shows that RRF overestimates PMI at the start point while underestimates PMI at the end point. These findings are consistent with results as shown in the scatter plots of Figure 3.

In previous analysis, prediction was conducted for each body location separately. As data are collected from four different body locations for each mouse, prediction of PMI can be conducted by RRF combining data from all four body locations. The results are shown in Figure 6. The average absolute error was significantly decreased to 0.58, about half day, which is more accurate than the individual prediction as shown in Figure 3. The estimates at the beginning and the end of the death time points are still less accurate comparing with those for inner time points, however, the estimates are closer to the true values when we combined four body locations together, the accuracy of prediction is greatly improved.

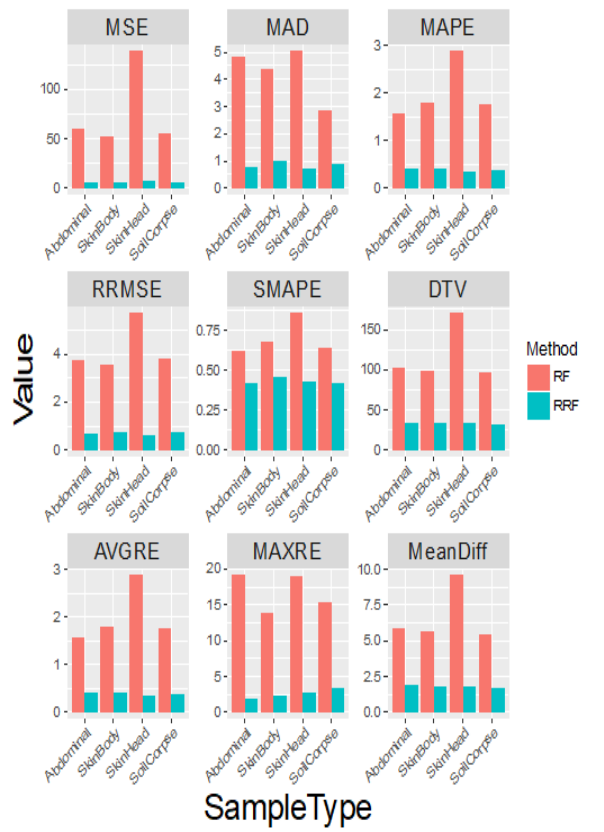

Figure 4: Bar plots of nine performance measurements for the case study 2. The x-axis shows sample types, including the Abdominal, SkinBody, Skin Head and Soil under the corpse. The smaller the performance value (shown on y-axis), the better the performance.

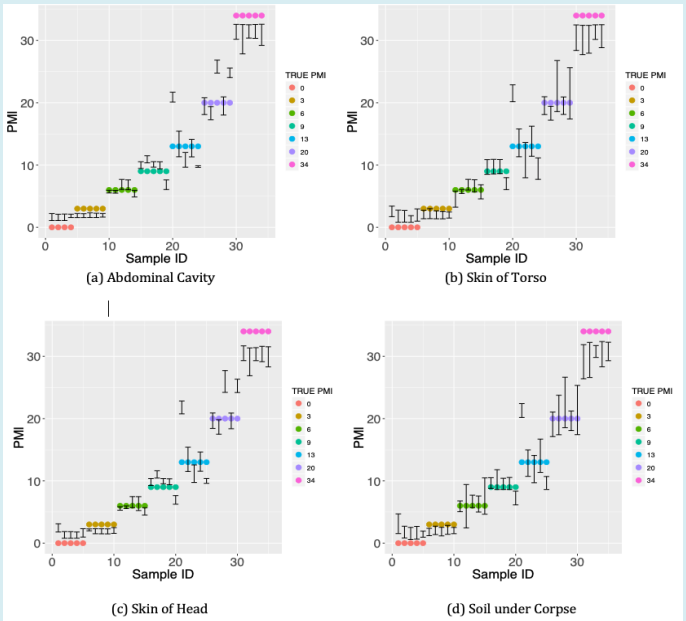

Figure 5: 95 % prediction interval at various time points. the x-axis represents the sample index. the y-axis represents the true value of pmi. the color points represent the true pmi at different time points and the error bars correspond to the 95 % prediction intervals for each sample. note: there are 5 mice at each time point.

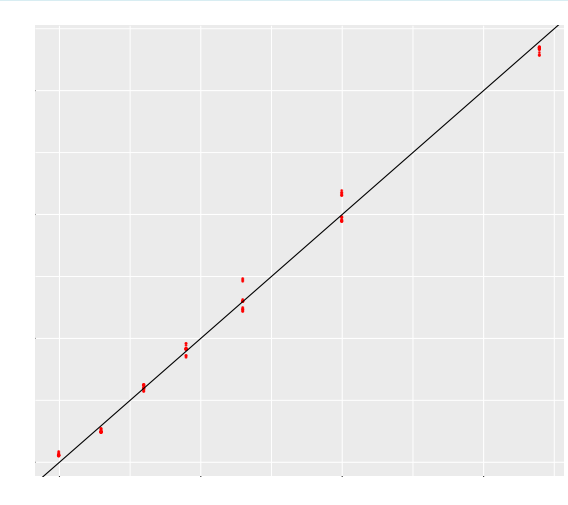

Figure 6: Estimation of PMI using Regularization Random Forest on the combined data from four body locations. The black line (45 degree line) represents a perfect estimation.

Importance of Taxonomical Features

RRF has a potential to identify important taxonomical features of decomposition microbiome which may influence PMI. We detected 17 features among four body sites in Metcalf, et al. dataset, and 145 features among three soil types in Metcalf et al. 2016 dataset. It is worthy to note that 12 overlapped features are found in both studies and they are likely associated with body decomposition.

A number of Bacillaceae family strains have been isolated from decomposing solid organic waste [22-24]. In another study, Brucellaceae was one of 10 most informative taxes which explained the most amount of variation in physiological time [25]. Comamonadaceae was reported that it could affect human decomposition [26] and was also detected during the bloat stage of decomposition, a stage readily attributed to microbial physiology [27]. Enterobacteriaceae was detected in the lower GI tract samples in both samples at pre-bloat and end-bloat stages [27]. The Clostridium, Bacteroides, Staphylococcus and Enterobacteriaceae were reported that associated with the destruction of soft tissue due to the enteric obligate and facultative anaerobic microorganisms [28,29]. The families Microbacteriaceae and Micromonosporaceae are reported dominant in the forest floor actinobacterial communities [29]; Actinobacterial is found an important predictor for predicting physiological time [25]. Moraxellaceae is commonly found in soil [30], which could explain its presence on both cadavers during the late stages of decomposition [31]. Hyde et al. found that Mocaxellaceae had significant effect on fresh pig corpses through 24 hours of decomposition [31]. They also suggested that Moraxellaceae family could be an important group of decomposition [31,32]. Pechal, et al. [25] found that Planococcaceae was the most dominant family on the fifth day of decomposition, Pseudomonadaceae one of the significant member of the skin microbial community, and Rhodocyclacear significantly affected physiological time. All of the evidence here shows that the detected taxa by RRF method have already been reported by other studies, which may give us a confirmation/support of our current findings.

Conclusion and Discussion

In this paper we aimed to estimate the PMI based on microbial changes and environmental factors. The pH value of soil under the body and body score were found highly correlated with PMI, which motivated us to include them as predictor variables in our model. Here we focused on utilizing a very powerful tool, Regularized Random Forest regression models. Regularized Random Forest method results in more accurate prediction than RF in both case studies. We reduced the average absolute error in prediction to only 2 days in the second study when including these environmental predictors. Using the combined data from four locations, we even reduced the average absolute error to only half day which shows a promising and exciting result.

Regularized Random Forest is like a sophisticated tool for predicting outcomes, standing out from the usual methods like traditional Random Forest, decision trees, support vector machines (SVM), and neural networks [32]. What makes it special is that it is really good at making predictions, especially when dealing with complicated data or when we do not have much information. It does this by automatically choosing the most important factors for making predictions, helping us understand why it predicts certain things. It is like finding the right balance between being easy to understand

and having the necessary complexity, making it a reliable choice for making sense of data. In simpler terms, it is a smart and advanced tool with features that make it both powerful and user-friendly for predicting outcomes based on data.

Regularized Random Forest, while advantageous for PMI estimation, has limitations. Tuning regularization parameters is complex, and interpretability, though improved, is not as straightforward as simpler methods. Adequate data is crucial for effective regularization, and the model may face challenges with limited data. Aggressive regularization may lead to information loss, and algorithmic complexity remains a consideration for large datasets. Sensitivity to outliers persists, and there is a risk of overfitting noise in the training data. The introduced complexity might not always be beneficial, particularly in cases where relationships are straightforward. However, considering that microbial data is typically high-dimensional, involving a large number of species, the Regularized Random Forest method holds promise.

Conflicts of Interest

None declared.

Funding

This work has been partially supported by the United States Department of Agriculture (ARZT-1361620-H22-149) to L.A.

References

1. Saks MJ, Koehler JJ (2005) The coming paradigm shift in forensic identification science. Science 309(5736): 892-895.

2. Goff ML, Flynn MM (1991) Determination of postmortem interval by arthropod succession: a case study from the Hawaiian Islands. Journal of Forensic Science 36(2): 607-614.

3. Sperling FA, Anderson GS, Hickey DA (1994) A DNA-based approach to the identification of insect species used for postmortem interval estimation. Journal of Forensic Science 39(2): 418-427.

4. Lord WD, Goff ML, Adkins TR, Haskell NH (1994) The black soldier fly Hermetia illucens (Diptera: Stratiomyidae) as a potential measure of human postmortem interval: observations and case histories. Journal of Forensic Science 39(1): 215-222.

5. Vass AA, Barshick SA, Sega G, Caton J, Skeen JT (2002) Decomposition chemistry of human remains: a new methodology for determining the postmortem interval. Journal of Forensic Science 47(3): 542-553.

6. Bauer M, Gramlich I, Polzin S, Patzelt D (2003) Quantification of mRNA degradation as possible indicator of postmortem interval-a pilot study. Legal medicine 5(4): 220-227.

7. Metcalf JL, Parfrey LW, Gonzalez A, Lauber CL, Knights D, et al. (2013) A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. Elife 2: e01104.

8. D'Angiolella G, Tozzo P, Gino S, Caenazzo L (2020) Trick or Treating in Forensics-The Challenge of the Saliva Microbiome: A Narrative Review. Microorganisms 8(10): 1501.

9. Procopio N, Lovisolo F, Sguazzi G, Ghignone S, Voyron S, et al. (2021) "Touch Microbiome" as a Potential Tool for Forensic Investigation: A Pilot Study. J Forensic Leg Med 82: 102223.

10. Alan G, Sarah JP (2012) Microbes as Forensic Indicators. Trop Biomed 29(3): 311-330.

11. Metcalf JL, Xu ZZ, Bouslimani A, Dorrestein P, Carter DO, et al. (2017) Microbiome Tools for Forensic Science. Trends Biotechnol 35(9): 814-823.

12. Oliveira M, Amorim A (2018) Microbial Forensics: New Breakthroughs and Future Prospects. Appl Microbiol Biotechnol 102(24): 10377-10391.

13. Metcalf JL, Xu ZZ, Weiss S, Lax S, Treuren WV, et al. (2016) Microbial community assembly and metabolic function during mammalian corpse decomposition. Science 351(6269): 158-162.

14. Belk A, Xu ZZ, Carter DO, Lynne A, Bucheli S, et al. (2018) Microbiome Data Accurately Predicts the Postmortem Interval Using Random Forest Regression Models. Genes 9(2): 104.

15. Yang F, Zhang X, Hu S, Nie H, Gui P, et al. (2023) Changes in Microbial Communities Using Pigs as a Model for Postmortem Interval Estimation. Microorganisms 11(11): 2811.

16. Zhang J, Wang M, Qi X, Shi L, Zhang J, et al. (2021) Predicting the postmortem interval of burial cadavers based on microbial community succession. Forensic science international: Genetics 52: 102488.

17. Houtao D, Runger G (2013) Gene selection with guided regularized random forest. Pattern Recognition 46(12): 3483-3489.

18. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, et al. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37(8): 852-857.

19. Meadows L, Robert WM, William MB (1990) Time since death and decomposition of the human body: variables and observations in case and experimental field studies. Journal of Forensic Science 35(1): 103-111.

20. Yvan S, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19): 2507-2517.

21. Luo Q, Lu M, Zhang M, Jiang H, An L (2023) A Regression-based approach for accurate source tracking using microbial communities. Int J Forens Sci 8(4): 000338.

22. Yusaku F, Kume S (1991) Isolation and identification of thermophilic bacteria from sewage sludge compost. Journal of fermentation and bioengineering 72(5): 334-337.

23. Jaak R, Mergaert J, Coosemans J, Deprins K, Swings J (2003) Microbiological aspects of biowaste during composting in a monitored compost bin. Journal of Applied Microbiology 94(1): 127-137.

24. Peter SF (1985) Effect of temperature on bacterial species diversity in thermophilic solid-waste composting. Applied and Environmental Microbiology 50(4): 899-905.

25. Pechal JL, Crippen TL, Benbow ME, Tarone AM, Dowd S, et al. (2014) The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. International Journal of Legal Medicine 128(1): 193-205.

26. Anne W, Ley JD, Gillis M, Kersters K (1991) NOTES: comamonadaceae, a new family encompassing the acidovorans rRNA complex, including variovorax paradoxus gen. nov., comb. nov., for Alcaligenes paradoxus (Davis 1969). International Journal of Systematic and Evolutionary Microbiology 41(3): 445-450.

27. Embriette HR, Haarmann DP, Lynne AM, Bucheli SR, Petrosino JF (2013) The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. PloS one 8(10): e77733.

28. Avrom IM, Melvin JR, Simson LR, Cronholm LS (1984) Bacterial transmigration as an indicator of time of death. Journal of Forensic Science 29(2): 412-417.

29. Sarah ED, Zak DR (2010) Simulated atmospheric nitrogen deposition alters actinobacterial community composition in forest soils. Soil Science Society of America Journal 74(4): 1157-1166.

30. Paul B (1968) Isolation of Acinetobacter from soil and water. Journal of bacteriology 96(1): 39-42.

31. Embriette HR, Haarmann DP, Petrosino JF, Lynne AM, Bucheli SR (2015) Initial insights into bacterial succession during human decomposition. International journal of legal medicine 129(3): 661-671.

32. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in bioinformatics 2: 927312.