Indonesian Journal of Electrical Engineering and Computer Science
Vol. 34, No. 1, April 2024, pp. 534∼547
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i1.pp534-547

Modeling agricultural and methane emission data: a finite mixture regression approach

Pattharaporn Thongnim[1,2], Ekkapot Charoenwanit[3]
[1]Department of Mathematics, Faculty of Science, Burapha University, Chonburi, Thailand
[2]Data Center, Faculty of Science and Arts, Burapha University, Chanthaburi, Thailand
[3]The Sirindhorn Thai-German Graduate School of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Article Info

Article history:

Received Dec 14, 2023
Revised Jan 17, 2024
Accepted Jan 19, 2024

Keywords:

Crop production
EM algorithm
Finite mixture model
Methane emissions
Unsupervised learning

ABSTRACT

In this paper, the method for unsupervised learning of finite mixture regression (FMR) models is presented for evaluation using agricultural and emissions data sets. The FMR models can be written as problems with incomplete data, and the expectation–maximization (EM) algorithm can be used to estimate unknown variables. The goals of this research are to find the best clustering model with different sets of training and test data and examine the relationship between crop production index and methane emissions in 22 countries from 1990 to 2019 using FMR. In this study also use machine learning process for a FMR model from real world data. According to the findings, the performance of the random training data (RDM) in time series is preferable to that of the fixed training data (FXM). In addition, both RDM and FXM are capable of classifying the 22 countries into two distinct groups and constructing the parameters for the regression model. However, selecting training and test data will result in a good prediction; it is dependent on the data collected. Picking the right training and test data is crucial for accurate predictions-it all comes down to having good data in the first place.

Corresponding Author:

Ekkapot Charoenwanit
The Sirindhorn Thai-German Graduate School of Engineering
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
Email: ekkapot.c@tggs.kmutnb.ac.th

1. INTRODUCTION

As a result of growing energy demands, industrialization, and climate change, environmental challenges have recently emerged as major worldwide concerns [1]-[3]. Global climate change produced by greenhouse gas emissions has become one of the most pressing issues facing contemporary society, threatening the global ecological environment and the economy's sustainable development [4]. It can be seen that the average temperature is expected to globally rise by two degrees Celsius by the end of the twenty-first century, owing to global warming and greenhouse gas (GHG) emissions [5]. Agriculture is one of the aspects of the economy that would suffer the most as a direct consequence of GHG emissions and agriculture is one of the major contributors to global warming, which is caused by human activities.

Furthermore, agricultural GHG emissions are constantly increasing as traditional farming methods are increasingly used to meet the world's growing food demands, in order to feed the world's growing population [6]. Therefore, about one-third of the world's total greenhouse gas emissions are caused by emissions from agricultural practices [7]. For example, GHGs come from the production of agricultural inputs such as fertilizer, pesticides, herbicides, and irrigation, as well as farm machinery used for spraying, tilling, harvesting, stocking, and shipping. GHGs such as methane (CH_4), carbon dioxide (CO_2), and nitrous oxide (N_2O) have been released in greater quantities as a result of agricultural processes [8]. Nitrous oxide (N_2O) and methane (CH_4) have a global warming potential, which is 296 and 34 times more than that of carbon dioxide (CO_2), respectively, despite the fact that their emissions are smaller than that of carbon dioxide (CO_2). In order to measure GHGs, the effects of crop production on and emissions must be evaluated. Methane is a greenhouse gas in the atmosphere and a major influence on global warming. Agriculture, waste, and energy are the three main sectors that have been identified as the major sources of methane emissions. It can be seen that methane emissions from farming seem to be a major, and possibly the main cause of the growth of the atmosphere over the past ten years [9]. For instance, agriculture is the largest man-made source of emissions, with ruminant livestock dominating the sector, and rice paddies also contribute to the problem [10]. Additionally, wetland and crop residue retention leads to increased methane (CH_4) emissions. As the amount of methane in the atmosphere keeps increasing quickly, there are an increasing number of methods for reducing it. This is important due to the need to find a balance between food security and environmental protection.

The relationship between climate change, and agriculture is a very important issue because the world's food production resources are already being put under a lot of pressure by a rapidly growing population around the world [11]. The influence of methane emissions on crop production is of particular interest, not only because of the global significance of crop production as a food source but also because the recent explosive growth of crop production, particularly in many regions, contributes to global warming through the emission of methane (CH_4) into the atmosphere. Consequently, Gaussian mixture models can usually be used to predict methane emissions and determine the relationship between crop production and methane emissions in different countries. How to construct the prediction models that require small amounts of climate change data from various regions of the world. In addition, how to split a data set into training data and test data in order to evaluate prediction models. Finally, how to categorize the countries into groups and determine the pattern of the mixture using a finite mixture model (FMM).

The paper outlines an approach to apply unsupervised learning techniques, specifically finite mixture regression (FMR) models, to analyze methane emissions across various country groups. FMR models are particularly suited to handling datasets with inherent groupings and can account for different regression relationships within these subsets. By treating the problem as one with incomplete data, the expectation-maximization (EM) algorithm is employed to estimate the unknown variables, which is a common approach in such statistical modeling to maximize the likelihood function when data are missing or latent variables are present [12]. Therefore, the research aims to construct an optimal predictive model for methane emissions, segmenting the data into distinct groups of countries. This stratification allows for more nuanced analysis, recognizing that the relationship between the crop production index and methane emissions may vary significantly across different national contexts.

The study spans data from 1990 to 2019, encompassing 22 countries, providing a broad temporal and spatial frame of reference for the FMR application. Moreover, the study ventures into the realm of machine learning by operationalizing the FMR model with real-world data, emphasizing the practical application of the methodology with technology in agriculture [13]. In this study, experimentation with various combinations of training and test datasets is a critical part of the model development process. The paper reports that random training data (RDM) sequences yield better performance in time-series modeling than fixed training data (FXM) sequences. This could suggest that the model benefits from the variability and potential non-linearity captured by random sampling. Furthermore, the FMR models prove effective in clustering the countries into two distinct groups, allowing for targeted regression parameter construction within these clusters.

In the realm of machine learning, the effectiveness of a predictive model is significantly dependent on the quality and representativeness of its training and test datasets [14]. This is vital because a model trained on limited or biased data might exhibit excellent performance during training but fail in real-world applications. Therefore, the process of carefully selecting and preprocessing data is not just a preliminary step but a cornerstone in building a model that not only performs well in theory but also functions effectively and reliably in diverse, real-world scenarios. This highlights the need for meticulous attention in preparing data, encompassing tasks like cleaning data, managing missing values, and ensuring the data used for training truly represents the environment in which the model will operate.

2. METHOD

2.1. Data collection

In this detailed investigation, the study delineates two primary factors influencing GHG emissions, specifically focusing on the agricultural sector. The research categorizes countries into distinct groups based on these factors, which encompass agricultural practices and GHG emissions, with a spotlight on crop production as a significant contributor to methane release into the atmosphere [15]. The crop production index serves as a proxy for agricultural output, and its impact on methane emissions, measured in kilotonnes (kt) of CO2 equivalent, forms the crux of the analysis. To facilitate this analysis, time-series data spanning from 1990 to 2019 were sourced from the World Bank , a decision underpinned by the need for reliable and comprehensive global datasets. The study's breadth is considerable, comprising 22 countries that present a mix of developed and developing nations, each with unique farming systems and practices. This diverse selection includes Austria, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, the United Kingdom, Greece, Italy, Norway, New Zealand, Portugal, Sweden, Cambodia, the Philippines, Vietnam, Japan, Thailand, Laos, and Myanmar. Choosing 22 countries from both Asia and Europe for research offers several advantages, especially given criteria of including countries with varying economic sizes and a mix of developed and developing nations.

Here are some reasons why this approach is beneficial. Firstly, by selecting countries with small, medium, and large economies, this study can capture a wide range of economic contexts. This diversity is crucial in understanding how economic factors influence agricultural practices and environmental impacts, such as methane emissions. Secondly, different countries have unique farming systems and practices. This variability can provide insights into how different agricultural methods contribute to or mitigate methane emissions. It can also offer a broader understanding of sustainable farming practices in diverse economic contexts. Next, different countries have varied cultural approaches to agriculture and differing environmental policies. Studying a broad range of countries allows for an examination of how these factors influence agricultural practices and methane emissions. Lastly, with a selection of 22 countries, the study can perform comparative analyses between nations with similar economic statuses or between developed and developing countries. This can help identify best practices and areas needing improvement in different contexts.

The dataset's temporal depth and geographical breadth, accounting for 660 observations across 30 years and 22 countries, allow for a robust statistical examination. The inclusion of both developed and developing countries provides a comparative lens to assess how different agricultural systems and stages of economic development correlate with methane emissions. The study's approach, which involves grouping countries and analyzing the time series within each cluster, could reveal patterns and trends that are not apparent in more generalized global models. By meticulously collating and analyzing this data, the research aims to offer insights into the complex dynamics between agricultural productivity and its environmental impact, potentially informing policy and practice in both national and international contexts.

2.2. Data processing and data visualization

Before being analyzed, the crop production index as well as the methane emissions data obtained are cleaned. The process of cleaning involved verifying the data to ensure that they are accurate and comprehensive [16], [17]. Every row of data contained an index of crop production and methane emissions for a specific country, as well as a time step. Rows that have erroneous data or data that is missing are removed. In addition, in order to analyze methane emissions, the logarithmic transformation is used. As a result, the application of data transformation can increase the efficiency of model training and remove numerical errors caused by the skewness of a measurement variable.

Plotting the data can provide a visual context and identify trends, patterns, and outliers in the data. Using scatter plots, relationships between crop production and methane emissions in varying countries over the years are observed in Figure 1. Employing scatter plots, a versatile tool in data visualization, allows for meticulous exploration and analysis of the relationships between crop production and methane emissions across different countries over a span of years. These scatter plots enable a multi-dimensional view, where each point represents a specific country's data at a given time. Adjusting variables such as color and pattern of the data points further distinguishes between countries, years, or other relevant categories [18]. This method provides a nuanced understanding of how crop production correlates with methane emissions, highlighting any direct or indirect associations. It also facilitates the identification of anomalies or exceptional cases, where the relationship deviates from the general trend. Such visualizations are instrumental in drawing meaningful conclusions and guiding subsequent data-driven strategies in agricultural and environmental management. Therefore, these plots allow for a clear visual comparison, highlighting correlations or anomalies in the data and aiding in understanding environmental impacts in different regions.

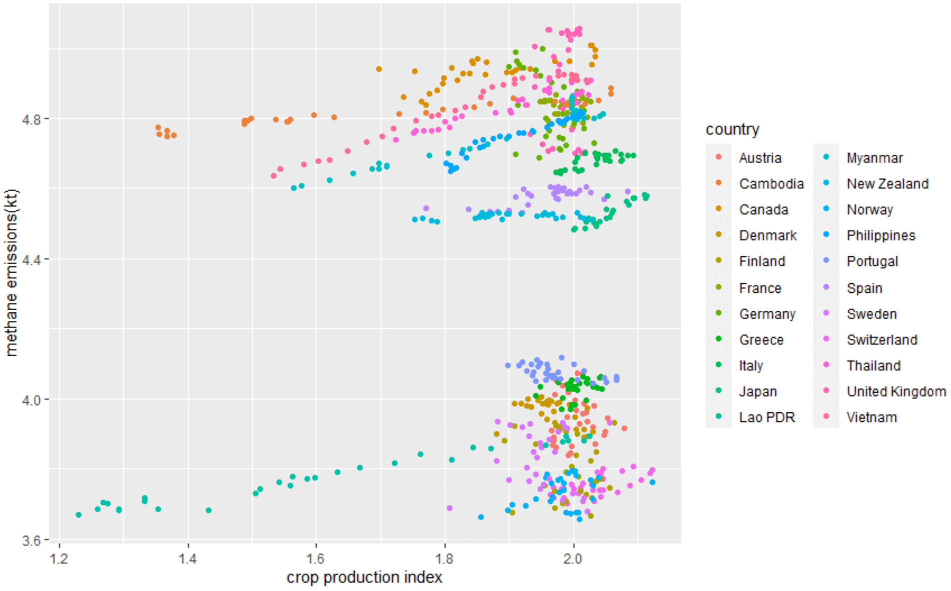

Figure 1. A scatter plot of the relationship between crop production index and methane emissions (kt of CO_2 equivalent)

2.3. Machine learning methods

Sometimes, it appears difficult to determine the optimum models using the available mathematical and statistical tools. Regression models are a standard technique for statistical procedures of prediction that are extensively used in numerous applications and fields [19]. However, machine learning models have been used to overcome the relationship of data and achieve improved prediction accuracy more efficiently [20], [21]. It can be said that one of the key benefits of utilizing statistical models is that they provide intuitive data visualizations that support the discovery of correlations between variables and the formulation of predictions. Machine learning focuses on making predictions by using general-purpose learning algorithms to find patterns in data that are more beneficial. Therefore, the use of these machine learning models in this analysis is justified by their ability to predict interactions between two variables.

An effective machine-learning solution is designed to fit data. A trained model must perform well on data available during the training data process as well as test data [22]. A split data technique is commonly utilized, in which a larger subset of existing records is used for learning models and a smaller sample for testing models [23]. Usually, these are referred to as training and test sets, respectively. Different approaches are employed in the data selection process to discover a satisfactory performance for the prediction models. Machine learning such as an unsupervised learning method is used to find the best fit model. The FMMs and EM algorithm are used to highlight the concepts of unsupervised learning in this study. Moreover, the FMMs, particularly Gaussian mixtures, have been widely utilized in statistical machine learning as an unsupervised technique for clustering data in a variety of scientific disciplines including agricultural and ecological science.

Therefore, this research uses two different methods for splitting data into training and testing sets. The first case study is based on data series. Therefore, the first 547 samples are collected as a training set, representing 80 percent of the length of the entire series from 1990 to 2013 (24 years) in study countries. The remaining data represents the test set, which is used to evaluate the accuracy of predictive models. This is the last six of the available data points in each country that fixes the number of years FXM. In the second case study, there are a total of 547 samples taken by a random sampling process RDM. These samples also accounted for 80 % of the total.

2.4. A finite mixture model

A FMM is a semi-parametric method employed for fitting complex data distributions and estimating density [24]. This model provides a flexible approach for representing data originating from a heterogeneous population. Notably, FMM offers the benefits of both analytic simplicity and modeling flexibility when dealing with complex data distributions.

Let $x = [x_1, ..., x_p]$ be a p-dimensional random variable. The probability density function $p(x)$ has the following form:

$$p(x|\Theta) = w_1 f_1(x|\theta_1) + w_2 f_2(x|\theta_2) + ... + w_K f_K(x|\theta_K)$$

where $\Theta = ((\theta_1, ..., \theta_K)^T, W)$ denotes the set of all unknown parameters, $W = (w_1, ..., w_k)^T$ is the mixing proportion with the condition $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$, and $k = 1, 2, \ldots, K$ is the number of components.

Given a set of n data samples and identically distributed samples $X = \{x^{(1)}, ..., x^{(n)}\}$, the log-likelihood can be written as a function.

$$\begin{aligned} \log p(x|\Theta) = & \log \Pi_{i=1}^{n} p(x^{(i)}|\Theta) \\ & = \log \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f(x^{(i)}|\theta_k) \end{aligned}$$

The maximum likelihood machine learning estimate of parameter values is expressed as:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log p(X|\Theta)$$

however, it is important to note that this expression cannot always be analytically determined.

The linear regression of mixture models is part of FMMs [25]. Let y be a response variable, x be a independent variable that has an effect on y, and $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(n)}, y^{(n)})$ be sample of observations from the linear mixture model. Let z be a latent variable with $P(z_i = k|x)$ for $k = 1, \ldots, K$. Given $z_i = k$, the linear regression is:

$$y_i = x^T \beta_k + \varepsilon_{ik}$$

a FMR model is defined as follows:

$$\begin{aligned} f(y|x, \Theta) = & \sum_{k=1}^{K} w_k f(y|x, \theta_k) \\ & w_k \geq 0, \sum_{k=1}^{K} w_k = 1 \end{aligned}$$

where Θ is the vector of all unknown parameters and $\theta_k = x^T \beta_k$, σ_k^2 and w_k is the mixing proportion. The presented approach and theoretical insights can be applied to the Gaussian mixture linear regression model.

$$f(y|x) = w_1 G(x^T \beta_1, \sigma_1^2) + ... + w_k G(x^T \beta_k, \sigma_k^2)$$

The log-likelihood of a sample of n observations can be defined as follows.

$$\log L = \sum_{i=1}^{n} \log \sum_{k=1}^{k} w_k f(y_i|x_i, \theta_k)$$

Typically, direct optimization of the log-likelihood function is not feasible. The EM algorithm, detailed in the following section, is the most common method for obtaining the maximum likelihood estimate of the parameter vector with latent variables.

2.5. The expectation-maximization algorithm

The EM algorithm [26] is an iterative approach commonly used for estimating the parameters of statistical models, especially in the context of FMMs, where latent variables, in addition to unknown parameters, are involved. Hence, the EM algorithm, which converges to a (local) maximum likelihood estimate of the parameters of mixture models, serves as a fundamental method employed in fitting mixture models to observations. The EM algorithm demonstrates its proficiency in handling incomplete data. As a result, it is suggested to apply the EM algorithm to the model for estimating the parameters of FMR by incorporating hidden variables."

The EM algorithm consists of two stages: the computation of a specific conditional expectation of the log-likelihood (E-step) and the maximization of this expectation over the relevant parameters (M-step). For the current iteration $t+1$, using arbitrary initial values $w_k^{(0)}, \beta_k^{(0)}, \sigma_k^{2(0)}; k = 1, ..., K$ and the parameter values from the previous iteration t, the EM algorithm is applied to maximize (2.4.) as outlined in [27], [28]:

E-step: evaluate the responsibility $w_{ik}^{(t+1)}$ of each component $k = 1, ..., K$ for each data point $i = 1, ..., n$ using the current value of the parameters $\beta_k^{(t)}$,$\sigma_k^{(t)}$ and $w_k^{(t)}$.

$$w_{ik}^{(t+1)} = \frac{w_k^{(t)} f(y_i|x_i^T \beta_k^{(t)}, \sigma_k^{2(t)})}{\sum_{j=1}^{K} w_j^{(t)} f(y_i|x_i^{(t)} \beta_j^{(t)}, \sigma_j^{2(t)})}$$

M-step: re-estimate the parameters $\beta_k^{(t+1)}$,$\sigma_k^{(t+1)}$ and $w^{(t+1)_k}$ for each component $k = 1, ..., K$ using the recently obtained values of $w_{ik}^{(t+1)}$ from the E-Step

$$\beta_k^{(t+1)} = \underset{\beta_k}{\text{argmax}} \sum_{i=1}^{n} w_{ik}^{(t+1)} (y_i - x_i^T \beta_k)^2$$

$$w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(t+1)}$$

$$\sigma_i^{2(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(t+1)} (y_i - x_i^T \beta_k^{(t+1)})^2$$

The EM algorithm begins by augmenting the given data with estimates of latent group membership. Implicitly, it assumes that k represents the number of components to be estimated from the data. In addition, the EM method iteratively updates the parameters using a weighted least-squares estimation. The E (Expectation) and M (Maximization) steps are repeated until a convergence condition is satisfied.

2.6. Model selection criterion

This work focuses on using FMR models when population heterogeneity is unknown. Identifying the number of latent classes that best match the data is the first step in performing regression mixture analysis. Then, the effects of independent variables on the dependent variable can be investigated within each latent class. In order to implement the EM method, it is necessary to know the appropriate number of components, k. Too many components result in overfitting, whereas too few results in underfitting. It is important to choose the right number of components. The right number of components can show some significant underlying structure that describes the data.

This application requires data-driven estimation. As a result, a problem in statistics is the number of components (k) estimation problem in mixture models. Therefore, this research indicates using the Bayesian Information Criterion (BIC) to find the number of Gaussian components. In general, it is known that BIC can be used to estimate k in a consistent way [29]. It can be said that the BIC is a reliable and efficient method for determining the number of mixture components when each component has a Gaussian distribution. For the k estimation problem in FMR models, information-theoretic methods like BIC are often used [30], [31]. For a given set of data, the estimation method is used to fit FMR models with different orders of $k = 1, 2, 3$ that are considered in this application. The best possible value for k can be determined by using the minimization [32]:

$$BIC_k = -2\log(L_k) + m\log(n^*)$$

where n^* is the sample of the training data set and m is the total number of parameters.

2.7. Evaluation models

For the purpose of the present investigation, 80% of the data is chosen to be used as the training set, and the other 20% is placed to use as the testing dataset. On the basis of the results of the three statistical evaluation models, the performance of the models is evaluated. Therefore, the performance of the models in this study is evaluated by utilizing the coefficient of correlation (R^2), the mean absolute error (MAE), and the root-mean square error (RMSE) [33].

R^2 is one of the most important measures for assessing the accuracy of prediction models with a value between 0 and 1. The models with the highest values are considered the best and are better at predicting than others.

$$R^2 = \left[\frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i^* - \bar{y}^*)}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y}^*)^2 \sum_{i=1}^{N}(y_i^* - \bar{y}^*)^2}} \right]^2$$

Where y_i is the actual data of sample i, $\bar{y}$ is the mean of the actual data, y_i^* is the predicted data of sample i, $\bar{y}^*$ is the mean of the predicted data and N is the number of data.

The MAE and RMSE both represent the error that happens between the actual data and the predicted data, and values that are lower indicate that the predictions have more forecast accuracy:

$$MAE = \frac{1}{n}\sum_{i=1}^{N}|y_i - y_i^*|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(y_i - y_i^*)^2}$$

The performance of models is evaluated based on their R^2, MAE, and RMSE values.

3. RESULTS AND DISCUSSION

3.1. Data analysis

The provided line graph offers a visual comparison between the mean values of crop production and methane emissions across a selection of countries in Figure 2. Displayed on the x-axis are the countries, while the y-axis quantifies the mean values for the two distinct datasets. The blue line, marked with circles, represents the mean crop production values, suggesting a relatively stable trend across the nations, with slight variations. In contrast, the green line, highlighted with squares, corresponds to mean methane emissions and shows more significant fluctuations. This visual representation underscores the variability and complexity of agricultural metrics and their environmental implications. The precise data points are not provided, which limits the ability to extract exact figures for each country directly from the graph.

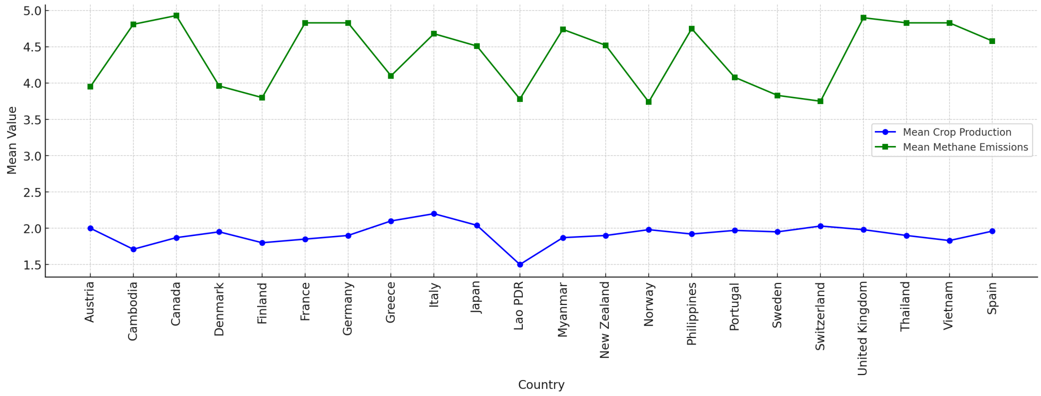

Figure 2. Comparison of crop production and methane emissions by countries

The graphs presented provide a comprehensive overview of the distribution of data related to crop production and methane emissions across different countries in Figures 1-4. In Figure 2, the mean values of the methane emissions index are specifically highlighted. The plots reveal an intriguing observation: one might anticipate a correlation between high crop production and elevated methane emissions, yet this is not consistently observed across the board. It is noteworthy that certain countries, such as Canada, the United Kingdom, France, Germany, Vietnam and Thailand, register high indices for methane emissions. Conversely, countries like Norway, Switzerland, Finland, Denmark, Sweden, and Lao report methane emission indices below 4, with Canada exhibiting the highest emission levels and Norway the lowest within the dataset. Moreover, the temporal and geographical variability in methane emissions is apparent, indicating that these patterns are not uniform. For instance, countries like Canada, Cambodia, Myanmar, Lao, the Philippines, Spain, Thailand, and Vietnam show an upward trajectory in their methane emissions over time. On the other hand, a set of countries, including Austria, Denmark, Finland, France, Germany, Sweden, and the UK, demonstrate a declining trend in emissions. This diversity in patterns underscores the complex interplay of factors influencing methane emissions and highlights the need for country-specific analyses to understand the underlying causes and potential mitigation strategies.

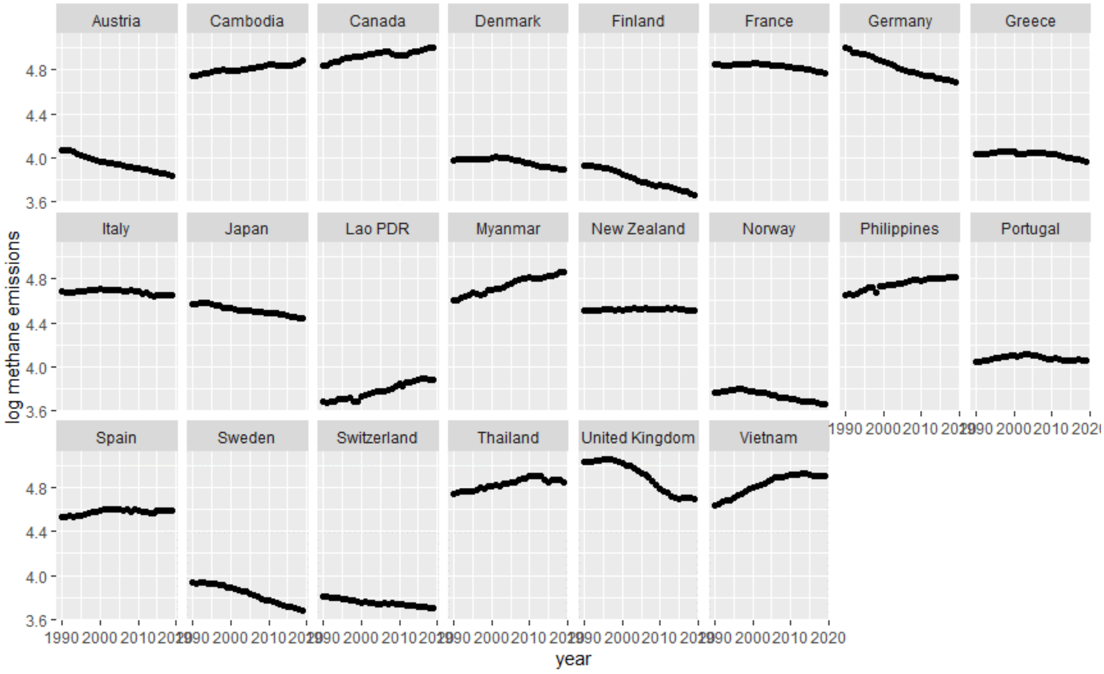

Figure 3. Scatter plots of year versus methane emissions (kt of equivalent) comparing 22 countries

The descriptive statistics of crop production and methane emissions are visualized in Figure 3. The image displays a series of time-series plots, each representing the log-transformed methane emissions for a range of countries from 1990 to 2019. The use of logarithmic scale on the y-axis is a common analytical approach to normalize data and highlight proportional changes over time. Each subplot is clearly labeled with the name of a country, and the x-axis marks the progression of years. From a high-level perspective, the plots exhibit varying trends: certain countries show a decrease in methane emissions over the examined period, others an increase, and some display relatively flat trends, indicating stable emissions over time. In analyzing the trends, it's evident that countries such as Austria, Finland, France, and Germany demonstrate a decline in methane emissions, which could suggest the effectiveness of environmental policies or advances in technology that lead to reduced emissions. On the other hand, countries like Cambodia, Myanmar, and Vietnam are on an upward trajectory, possibly reflecting increased industrial or agricultural activities that have not been offset by emissions control measures. The data for Japan, Norway, and Switzerland, with their minimal fluctuations, might indicate consistent emissions standards or a balance between emissions sources and mitigations.

To draw deeper conclusions from these graphs, additional context regarding the countries' economic development, environmental policies, and sector-specific activities would be required, as these factors significantly influence methane emission levels.

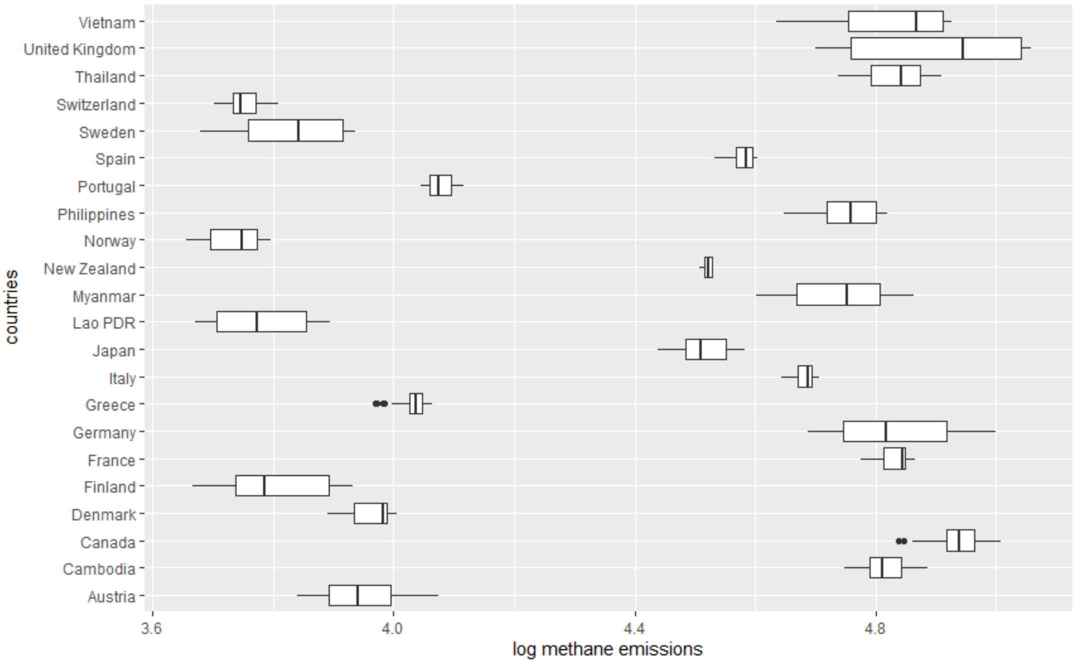

Figure 4. Box plot methane emissions compared with countries

The image presents a series of box plots illustrating the distribution of log-transformed methane emissions data for various countries in Figure 4. Each box plot aligns with a specific country, allowing for direct comparison of methane emission statistics among them. The central line within each box signifies the median value, offering a quick visual cue to the central tendency of the data. The spread of each box, demarcating the interquartile range, indicates the middle of the data, providing insight into the variability of emissions within each country. Notably, certain countries display outliers, depicted as individual points, suggesting emissions events that are significantly different from the typical range observed. Analyzing these plots, one can discern that countries such as Greece and Germany have high-value outliers, which could indicate sporadic instances of elevated methane emissions. On the other end of the spectrum, some countries exhibit a narrow interquartile range, reflecting more uniform methane emissions over the period studied. The use of a logarithmic scale is indicative of a focus on relative changes and can be especially useful in revealing patterns in data that span several orders of magnitude. This type of visualization is particularly adept at highlighting variations in data distribution and can be instrumental for researchers and policymakers in identifying trends, assessing policy effectiveness, and prioritizing areas for environmental intervention.

The findings from this section of the research elucidate distinct patterns of methane emissions related to agricultural practices, particularly crop production. Utilizing the FMM, the study categorizes the 22 countries into separate groups, each demonstrating a unique emission pattern. This model is adept at handling heterogeneous data, identifying subpopulations within the overall dataset, and fitting multiple distributions to the collected observations. By applying this methodology, the research discerns the underlying structures and categorizations that might not be visible in a more homogenized analysis. For each group of countries, the FMM assigns classification based on the similarity in their time-series patterns of methane emissions from agriculture. These classifications could be influenced by factors such as the intensity of agricultural practices, the type of crops produced, the scale of farming operations, and the implementation of methane reduction technologies or strategies. Developed countries might exhibit one pattern, potentially reflecting advanced agricultural techniques and stringent environmental policies, while developing nations may show another, possibly due to traditional farming methods and different regulatory frameworks. Therefore, taking into consideration the results of this section, the pattern can be broken up into groups of countries each of which classifies the pattern into a distinct category. This applies to 22 countries by the FMM.

3.2. Application results

In the realm of machine learning, particularly within the scope of this application, the robustness of a model is critically judged by its ability to generalize well to new, unseen data. A model that performs well not only on the training and test datasets but also on future data exemplifies good machine learning practice. The goal is to create a model that captures the underlying patterns and relationships within the data without overfitting to the noise or specificities of the training set. In this study, the finite mixture regression (FMR) model is calibrated to cluster the dataset of methane emissions and crop production indices from the 22 countries. The model's training involves an iterative process where the expectation-maximization (EM) algorithm is pivotal. This algorithm optimizes the log-likelihood of the FMR model, effectively enabling it to estimate the parameters that best fit the data. Upon completion of the learning process, the model iteration that yields the highest log-likelihood is selected as the optimal model for data clustering. This selected model is presumed to be the most representative of the complex, multidimensional data, encompassing the intricacies of methane emission patterns across varying agricultural contexts.

The methodology applies a two-pronged approach to data segmentation. The first method involves a random division of the dataset into training and testing sets, which is conducive to assessing the model's predictive power. The second method could involve a more systematic partitioning of data, perhaps chronological, where earlier years are used for training and later years for testing. This temporal split is particularly relevant for time-series data, ensuring that the model is tested against the most recent and, presumably, the most evolved patterns of emissions. Through this dual approach, the research ensures a comprehensive evaluation of the FMR model's performance. By effectively estimating the number of components (or clusters) and identifying the best-fit model, the study paves the way for more informed, data-driven decisions in environmental policy and agricultural management. Such a model, when applied successfully, can become an indispensable tool for understanding and mitigating the impact of agriculture on climate change.

In this application, a good machine learning solution should be able to fit new data. A model that has been trained should first work well on both the data that was used to train and test it and the data that will be available in the future. After learning, the particle with the highest log-likelihood is chosen, and its FMR is considered to be the best model for clustering the input data set. The mixture model and EM algorithm are used to find the number of components and the best fit model. Therefore, two methods are utilized to divide data into training and testing sets with FMR.

The BIC values are given between FXM and RDM in Table 1. Through the use of BIC, the number of Gaussian components can be determined. As a result, both FXM and RDM have the lowest BIC for two components ($k = 2$), which is -486.6049 for FXM and -474.5145 for RDM. The optimal solution corresponds to $k = 2$ when evaluated using BIC. Through the use of regression models, it is possible to further demonstrate that the model with $k = 2$ is appropriate. It is possible to say that the data on crop output and methane emissions can be divided into two distinct groups. In the next step, the coefficients of models should be analyzed in each component.

Table 1. The BIC values compared between FXM and RDM in different the number of components ($k = 1$, $k = 2$, and $k = 3$)

Model	$k = 1$	$k = 2$	$k = 3$
FXM	-413.8567	-486.6049	-467.0757
RDM	-408.7911	-474.5145	-461.5694

A finite mixture of two regressions fit data sets provides the optimum solution. The models with two components are fitted, and the parameters are computed using the data set in Table 2. The linear mixture model identifies the parameters ($\hat{\beta_{0k}}$, $\hat{\beta_{1k}}$, σ_k^2, and w_k) that need to have estimates calculated for them. In the case of FXM, the parameters of variance σ_1^2 and σ_2^2 are 0.043218 and 0.047868, while the parameters of mixing proportion w_1 and w_2 are 0.57 and 0.43, respectively. Then, a mixture of two linear models in this case is defined by (1) and (2).

Table 2. The coefficients of two models (FXM and RDM) in different two components ($k = 2$)

Coeffcient of models	FXM	RDM
Equation (1)		
$\hat{\beta}_0$	4.882786	4.943889
$\hat{\beta}_1$	-0.075938	-0.1180935
Equation (2)		
$\hat{\beta}_0$	3.421983	3.55212
$\hat{\beta}_1$	0.27306856	0.22692

$$\hat{y}_i = 4.882786 - 0.075938x_i \tag{1}$$

$$\hat{y}_i = 3.421983 - 0.273068x_i \tag{2}$$

Likewise, for the case of RDM, the parameters of variance σ_1^2 and σ_2^2 are 0.060890 and 0.083298, whereas the parameters of mixing proportion w_1 and w_2 are 0.55420 and 0.44579, respectively. In this model, the two linear models are written as (3) and (4).

$$\hat{y}_i = 4.943889 - 0.1180935x_i \tag{3}$$

$$\hat{y}_i = 3.555212 - 0.2269237x_i \tag{4}$$

After that, (1)-(4) are compared by evaluating predictive performance on training and test data set.

3.3. Accuracy evaluation

To evaluation the prediction performance of selected models, model performance is measured using the coefficient of determination (R^2), MAE, and RMSE. Tables 3 and 4 summarize the findings of the performance metrics for estimating the number of components $k = 2$ based on its relevance on the accuracy of the parameter estimation models.

Table 3. Performance metrics (R-Squared, MAE, and RMSE) of the FX models for training and test data

Equation	R^2	MAE	RMSE
Equation (1)			
Train	0.347770	0.014018	0.148211
Test	0.274452	0.014026	0.153503
Equation (2)			
Train	0.757507	0.05844	0.134729
Test	0.289443	0.13528	0.188063
All	0.332153		

Table 4. Performance metrics (R-squared, MAE, and RMSE) of the RD models for training and test data

Equation	R^2	MAE	RMSE
Equation (3)			
Train	0.321076	0.011077	0.151784
Test	0.313210	0.046520	0.114250
Equation (4)			
Train	0.664368	0.070090	0.143313
Test	0.538800	0.127640	0.185254
All	0.325803		

For training data performance, (2) and (4), ($R^2 = 0.76$ and R^2 = 0.66) are better than test data performance (R^2= 0.29 and R^2 = 0.54) in (1) and (3). All training data performances in FXM and RDM are stronger than test data while test data in (3) and (4) ($R^2 = 0.31$ and $R^2 = 0.54$) for RDMs are better than test data in (1) and (2) ($R^2 = 0.27$ and $R^2 = 0.29$) for FXM. Overall, the performance metrics of the FXM and RDM models are the same because their averages are so close. However, the model will use for prediction in the future and the test data are considered. Therefore, the best performance predictor model is RDM.

A detailed explanation of the individuals that make up different subpopulations within a sample can be obtained through the use of FMR. The illustration demonstrates that the error is relatively considerable, while the value of R^2 in the findings of the linear regression mixture analysis is quite low in some conditions but the models can be considered [34]. It is important to note that regression mixture modeling is not the same as multiple-group modeling when it comes to analyzing population diversity [35]. This means FMR can represent some distinct subgroups in the data. However, the RDM for selecting training data by a random process is a good choice for prediction because it considers the performance of test data. The models are good because the predictions are much closer to the actual values of the parameters, making forecasts more accurate [36].

3.4. Agricultural and methane emissions predictors

Table 5 shows that there are two different groups based on the crop production index and methane emissions data. In (1) and (2) list Cambodia, Canada, France, Germany, Italy, Japan, Myanmar, New Zealand, Philippines, Spain, Thailand, the United Kingdom, and Vietnam in Group 1 (Equation 3). Group 2 includes Austria, Finland, Greece, Laos, Norway, Portugal, Sweden, Switzerland, and Denmark. Comparatively, the methane emissions from Group 1 are quite considerable, but those from Group 2 are quite low. In Group 2, crop production has a small negative effect on methane emissions, but the emissions are still low compared to Group 1. On the other hand, methane emissions are almost stable in Group 1. In the past, agriculture had a big negative effect on methane emissions in most countries in Group 1. Because of this, they have tried to fix the problem [37], [38]. The effect of climate change on crop production is of particular interest, not only because crops are an important source of food all over the world, but also because recent intensification of crop production, especially in Asia, releases methane, which contributes to global warming. Only Laos has small emissions of methane, which is in Group 2 because of the small size of agricultural land. However, the pattern of methane emissions in Laos is rising.

In Group 1, which includes countries in Asia such as Thailand, Vietnam, Cambodia, and Philippines, as well as Japan, approximately 90 percent of the world's agricultural area, including rice fields, is located in monsoon Asian countries. Consequently, during the 1990s, CH_4 emissions from agricultural fields in these countries have been intensively measured [39]-[41]. Group 2 is comprised almost mainly of countries that are members of the European Union. This group is charged with implementing a policy that was recently passed and is intended to encourage activities that have no impact on the environment. Therefore, it would be beneficial for countries all around the world to severely reduce their CH_4 emissions.

Table 5. The two groups of 22 countries divided by the finite mixture regression model

Group 1	Group 2
Cambodia	Australia
Canada	Finland
France	Greece
Germany	Laos
Italy	Norway
Japan	Portugal
Myanmar	Sweden
New Zealand	Switzerland
Philippines	Denmark
Spain	
Thailand	
UK	
Vietnam	

4. CONCLUSIONS

To address the research question regarding relationship between crop production and methane emissions index, the FMR and EM algorithm successfully classified 22 countries into two groups based on crop production and methane emissions. This classification reveals that the choice of algorithm and the nature of the training data significantly impact predictive accuracy. The RDM was found to be more effective than the FXM in modeling these complex relationships. However, this prediction model depends on the type of data. This study highlighted the application of agricultural data to estimate emissions of methane in two different groups of countries with random training data. The research suggests that mixture models, like the Gaussian process, could further enhance prediction accuracy in time series data. This advancement is particularly pertinent in agricultural studies, where large and diverse datasets are vital. The study underscores the need for extensive data collection to improve model training and prediction capabilities. These findings have broad implications for environmental research, highlighting the critical role of advanced statistical methods in understanding and mitigating the impacts of agricultural practices on the environment. In the future work, the findings suggest that mixture models can improve predictions from nonparametric regression models in time series data such as Gaussian process. In addition to this, it is important to be concerned about the larger collection of training data.

REFERENCES

[1] H. Lee *et al.*, "Climate change 2023: synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change," 2023.

[2] A. Raihan, "A review of the global climate change impacts, adaptation strategies, and mitigation options in the socio-economic and environmental sectors," *Journal of Environmental Science and Economics*, vol. 2, no. 3, pp. 36–58, 2023, doi: 10.56556/jescae.v2i3.587.

[3] F. Bibi and A. Rahman, "An overview of climate change impacts on agriculture and their mitigation strategies," *Agriculture (Switzerland)*, vol. 13, no. 8, p. 1508, 2023, doi: 10.3390/agriculture13081508.

[4] S. Singh, R. S. Gill, V. S. Hans, and T. C. Mittal, "Experimental performance and economic viability of evacuated tube solar collector assisted greenhouse dryer for sustainable development," *Energy*, vol. 241, p. 122794, 2022, doi: 10.1016/j.energy.2021.122794.

[5] Q. Zhang, J. Xiao, J. Xue, and L. Zhang, "Quantifying the effects of biochar application on greenhouse gas emissions from agricultural soils: a global meta-analysis," *Sustainability (Switzerland)*, vol. 12, no. 8, p. 3436, 2020, doi: 10.3390/SU12083436.

[6] H. Pathak, "Impact, adaptation, and mitigation of climate change in Indian agriculture," *Environmental Monitoring and Assessment*, vol. 195, no. 1, p. 52, 2023, doi: 10.1007/s10661-022-10537-3.

[7] D. G. Oladipo *et al.*, "Short-term assessment of nitrous oxide and methane emissions on a crop yield basis in response to different organic amendment types in sichuan basin," *Atmosphere*, vol. 12, no. 9, p. 1104, 2021, doi: 10.3390/atmos12091104.

[8] Y. Huang *et al.*, "Greenhouse gas emissions and crop yield in no-tillage systems: a meta-analysis," *Agriculture, Ecosystems and Environment*, vol. 268, pp. 144–153, 2018, doi: 10.1016/j.agee.2018.09.002.

[9] M. Saunois, R. B. Jackson, P. Bousquet, B. Poulter, and J. G. Canadell, "The growing role of methane in anthropogenic climate change," *Environmental Research Letters*, vol. 11, no. 12, p. 120207, 2016, doi: 10.1088/1748-9326/11/12/120207.

[10] H. Schaefer, "On the causes and consequences of recent trends in atmospheric methane," Current Climate Change Reports, vol. 5, no. 4, pp. 259–274, 2019, doi: 10.1007/s40641-019-00140-z.

[11] M. Agovino, M. Casaccia, M. Ciommi, M. Ferrara, and K. Marchesano, "Agriculture, climate change and sustainability: the case of EU-28," *Ecological Indicators*, vol. 105, pp. 525–543, 2019, doi: 10.1016/j.ecolind.2018.04.064.

[12] P. Thongnim, "Parametric and non-parametric mixture of regression models for agricultural economic data," University of Leicester, 2022.

[13] P. Thongnim, V. Yuvanatemiya, and P. Srinil, "Smart agriculture: transforming agriculture with technology," in *Communications in Computer and Information Science*, 2024, vol. 1911 CCIS, pp. 362–376, doi: 10.1007/978-981-99-7240-1_29.

[14] M. K. Uçar, M. Nour, H. Sindi, and K. Polat, "The effect of training and testing process on machine learning in biomedical datasets," *Mathematical Problems in Engineering*, vol. 2020, 2020, doi: 10.1155/2020/2836236.

[15] A. A. Warsame and A. H. Abdi, "Towards sustainable crop production in Somalia: examining the role of environmental pollution and degradation," *Cogent Food and Agriculture*, vol. 9, no. 1, p. 2161776, 2023, doi: 10.1080/23311932.2022.2161776.

[16] R. A. Genedy and J. A. Ogejo, "Using machine learning techniques to predict liquid dairy manure temperature during storage," *Computers and Electronics in Agriculture*, vol. 187, p. 106234, 2021, doi: 10.1016/j.compag.2021.106234.

[17] A. E. Danganan, A. M. Sison, and R. P. Medina, "OCA: overlapping clustering application unsupervised approach for data analysis," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 3, pp. 1471–1478, 2019, doi: 10.11591/ijeecs.v14.i3.pp1471-1478.

[18] P. Thongnim, V. Yuvanatemiya, P. Srinil, and T. Phukseng, "An exploration of emission data visualization in Southeast Asian Countries," in *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2023*, 2023, pp. 160–165, doi: 10.1109/ICITACEE58587.2023.10277282.

[19] J. K. Basak *et al.*, "Regression analysis to estimate morphology parameters of pepper plant in a controlled greenhouse system," *Journal of Biosystems Engineering*, vol. 44, no. 2, pp. 57–68, 2019, doi: 10.1007/s42853-019-00014-0.

[20] J. K. Basak, E. Arulmozhi, B. E. Moon, A. Bhujel, and H. T. Kim, "Modelling methane emissions from pig manure using statistical and machine learning methods," *Air Quality, Atmosphere and Health*, vol. 15, no. 4, pp. 575–589, 2022, doi: 10.1007/s11869-022-01169-0.

[21] O. V. Lee *et al.*, "A malicious URLs detection system using optimization and machine learning classifiers," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 17, no. 3, pp. 1210–1214, 2020, doi: 10.11591/ijeecs.v17.i3.pp1210-1214.

[22] M. Ali and S. Shatabda, "A data selection methodology to train linear regression model to predict bitcoin price," in *2020 2nd International Conference on Advanced Information and Communication Technology, ICAICT 2020*, 2020, pp. 330–335, doi: 10.1109/ICAICT51780.2020.9333525.

[23] M. Norel, K. Krawiec, and Z. W. Kundzewicz, "Machine learning modeling of climate variability impact on river runoff," *Water (Switzerland)*, vol. 13, no. 9, p. 1177, 2021, doi: 10.3390/w13091177.

[24] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annual Review of Statistics and Its Application*, vol. 6, pp. 355–378, 2019, doi: 10.1146/annurev-statistics-031017-100325.

[25] B. Muthén, "Second-generation structural equation modeling with a combination of categorical and continuous latent variables: new opportunities for latent class–latent growth modeling," *New methods for the analysis of change*, pp. 291–322, 2004, doi: 10.1037/10409-010.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, " Maximum Likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[27] X. Bai, W. Yao, and J. E. Boyer, "Robust fitting of mixture regression models," *Computational Statistics and Data Analysis*, vol. 56, no. 7, pp. 2347–2359, 2012, doi: 10.1016/j.csda.2012.01.016.

[28] H. Hu, W. Yao, and Y. Wu, "The robust EM-type algorithms for log-concave mixtures of regression models," *Computational Statistics and Data Analysis*, vol. 111, pp. 14–26, 2017, doi: 10.1016/j.csda.2017.01.004.

[29] V. Melnykov and R. Maitra, "Finite mixture models and model-based clustering," *Statistics Surveys*, vol. 4, no. none, pp. 80–116, 2010, doi: 10.1214/09-SS053.

[30] A. Khalili and S. Lin, "Regularization in finite mixture of regression models with diverging number of parameters," *Biometrics*, vol. 69, no. 2, pp. 436–446, 2013, doi: 10.1111/biom.12020.

[31] Q. Liu, M. A. Charleston, S. A. Richards, and B. R. Holland, "Performance of akaike information criterion and bayesian information criterion in selecting partition models and mixture models," *Systematic Biology*, vol. 72, no. 1, pp. 92–105, 2023, doi: 10.1093/sysbio/syac081.

[32] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 2007, doi: 10.1214/aos/1176344136.

[33] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, pp. 1–24, Jul. 2021, doi: 10.7717/PEERJ-CS.623.

[34] P. K. Ozili, "The acceptable R-square in empirical modelling for social science research," in *Social Research Methodology and Publishing Results: A Guide to Non-Native English Speakers*, IGI Global, 2023, pp. 134–143.

[35] C. S. Ding, "Using regression mixture analysis in educational research," *Practical Assessment, Research and Evaluation*, vol. 11, no. 11, p. 11, 2006.

[36] R. Medar, V. S. Rajpurohit, and B. Rashmi, "Impact of training and testing data splits on accuracy of time series forecasting in machine learning," in *2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017*, 2017, pp. 1–6, doi: 10.1109/ICCUBEA.2017.8463779.

[37] A. E. Milne *et al.*, "Analysis of uncertainties in the estimates of nitrous oxide and methane emissions in the UK's greenhouse gas inventory for agriculture," *Atmospheric Environment*, vol. 82, pp. 94–105, 2014, doi: 10.1016/j.atmosenv.2013.10.012.

[38] M. Azhar Khan, M. Zahir Khan, K. Zaman, and L. Naz, "Global estimates of energy consumption and greenhouse gas emissions," *Renewable and Sustainable Energy Reviews*, vol. 29, pp. 336–344, 2014, doi: 10.1016/j.rser.2013.08.091.

[39] X. Yan, H. Akiyama, K. Yagi, and H. Akimoto, "Global estimations of the inventory and mitigation potential of methane emissions from rice cultivation conducted using the 2006 intergovernmental panel on climate change guidelines," *Global Biogeochemical Cycles*, vol. 23, no. 2, 2009, doi: 10.1029/2008GB003299.

[40] R. Matthews and R. Wassmann, "Modelling the impacts of climate change and methane emission reductions on rice production: a review," *European Journal of Agronomy*, vol. 19, no. 4, pp. 573–598, 2003, doi: 10.1016/S1161-0301(03)00005-4.

[41] R. W. Howarth, "Methane and climate change," *Environmental Impacts from the Development of Unconventional Oil and Gas Reserves; Stolz, J., Bain, D., Griffin, M., Eds*, pp. 132–149, 2021.

BIOGRAPHIES OF AUTHORS

Pattharaporn Thongnim is a distinguished figure in her academic and professional capacities, notably as a lecturer at Burapha University in Thailand. She earned her Ph.D. in Statistics and Data Science from the University of Leicester, UK, where her focus was on Gaussian Mixture regression models pertinent to agricultural economic data, integrating elements of data science, statistics, and machine learning. Her projects reflect a keen interest in the interplay between climate data and agriculture, employing advanced techniques like Gaussian process mixture regression models for analyzing agricultural economies in Asia and Europe, and utilizing drones for innovative agricultural applications. Thongnim's academic credentials are impressive, encompassing a Master's in Statistics from Chulalongkorn University and a Bachelor's in Mathematics from Mahidol University. Her career is characterized by her commitment to merging cutting-edge statistical and machine learning methods with climate and agricultural data, demonstrating a sophisticated understanding of the environmental factors that influence agricultural economics. She can be contacted at email: pattharaporn@buu.ac.th.

Ekkapot Charoenwanit is a computer scientist with a B.Eng. in Computing and an M.Sc. in Advanced Computing from Imperial College London, UK, as well as a Ph.D. in Computer Science from RWTH Aachen University, Germany. Currently, he serves as a lecturer at the Electrical and Computer Engineering Programme (ECE) at the The Sirindhorn International Thai-German Graduate School of Engineering (TGGS), King Mongkut's University of Technology North Bangkok (KMUTNB), Bangkok, Thailand. His research work focuses on algorithmic differentiation, high-perfomance computing, data science and machine learning. He can be contacted at email: ekkapot.c@tggs.kmutnb.ac.th.