RESEARCH ARTICLE • OPEN ACCESS

Modeling and Optimizing State-Based Electricity Consumption Distributions in Dodoma Region, Tanzania Using Hidden Markov Models

Eliasi M. Jeremiah, Ramkumar T. Balan, and Jairos K. Shinzeh

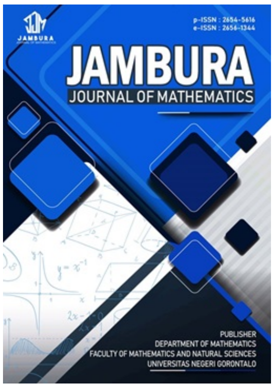
JAMBURA
JOURNAL OF MATHEMATICS

Volume 8, Issue 2, Pages 169–187, August 2026

Received 22 August 2025, Revised 27 December 2025, Accepted 8 January 2026, Published 20 July 2026

To Cite this Article : E. M. Jeremiah, R. T. Balan, and J. K. Shinzeh,“Modeling and Optimizing State-Based Electricity Consumption Distributions in Dodoma Region, Tanzania Using Hidden Markov Models”, *Jambura J. Math*, vol. 8, no. 2, pp. 169–187, 2026, *https://doi.org/10.37905/jjom.v8i2.34047*

JOURNAL INFO • JAMBURA JOURNAL OF MATHEMATICS



Homepage	:	http://ejurnal.ung.ac.id/index.php/jjom/index
Journal Abbreviation	:	Jambura J. Math.
Frequency	:	Biannual (February and August)
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/jjom
Online ISSN	:	2656-1344
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjom/oai
Google Scholar ID	:	iWLjgaUAAAAJ
Email	:	info.jjom@ung.ac.id

Jambura Journal of Mathematics, Volume 8, Issue 2, Pages 169–187, August 2026 *https://doi.org/10.37905/jjom.v8i2.34047*

Research Article

Check for updates

Modeling and Optimizing State-Based Electricity Consumption Distributions in Dodoma Region, Tanzania Using Hidden Markov Models

Eliasi M. Jeremiah[1,*], **Ramkumar T. Balan**[2], **Jairos K. Shinzeh**[2]

[1]*Department of Mathematics and Statistics, University of Mzumbe, Morogoro, United Republic of Tanzania*
[2]*Department of Mathematics and Statistics, The University of Dodoma, Dodoma, United Republic of Tanzania*

ARTICLE HISTORY

Received 22 August 2025
Revised 27 December 2025
Accepted 8 January 2026
Published 20 July 2026

KEYWORDS

Hidden Markov Model
Hidden states
Akaike Information Criterion
Bayesian Information Criterion
Homoscedasticity
Normal Distribution

ABSTRACT. *The aim of this study is to model and optimize state-based Electricity Consumption Distributions in Dodoma Region, Tanzania using Hidden Markov Models. The research is specifically aimed at the identification of hidden consumption states, and the determination of the best number of hidden states to enhance the state distributions identification. A quantitative approach was utilized based on monthly TANESCO electricity demand time series data for the years 2010-2025. The estimation involved the calculation of state transition probabilities, model diagnostics testing, and evaluation of forecasting performance based on different hidden state configurations. Model selection was based on extensively documented statistical criteria, namely the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Empirical results revealed that a three-state HMM achieved the best performance with the lowest AIC (5492.480) and BIC (5707.104) values relative to models of two to ten states. The diagnostic tests concluded that segmenting the series into latent states improved statistical attributes such as normality, homoscedasticity, and stationarity that otherwise failed in the original unsegmented data. For instance, State 1 residuals were homoscedastic (Breusch-Pagan $p = 0.22$), and State 3 demonstrated stationarity (ADF $p \approx 0.01$), enhancing interpretability and model fit. These findings show that three states bring optimal fit for prediction, and each state fulfils the assumption of homoscedasticity, as all states follow a normal distribution. The study recommends that energy policymakers and utility providers integrate HMM-based forecasting approaches to improve decision-making in regions with dynamic and complex electricity usage patterns.**The aim of this study is to model and optimize state-based Electricity Consumption Distributions in Dodoma Region, Tanzania using Hidden Markov Models. The research is specifically aimed at the identification of hidden consumption states, and the determination of the best number of hidden states to enhance the state distributions identification. A quantitative approach was utilized based on monthly TANESCO electricity demand time series data for the years 2010-2025. The estimation involved the calculation of state transition probabilities, model diagnostics testing, and evaluation of forecasting performance based on different hidden state configurations. Model selection was based on extensively documented statistical criteria, namely the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Empirical results revealed that a three-state HMM achieved the best performance with the lowest AIC (5492.480) and BIC (5707.104) values relative to models of two to ten states. The diagnostic tests concluded that segmenting the series into latent states improved statistical attributes such as normality, homoscedasticity, and stationarity that otherwise failed in the original unsegmented data. For instance, State 1 residuals were homoscedastic (Breusch-Pagan $p = 0.22$), and State 3 demonstrated stationarity (ADF $p \approx 0.01$), enhancing interpretability and model fit. These findings show that three states bring optimal fit for prediction, and each state fulfils the assumption of homoscedasticity, as all states follow a normal distribution. The study recommends that energy policymakers and utility providers integrate HMM-based forecasting approaches to improve decision-making in regions with dynamic and complex electricity usage patterns.*

1. Introduction

The growing sophistication of global energy systems, driven by increasing electrification, climate change mitigation policies, and smart grid technologies, has heightened the need for accurate electricity consumption forecasting [1]. This need is particularly acute in developing economies like Tanzania, where electricity consumption is projected to more than double by 2030 [2]. In this context, reliable forecasting is indispensable for achieving optimal load balancing, guiding infrastructure investment, and ensuring energy security [3]. Unique local drivers including rural electrification programs, rapid urbanization, population growth, and hydroelectric generation variability create irregular and multi-patterned consumption behaviors that are less common in stable, high-income grids [4]. These complexities necessitate forecasting models capable of adapting to and capturing

*Corresponding Author.

Email : *eliasijrmh@mail.com* (E. M. Jeremiah)
Homepage : http://ejurnal.ung.ac.id/index.php/jjom/index / E-ISSN : 2656-1344

these distinct latent patterns.

Among various forecasting techniques, Hidden Markov Models (HMMs) offer a powerful framework for modelling time series data characterized by underlying, unobservable states [5, 6]. Unlike traditional models such as ARIMA that rely on assumptions of linearity and stationarity, HMMs operate on the principle that a system transitions between a finite set of hidden states such as peak demand, off peak, and outage periods each emitting observable data according to a specific probability distribution [5]. This structure makes them exceptionally suited for capturing the non-linear dependencies and regime shifts inherent in electricity consumption, which are influenced by behavioral patterns, appliance usage, and temporal cycles [7]. Despite their promise, the application of HMMs to electricity forecasting, especially in Sub-Saharan Africa, remains underexplored compared to high-income economies [8, 9].

A central methodological challenge hindering the broader application of HMMs is the determination of the optimal number of hidden states [10]. An insufficient number of states may lead to underfitting, obscuring important variations in consumption, while an excessive number can cause overfitting, reduce the model's generalizability and increase sensitivity to noise [11, 12]. This trade-off has profound implications for forecast accuracy and real-world usability, particularly in data-scarce environments with infrastructure limitations.

The literature provides little consensus on a systematic approach for selecting the number of hidden states. Researchers often rely on heuristics or assumptions without rigorously testing performance across different state configurations [13, 14]. While information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are established tools for balancing model fit against complexity [6], their behavior in the specific context of electricity consumption in developing economies has not been thoroughly investigated [10].

Several empirical studies demonstrate the successful application of HMMs for energy forecasting. Electricity demand classification into distinct states using HMM-based approaches has been shown to improve predictive accuracy by accounting for time-evolving usage patterns [15]. Methodological advances further emphasize the importance of information criteria such as AIC and BIC in selecting models that balance complexity and interpretability [6, 16]. Recent developments also include hierarchical and multi-scale HMMs capable of handling consumption data across multiple temporal resolutions, thereby expanding the applicability of the model [1, 14]. Comparative studies further support the superiority of HMMs over conventional machine learning approaches in capturing temporal dependencies within complex time series data [6].

The HMM framework is built on the Markovian assumption that future system states depend only on the present state, allowing effective modeling of latent consumption regimes [5]. Model selection theory, operationalized through AIC and BIC, provides a principled method for identifying models that best explain observed data without unnecessary complexity [6, 10]. This study examines the impact of varying the number of hidden states on the predictive accuracy of HMMs for electricity consumption in Tanzania. By leveraging AIC, BIC, and cross validation on real consumption data, the study seeks to establish a methodological protocol for state optimization, thereby providing utility providers and policymakers with a data driven tool for improving load forecasting, investment planning, and grid resilience [16].

Accurate electricity consumption forecasting is essential for energy planning, grid stability, and development, but remains especially challenging in rapidly evolving regions such as Dodoma, Tanzania. Demand exhibits non-linear and volatile patterns driven by urbanization, electrification initiatives, climate variability, and socioeconomic factors. Conventional models such as ARIMA are limited by assumptions of stationarity and linearity, making them less suitable for such dynamics [17, 18]. Hidden Markov Models offer a robust alternative by representing consumption data as emissions from hidden states with distinct probability distributions [5, 6]. Their effectiveness depends critically on selecting an optimal number of hidden states and accurately estimating state-specific parameters. In data-scarce environments such as Dodoma, forecast errors can lead to poor planning, financial losses, and unreliable electricity supply. By rigorously evaluating alternative state configurations, this research aims to establish a scientific basis for building accurate and reliable HMM based electricity forecasting models.

2. Methods

This study employed a quantitative research approach using electricity consumption data obtained from TANESCO for the period 2010–2025. A Hidden Markov Model was applied to identify temporal dynamics and latent state-based patterns underlying electricity consumption in the Dodoma region. To determine the optimal number of hidden states, the model was iteratively trained using varying state configurations, with the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) employed as model selection metrics [6, 10]. The state configuration yielding the lowest information criterion values was selected as optimal. Subsequently, state specific probability distributions were estimated to characterize distinct electricity consumption regimes, providing a comprehensive understanding of consumption behavior across different usage patterns.

2.1. *Model Formulation*

In this subsection, we introduce the step-by-step process on how many states are available in electricity consumption, how many optimal states are obtained, and the probability distribution for each state as shown in equations.

$$X^T = (X_1, X_2, X_3, ..., X_\mathrm{T}),$$

$$\mathrm{P}(X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_0 = x_0) = \mathrm{P}(X_t = x_t \mid X_{t-1} = x_{t-1}),$$

In a first-order Markov chain, the probability of moving to the next state depends only on the current state. This is the most commonly used form of Markov chain in time series modelling and Hidden Markov Models.

$$\mathrm{P}(X_\mathrm{t} = x_\mathrm{t} | X_{\mathrm{t}-1}, X_{\mathrm{t}-2}, ..., X_0) = \mathrm{P}(X_\mathrm{t} = x_\mathrm{t} | X_{\mathrm{t}-1}),$$

In a second-order Markov chain, the current state depends on the two previous states

$$\mathrm{P}(X_\mathrm{t} = x_\mathrm{t} | X_{\mathrm{t}-1}, X_{\mathrm{t}-2}, ..., X_0 = x_0) = \mathrm{P}(X_\mathrm{t} = x_\mathrm{t} | X_{\mathrm{t}-1}, X_{\mathrm{t}-2}),$$

A third-order Markov chain depends on the last three states

$$\begin{aligned}&\mathrm{P}(X_t = x_t \mid X_{t-1}, X_{t-2}, \ldots, X_0 = x_0)\\ &= \mathrm{P}(X_t = x_t \mid X_{t-1}, X_{t-2}, X_{t-3}).\end{aligned}$$

Hidden Markov Model is a statistical model that describes a time-varying system with a non-observable (hidden) underlying Markov chain and an observable process whose dynamics are specified by the hidden states. The system is in a hidden state C_t, at every time point t, and produces an observable outcome X_t generated by a conditional probability distribution under the hidden state [6].

$$C_t = (c_1, c_2, c_3, ..., c_{\mathrm{T}}).$$

A transition probability is the probability of transitioning from one hidden state to another at the next time step. Mathematically, the transition probability γ_{ij} is given by equation below, which indicates the likelihood that the system moves from state i to state j at the next time point.

$$\gamma_{\mathrm{ij}} = \mathrm{P}(C_{\mathrm{t}+1} = j \mid C_{\mathrm{t}} = i).$$

Once the model has been trained, algorithms such as the Viterbi algorithm can be employed in order to determine the most likely sequence of hidden states at any given time. This assists us in making sense of patterns within the data instances of low, medium, or high electricity usage without ever directly observing the hidden states [6].

$$\Gamma = [\gamma_{\mathrm{ij}}] = \mathrm{P}(C_{\mathrm{t}+1} = j \mid C_{\mathrm{t}} = i),$$

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1\mathrm{m}} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \cdots & \gamma_{2\mathrm{m}} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \gamma_{\mathrm{m}1} & \gamma_{\mathrm{m}2} & \gamma_{\mathrm{m}3} & \cdots & \gamma_{\mathrm{mm}} \end{pmatrix}$$

for $i = 1, 2 \ldots m$ and $j = 1, 2 \ldots m$.

For a Hidden Markov chain with m states, all transition probabilities can be represented in a transition probability matrix (Γ) in which each row sum is 1. These probabilities are organized into a transition probability matrix (Γ), which is a matrix $m \times m$ (for m hidden states), where each element γ_{ij} satisfies $0 \leq \gamma_{\mathrm{ij}} \leq 1$ and each row sums to 1. Each row in the matrix represents the current state, while each column represents the possible next state.

$$\sum_{j=1}^{m} \gamma_{\mathrm{ij}} = 1 \qquad\qquad for\ all\ i.$$

Initial State Probability Vector (δ), was defined as the probability distribution over the initial states of the hidden Markov chain at time $t = 1$. It provides the prior belief about the state of the system before any observations are made. where $\delta_i = P(C_i = i)$, which means the probability that the initial hidden state is i and $\sum_{i=1}^{\mathrm{m}} \delta_{\mathrm{ij}} = 1$ because the system must start in one of the states.

$$\delta = [\delta_1, \delta_2, \delta_3, \ldots, \delta_{\mathrm{m}}].$$

Emission Probability Matrix (B) describes the relationship between the hidden states and the observed outputs. Specifically, it defines the probability of observing a particular symbol (or value) given the hidden state of the system.

$$B = [b_{\mathrm{i}}(x_{\mathrm{t}})] = \mathrm{P}(X_{\mathrm{t}} = x_{\mathrm{t}} \mid C_{\mathrm{t}} = i),$$

where $i = 1, 2 \ldots m$ and $t = 1, 2 \ldots k$.

$$B = \begin{pmatrix} b_1(x_1) & b_1(x_2) & \cdots & b_1(x_k) \\ b_2(x_1) & b_2(x_2) & \cdots & b_2(x_k) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ b_{\mathrm{m}}(x_1) & b_{\mathrm{m}}(x_2) & \cdots & b_{\mathrm{m}}(x_m) \end{pmatrix}$$

Hidden Markov Model equation for the joint probability of observations and states

$$\mathrm{P}(x) = \sum_{t=1}^{T} \mathrm{P}(X_{\mathrm{t}} \mid C_{\mathrm{t}}) \mathrm{P}(C_{\mathrm{t}}),$$

HMM is capable of modelling temporal dependence and predicting future states from observed sequences, which is highly appropriate to model electricity consumption data [12, 19]. The Hidden Markov Chain (HMM) model was employed to analyse the electricity consumption data, a highly suitable model for time series forecasting when data violate some of the forecasting assumptions, including linearity and stationarity [14]. The observation sequence consists of the actual observed values (or symbols) at each time step, which are generated by the hidden states [20].

The use of the Baum-Welch algorithm (a form of Expectation-Maximisation) was to estimate the model parameters (transition probabilities Γ, emission parameters B, and initial state distribution δ).

Transition probabilities γ_{ij} is given by

$$\gamma_{\mathrm{ij}} = \frac{\sum_{\mathrm{t}}^{T-1} \gamma(\mathrm{t}, c_{\mathrm{i}}) \xi(\mathrm{t}, c_{\mathrm{i}}, c_{\mathrm{j}})}{\sum_{\mathrm{t}}^{T-1} \gamma(\mathrm{t}, s_{\mathrm{i}})},$$

Emission probabilities for continuous variables

$$b_{\mathrm{ij}} = \frac{\sum_{\mathrm{t}}^{T} \gamma(\mathrm{t}, c_{\mathrm{i}}) \cdot \Pi(X_{\mathrm{t}} = x)}{\sum_{\mathrm{t}}^{T} \gamma(\mathrm{t}, c_{\mathrm{i}})},$$

where $\gamma(t, c_i)$ is the probability of being in state c_i at time t, and $\xi(t, c_i, c_j)$ is the probability of transitioning from state c_i to state c_j at time t, followed by the use of the Viterbi algorithm to find the most likely sequence of hidden states.

$$\delta_{\mathrm{t}}(c_{\mathrm{i}}) = \max_{c_1, c_2, c_3, \ldots, c_{\mathrm{t}-1}} \mathrm{P}(x_1, x_2, x_3, \ldots, x_{\mathrm{t}}, c_{\mathrm{t}} = c_{\mathrm{i}}).$$

Evaluating multiple models with different numbers of hidden states m, and choosing the one that maximizes model accuracy by using Log-Likelihood for each model

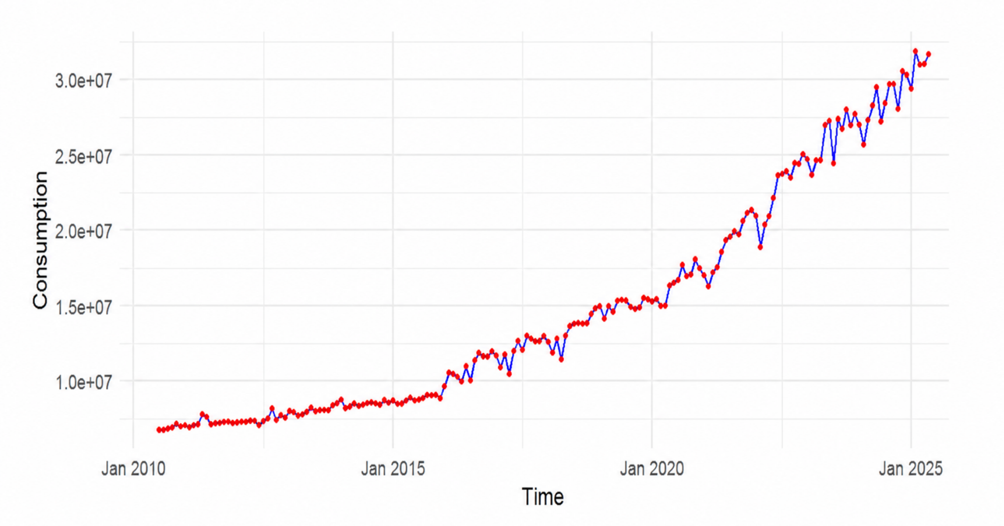


Figure 1. Electricity Consumption Over Time.

$$\log \mathrm{L}(x \mid \lambda) = \sum_{t=1}^{T} \log(\alpha_{\mathrm{T}}(i)),$$

where $\lambda = \{\mathrm{A}, \mathrm{B}, \Pi\}$ is the model parameter set, and $\alpha_{\mathrm{T}}(i)$ represents the forward variable (probability of being in state i at time t given the previous observations), and P represents the probability function.

Bayesian Information Criterion (BIC), Use of BIC to compare models with different numbers of states. The formula for BIC is:

$$\mathrm{BIC} = -2 \log \mathrm{L}(\lambda) + k \log \mathrm{T},$$

where:
$\mathrm{L}(\lambda)$ is the likelihood of the model, k is the number of parameters in the model, and T is the number of observations. The model with the lowest BIC is preferred.
Model the probability distributions of electricity consumption in each hidden state within the HMM framework, for each hidden state C_i, we have to model the electricity consumption distribution. Since the observed variable (electricity consumption) is continuous, and without loss of generality, normal (Gaussian) distribution. The assumption is valid since each state contains at least 30 observations, and variance is a constant within each state.

$$\mathrm{P}(X_{\mathrm{t}} \mid C_{\mathrm{i}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x_{\mathrm{i}}-\mu}{\sigma_{\mathrm{i}}}\right)^2},$$

$$\text{where, } -\infty < x < +\infty, \quad \sigma > 0, \quad -\infty < \mu < +\infty,$$

where μ_i is the mean of the normal distribution for the state C_i, and σ_i^2 is the variance of the normal distribution for the state C_i, followed by the use of the Baum-Welch algorithm to estimate the parameters μ_i and σ_i for each hidden state.

Maximum likelihood estimation for the mean μ_{i} and variance σ^2 be performed as follows.

$$\mu_{\mathrm{i}} = \frac{\sum_{\mathrm{t}}^{T} \gamma(\mathrm{t}, c_{\mathrm{i}}) \cdot x_{\mathrm{t}}}{\sum_{t=1}^{T} \gamma(\mathrm{t}, c_{\mathrm{i}})},$$

$$\sigma_{\mathrm{i}}^2 = \frac{\sum_{\mathrm{t}}^{T} \gamma(\mathrm{t}, c_{\mathrm{i}}) \cdot (x_{\mathrm{t}} - \mu_{\mathrm{i}})^2}{\sum_{\mathrm{t}}^{T} \gamma(\mathrm{t}, c_{\mathrm{i}})}.$$

After fitting the HMM, a set of Gaussian distributions for each hidden state s_{i} with corresponding means and variances μ_{i} sigma σ^2. These distributions describe the electricity consumption behaviour for each hidden state.

3. Results and Discussion

The July 2010 to May 2025 monthly electricity consumption in Figure 1 reveal a sharp and chronic rising trend. Starting at around 7 million units in 2010, electricity consumption slowly picked up pace, heading towards over 30 million units in 2025. In the first half of the series (2010–2015), the data suggest relatively stable month-to-month changes. However, beginning around 2016, the size of the changes grew incredibly.

In exploratory distribution fitting into electricity consumption data analysis, among the significant observations was that electricity consumption joint distribution is lognormal. The implication of the same is that a logarithm of the consumption measures is normally distributed, the original data is positively skewed, something which is typical in energy demand patterns due to outlier values during high usage [21]. X_{t} denotes electricity consumption, then $X_{\mathrm{t}} \sim Lognormal(\mu, \sigma^2)$ where u and σ^2 are the mean and variance of the parent normal distribution. Identification and validation of this model were done by applying the *fitdistrplus* package in R to compare various distributions, normal, gamma, and lognormal on the basis of their fitted density plots and information criteria. The density comparison and goodness-of-fit statistics such as Akaike Information Criterion and Bayesian Information Criterion revealed that the minimum values for AIC and BIC were for lognormal distribution, which, by default, is the best fit distribution as it shown in equation 28. With a μ (meanlog) = 16.3913405 and σ (sdlog) = 0.4784187.

$$f(X_{\mathrm{t}}, \mu, \sigma) = \frac{1}{X_{\mathrm{t}} \times 1.198237} \exp\left(-\frac{(\ln X_{\mathrm{t}} - 16.39134)^2}{0.457899}\right),$$

where, $X_{\mathrm{t}} > 0$.

3.1. *Density plot*

The Histogram and plot of theoretical densities in Figure 2 superimpose observed electricity consumption data distribution

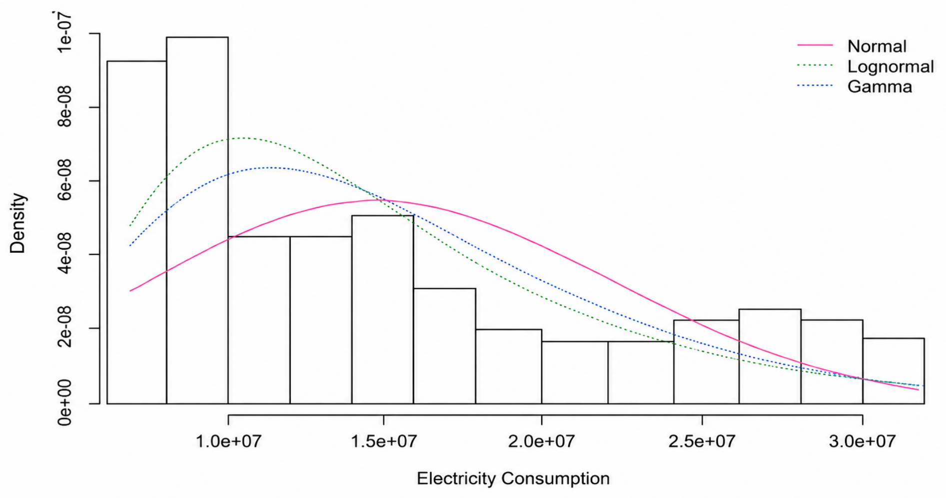


Figure 2. Fitted Distributions for Electricity Consumption.

with three fitted theoretical distributions, normal (red solid line), lognormal (green dashed line), and gamma (blue dotted line). Graphically, the lognormal distribution also has the best fit for the data, specifically the peak and skewness in the left tail and smoothly sloping right tail.

The probability distribution of the electricity consumption as it shown in Table 1, is a comparison of the goodness-of-fit of three statistical distributions, normal, lognormal, and gamma, based on log-likelihood, AIC, and BIC I. Log-likelihood values indicate how well a distribution fits the observed data, with higher (less negative) values indicating better fits. Lognormal distribution also has the lowest AIC (6116.138) and BIC (6122.512) values, in contrast, the Normal distribution performs the worst across all three metrics, inaccurately capturing the skewed and heavy-tailed nature of electricity consumption. The gamma distribution, while superior to normal, does not attain the goodness-of-fit obtained by the lognormal.

Table 1. Distribution of the Data.

Distribution	Log-Likelihood	AIC	BIC
Normal	-3083.982	6171.963	6178.338
Lognormal	-3056.069	6116.138	6122.512
Gamma	-3061.45	6126.901	6133.276

This expression allows us to use either hard or soft decoding techniques to allocate every observation to some hidden state. For the case of using the Viterbi algorithm, the optimum state sequence $\hat{C}_t$ is found by computing the maximum probability of each state at time t using $\delta_{\mathrm{t}}(i) = \max_j \mathrm{P}[\delta_{\mathrm{t}-1}(j)\cdot\gamma_{\mathrm{j,i}}]\cdot b_{\mathrm{i}}(X_t)$ where $\delta_{\mathrm{t}}(i)$ is the highest probability of any state sequence ending in state i at time t [22]. Alternatively, a soft-decoding approach approximates the posterior probability

$$P(C_i = \mathrm{i} \mid X_t) = \frac{\alpha_{\mathrm{t}} \cdot \beta_{\mathrm{t}}}{P(X_t)},$$

with α_{t} and β_{t} being the forward and backward probabilities, respectively, thereby illustrating a probabilistic assignment of every observation to the possible states [18, 23]. Not only does the method identify the most likely sequence of hidden states, but it quantifies the state assignment uncertainty, thereby illustrating firm insight into electricity consumption regime-switching dynamics [16].

3.2. Hidden Markov Model

Hidden Markov Models find widespread use in time-series modelling with observed data assumed to be generated by a process controlled by an unseen sequence of states. For electric consumption data, HMMs are particularly useful to model regime-switching behaviour, shift from peak to off-peak usage periods, by mapping each regime to a hidden state [17, 22]. These models are particularly useful in scenarios where system dynamics cannot be observed but can be estimated from the observable data stream a common scenario for electricity demand forecasting [20]. Mathematically, an HMM is defined by a set of parameters $\lambda = (\Pi, \Gamma, \mathrm{B})$ where Π is the initial state distribution, $\Gamma = [\gamma_{\mathrm{ij}}]$ is the state transition probability matrix with $\gamma_{\mathrm{ij}} = \mathrm{P}(C_{\mathrm{t}+1} = j \mid C_{\mathrm{t}} = i)$ and $B = [b_{\mathrm{i}}(x_{\mathrm{t}})] = \mathrm{P}(X_{\mathrm{t}} = x_{\mathrm{t}} \mid C_{\mathrm{t}} = i)$ are the emission probabilities (the probability of emitting x when in state i). The chain of the latent states $\{C_{\mathrm{t}}\}$ is modelled as a first-order Markov chain, while the observations $\{X_{\mathrm{t}}\}$ (electricity consumption) They are independent conditional on the states [22, 24]. Within the framework of the Hidden Markov Model HMM structure employed in electricity consumption data, the observed sequence $X^T = (X_1, X_2, X_3, \ldots, X_{\mathrm{T}})$ is assumed to be generated by an underlying sequence of hidden states $C_{\mathrm{t}} = (c_1, c_2, c_3, \ldots, c_{\mathrm{T}})$. Algebraically, the joint distribution of observations and states is given by

$$P(X_1, \ldots, X_{\mathrm{T}}, C_1 \ldots, C_{\mathrm{T}}) = \Pi c_1 \prod_{t=2}^{T} \gamma_{C_{\mathrm{t}-1}, C_{\mathrm{t}}} \prod_{t=1}^{T} b_{C_{\mathrm{t}}}(X_{\mathrm{t}}),$$

where Πc_1 represents the initial state probability, γ_{ij} denotes the transition probability from state i to state j, $b_{\mathrm{i}}(X_{\mathrm{t}})$ is the emission probability of observing X_{t} given that the system is in state i. The ability of HMMs to decompose electricity consumption into distinct regimes and detect temporal relationships makes them suitable for application in the analysis and forecasting of

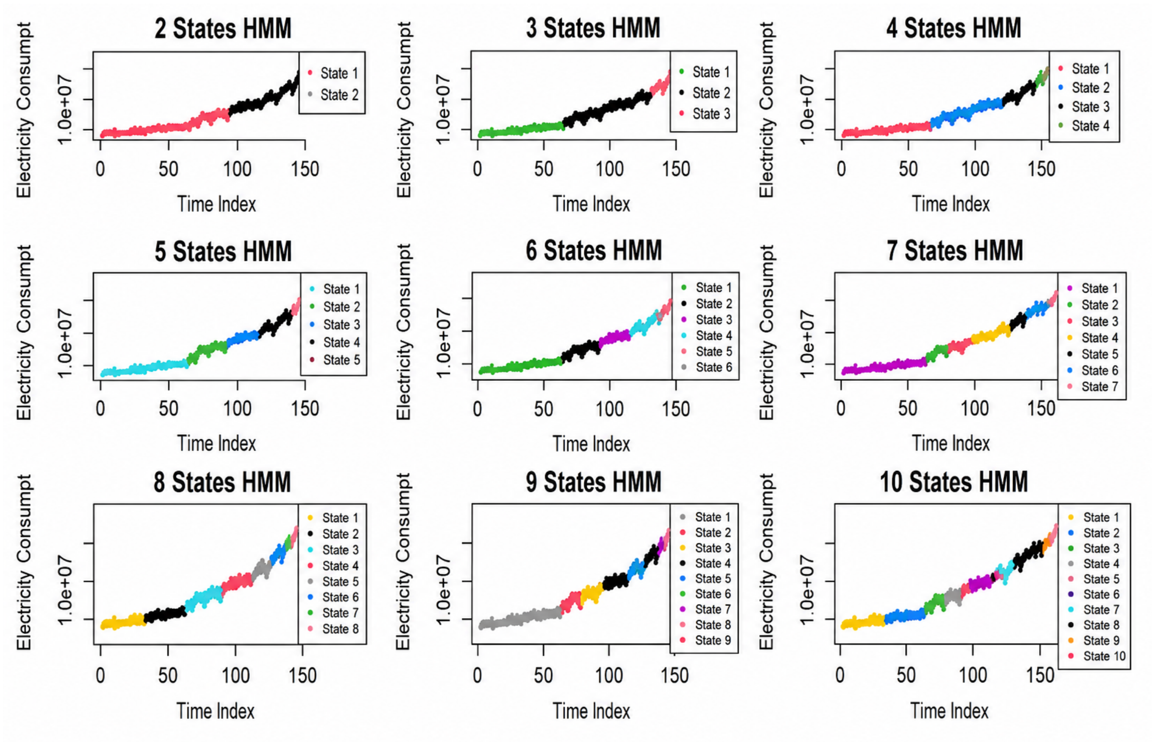


Figure 3. States Plot.

electricity consumption in short- and long-term frames.

Relationship Between Hidden States and Homoscedasticity, Homoscedasticity is where the variance of a time series remains constant over time [15]. For an HMM applied to modelling electricity usage, this would indicate that the variability in observed data is the same across all hidden states, but that the levels can differ. This is where the emission distributions for every state share a common variance. $X_t \mid (C_t = i) \sim N(u_i, \sigma^2), \forall_i$ Here, every state i introduces a different mean u_i, say low or high energy demand, but the variance within each state is constant at σ^2 is given by $Var(X_t) = \sigma^2 + Var(u_i)$.

$$Var(X_t) = \sum_{i=1}^{m} \mathsf{P}(C_t = i) \cdot Var(X_t \mid C_t = i) \\ + Var(E[X_t \mid C_t = i]).$$

If the state-specific values u_i are not very dispersed, then overall variance does not vary significantly over time. Homoscedasticity in HMMs, thus, means that the hidden states are the cause of changes in level in the observed series (differences in means) but not for differences in volatility or noise differences. This is appropriate in systems whose underlying process (appliance usage behavior) is inherently stable with respect to variability even though levels of consumption do vary. This type of modelling has been applied in standard power demand forecasting where short-term variances don't deviate far off [20, 24]. As opposed to homoscedasticity, heteroscedasticity implies that the variability of the under-observation variable is a function of time, sometimes in relation to the underlying latent states [10]. This is modelled in HMMs through allowing each hidden state to have its own emission variance $X_t \mid (C_t = i) \sim N(u_i, \sigma_i^2), \forall_i$ In this case, observation variance is conditioned on state C_t, and time variance as a whole is given by $Var(X_t) = \sum_{i=1}^{m} \mathsf{P}(C_t = i) \cdot \sigma_i^2 + Var(u_i)$ if the variances σ_i^2 differ substantially, which implies that the model exhibits state-dependent volatility. The peak-consumption state can be more volatile with unstable commercial or industrial load demand, while the off-peak consumption state can show more consistent residential consumption [4]. The heteroscedastic HMM has its conceptual setup in common with Markov-Switching GARCH models [4], whose clustering of volatility is explained by changes in latent states.

Table 2. State Variability.

No. States	Min Std Dev (Value, State)	Max Std Dev (Value, State)	Range
2	1,860,147 (State 2)	5,751,536 (State 1)	3,891,389
3	584,867 (State 3)	1,728,030 (State 2)	1,143,163
4	684,867 (State 2)	2,557,704 (State 3)	1,872,837
5	617,345 (State 4)	2,557,704 (State 2)	1,940,359
6	532,803 (State 2)	1,907,900 (State 4)	1,375,097
7	532,803 (State 5)	1,903,900 (State 4)	1,371,097
8	296,377 (State 1)	1,903,900 (State 6)	1,607,523
9	380,514 (State 9)	1,903,900 (State 3)	1,523,386
10	167,993 (State 9)	3,420,782 (State 5)	3,252,789

Table 2 presents the lowest and highest standard deviations for different numbers of states in a model and the range between them, providing some indication of heteroscedasticity by state. Heteroscedasticity is a case where the variance of errors or residuals differs among groups or states, rather than being constant. A high range between the lowest and highest standard deviations here is a sign of high heteroscedasticity. For the 2-state mod-

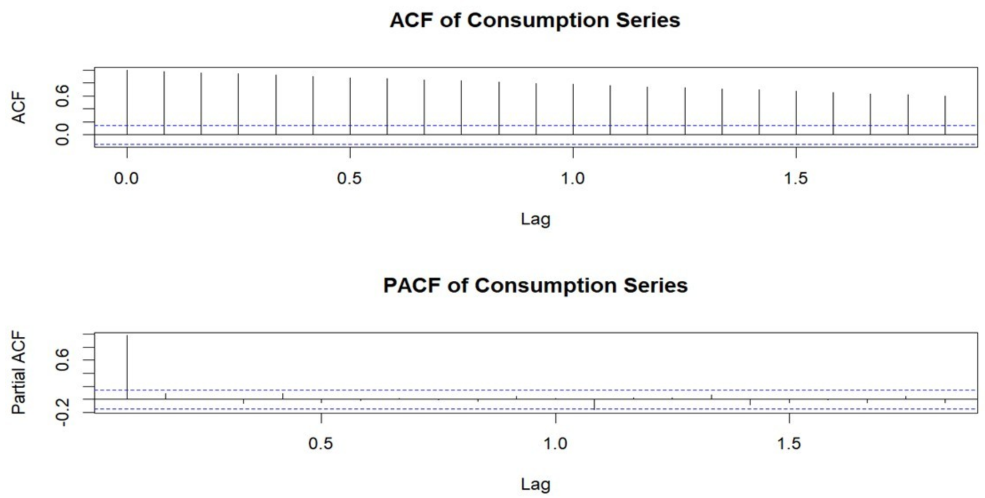


Figure 4. ACF and PACF Consumption Series.

els, the range is highest at about 3.9 million, where the lowest standard deviation occurs in State 2 and the highest one in State 1. Such a large difference implies that the variability of residuals changes considerably across the two states, meaning that there is strong heteroscedasticity. Similarly, the 10-state model also has an enormous range of about 3.25 million, with the smallest being State 9 and the largest being State 5, again reflecting intense heteroscedasticity. Both models have the highest variability differences across states. The 3- to 9-state models have smaller but still wide ranges of standard deviations, typically between roughly 1.1 million and 1.9 million. The 3-state model possesses the shortest range of approximately 1.14 million with the smallest standard deviation in State 3 and the biggest in State 2. Although it is less than the range in the 2- and 10-state models, this still indicates heteroscedasticity. The remaining models within this range possess similarly moderate residuals variability, which explains that although heteroscedasticity is not as bad, it remains significant and must not be neglected.

Figure 3 presents plots given in sequence that illustrate the application of Hidden Markov Models to electricity usage data and assess model performance for different state counts from 2 through 10. Such models are a robust methodology for detecting hidden patterns in time series data that enable the partitioning of consumption behavior into distinct states. In the simplest 2 or 3-state models, the patterns of electricity consumption appear relatively straightforward, typically identifying peak and trough usage times. Such models are useful for the rough-grained categorization of electricity consumption but may be too simplistically formulated to detect nuanced behavior or short-term dynamics. As the number of states increases from 4 to 10, the HMMs detect more complex and nuanced patterns within the electricity consumption data.

Diagnosis test for Stationarity, in determining the stationarity of the electricity consumption time series, a combination of statistical and graphical methods was employed as it shown in Figure 4. Initially, a graph of the raw data was produced as a time series and indicated clear trends and potential non-constant variance, initial indications of non-stationarity. This was also verified by Autocorrelation Function and Partial Autocorrelation Function plots, whose ACF revealed the slow decay characteristic of a non-stationary process. Statistical confirmation of this finding was done through the Augmented Dickey-Fuller test.

To assess the stationarity of the electricity consumption time series, both graphical and statistical techniques were used to confirm effective assessment. Initially, the raw data were plotted in a time series plot and revealed a clear pattern of increase and evidence of heteroscedasticity (non-constant variance) over time both typical signs of non-stationarity. Augmenting the visual diagnostics, the ACF and PACF plots were examined. The ACF presented a slow decline trend, typical of unit root processes and non-stationary series.

$$\Delta Y_{\mathrm{t}} = \alpha + \beta_{\mathrm{t}-1} + \gamma Y_{\mathrm{t}-1} + \sum_{i=1}^{m} \delta_{\mathrm{t}} \Delta Y_{\mathrm{t}-1} + \epsilon_{\mathrm{t}},$$

where Y_t is the electricity consumption at time t, Δ is the first difference operator, α is the constant (drift), β_t is the deterministic trend, γ is the coefficient tested for unit root ($H_0 : \gamma = 0$), and ϵ_t is white noise, with $\Delta Y_t = Y_t - Y_{t-1}$.

To strictly verify these results, the Augmented Dickey-Fuller (ADF) test as shown in Table 3, was carried out, which is a test against the null hypothesis of a unit root (or non-stationarity) with an alternative of stationarity. The test regression model used has a deterministic trend and lagged differences in an attempt to mitigate possible autocorrelation. The obtained ADF test statistic of -0.52669, at lag order 5 and p-value 0.9796, failed to reject the null hypothesis at 5% level of significance, thus determining that electricity consumption series is non-stationary. This result conforms to what has been found in earlier research concerning energy consumption time series data, where seasonality and trends are liable to induce non-stationarity.

Table 3. Augmented Dickey-Fuller.

Test	Test Statistic	Lag Order	p-value
Augmented Dickey-Fuller	-0.52669	5	0.9796

Residual analysis was necessary to check the assumptions of the model fitted to the time series. The residual plot in Fig-

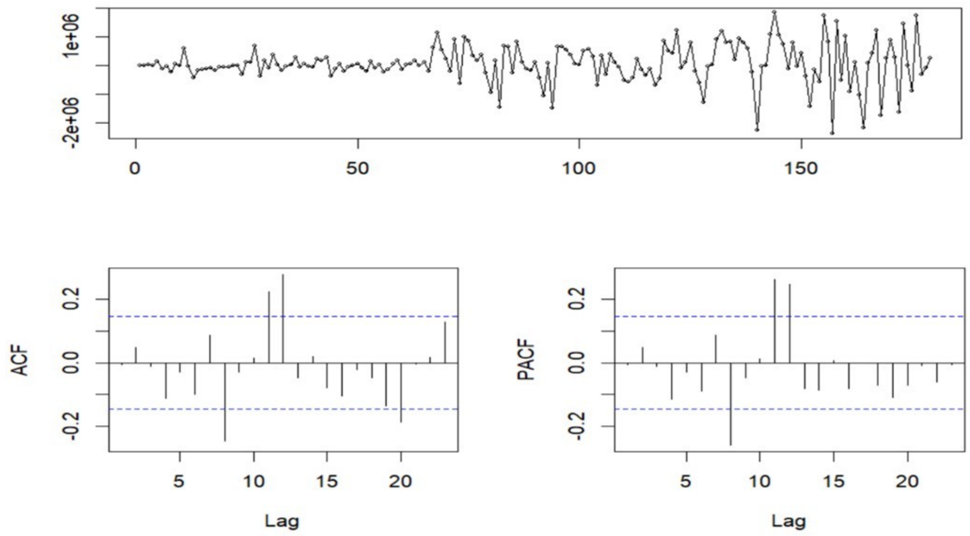


Figure 5. Residual Diagnosis.

Table 4. Residual Statistical Test.

Test	Test Statistic	df	p-value
Teräsvirta Test of Nonlinearity	5.828	2	0.05426
Box-Ljung Test	1493.30	10	< 2.2e-16
ARCH LM Test	165.74	12	< 2.2e-16

ure 5 showed a pattern of growing variance with time, particularly after observation 80, which is one of the signs of heteroscedasticity (non-constant error variance).

Table 4 presents the residual statistical test. First, the Teräsvirta Test of nonlinearity for the Neural Network test gives a test statistic of 5.828 with 2 degrees of freedom and p-value of 0.05426. Since the p-value is slightly above 0.05, we fail to reject the null hypothesis of linearity. This means that there is minimal evidence of nonlinearity in the model residuals, although the result is marginal and might possibly warrant further investigation. Secondly, the Box-Ljung Test for autocorrelation yields a test statistic of 1493.30 with 10 degrees of freedom and a p-value of less than $2.2e-16$. Such an extremely low p-value indicates that we reject the null hypothesis of no autocorrelation of the residuals. Therefore, there is clear evidence of autocorrelation, indicating that the model is unable to identify the temporal pattern in the data.

3.3. The Optimal Number of States

Table 5 below presents model fit indices for different state numbers, from 2 to 10, assessed by the Akaike Information Criterion and the Bayesian Information Criterion. Both AIC and BIC are widely used measures of model choice in which smaller values indicate a best trade-off between model fit and model complexity. Note the values; the 3-state model is significant in that it has the lowest AIC (5492.480) and also the best relative BIC (5707.104) compared to other models. This shows that the 3-state model achieves more goodness of fit and complexity than models with fewer or more states. Models with fewer states, such as the 2-state model, exhibit higher AIC and BIC values, which reflect worse fit or oversimplification. The 4-state model has an incremental increase in both the criteria compared to the 3-state model, and the values continue to increase with increasing states, hence demonstrating diminishing returns to increasing the number of states beyond 3.

Table 5. Optimal State.

Index	States	AIC	BIC
1	2	5889.035	5911.347
2	3	**5492.480**	**5707.104**
3	4	5627.991	5771.300
4	5	5775.202	5783.574
5	6	5788.658	5808.465
6	7	5870.597	5968.215
7	8	5891.645	5943.448
8	9	5941.611	6053.975
9	10	6398.515	6777.813

The graph in Figure 6 illustrates electricity consumption from July 2010 to May 2025, categorized by a 3-state Hidden Markov Model. The model separates the data into three unknown states, distinguished by different colours green, orange, and purple. The green state appears to dominate most of the initial years, demonstrating relatively lower rates of consumption. This state is likely an indicator of a stable or reduced electricity demand phase. Statistically, it is State 3 with the lowest mean consumption (7,922,075) and lowest variation (SD = 684,867), which reflects a stable and low-demand phase for the entire period under observation. Hidden Markov models have extensively been used in energy consumption research to classify efficiently consumption regimes with an eye on structural time changes [8, 22]. With time, and particularly around 2015, there is an evident move towards the orange state, which reflects increasing consumption. This is in line with State 1, which exhibits a moderate mean level of consumption (13,837,556) and a larger spread (SD = 2,323,326), representing increasing and more heterogeneous electricity demand, which may be from economic growth, urbanization, or broader access to energy infras-

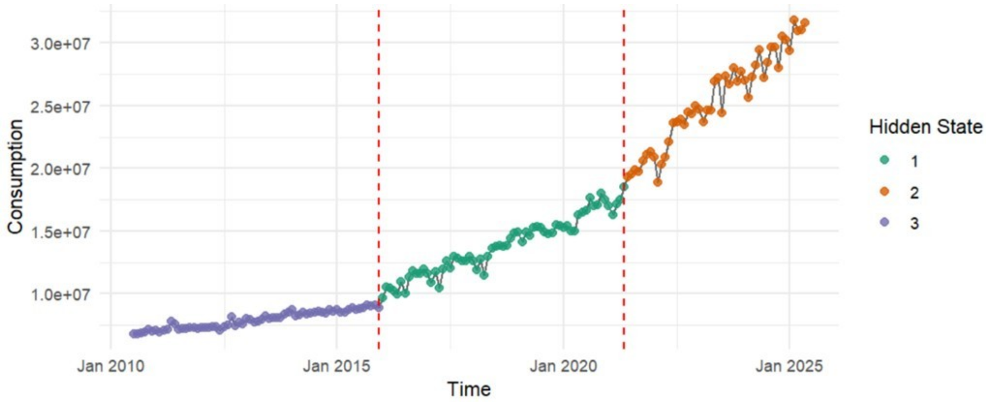


Figure 6. Electricity Consumption with Three State HMM Classification.

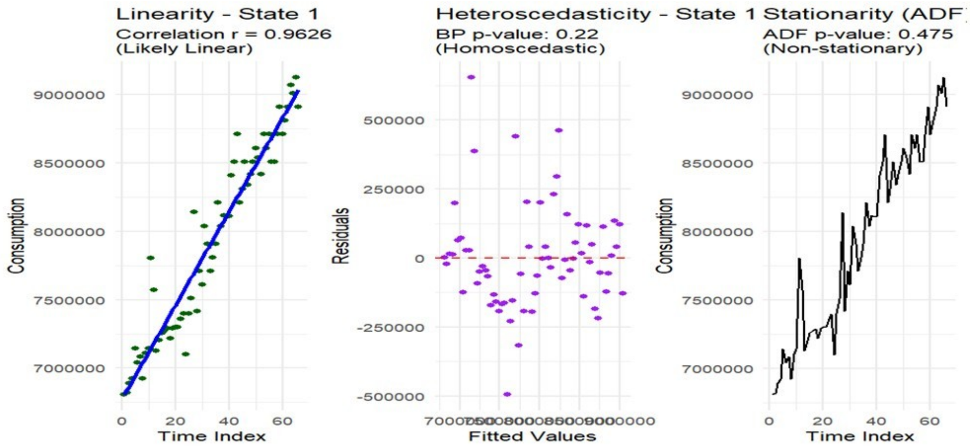


Figure 7. State One Diagnosis and Statistical Test.

tructure [13, 15]. The purple state, also prominent, seems to be a peak or transitional phase. This aligns with State 2, having the highest mean consumption (25,460,182) and highest variability (SD = 3,728,030).

Table 6 below presents summary statistical characteristics of three alternative states based on their count and population-related measures. State 2 possesses the highest mean within-group value of 25,460,182 with a range of 1,728,030 units of SD, indicating a relatively larger and more dispersed population size, between 18,898,275 and 31,825,574. State 1 possesses a mean of 13,837,556 with a moderate range (SD = 2,323,326), between 9,662,564 and 18,509,053. State 3 has the lowest mean of 7,922,075 and smallest variability (SD = 584,867), which indicates a smaller population size with less variability across it, from 6,808,209 to 9,126,934.

Table 6. State Summary Statistics.

State	Count	Mean	SD	Min	Max	Median
1	65	13,837,556	2,323,326	9,662,564	18,509,053	13,800,000
2	48	25,460,182	1,728,030	18,898,275	31,825,574	25,300,000
3	66	7,922,075	584,867	6,808,209	9,126,934	7,980,000

The Levene's test value that appears in Table 7 is used to check for homogeneity of variances (groups have equal variances). The test here has used the median as the middle value (a robust choice against outliers), and the degrees of freedom (Df) given is 2, indicating three groups are being compared. The F value of 60.329 is extremely large, and the associated p-value is less than $2.2e-16$, which is extremely small and much lower than the classical significance level of 0.05. This suggests that the null hypothesis of equal variances among the groups is strongly rejected.

Table 7. Levene's Test.

Test	Df	F value	Pr(>F)
Levene's Test (center = median)	2	60.329	< 2.2e-16

State 1 diagnosis plots as it shown in Figure 7 provide a comprehensive examination of the inherent time series characteristics of electricity consumption data. In the first panel, there is a prominent linear trend indicated by a high Pearson correlation coefficient ($r = 0.9626$) of time index and consumption values. With respect to heteroscedasticity, the p-value of the Breusch-Pagan (BP) test is 0.22, which supports homoscedasticity for this state. This implies homoscedastic error variance, a critical assumption to be made in linear regression model construction. For stationarity, the Augmented Dickey-Fuller (ADF) test has a p-value of 0.475, which means non-stationarity of the time series.

The Shapiro-Wilk test in State 2 as it shown in Figure 8 has a p-value of 0.2383, indicating that the residuals are approx-

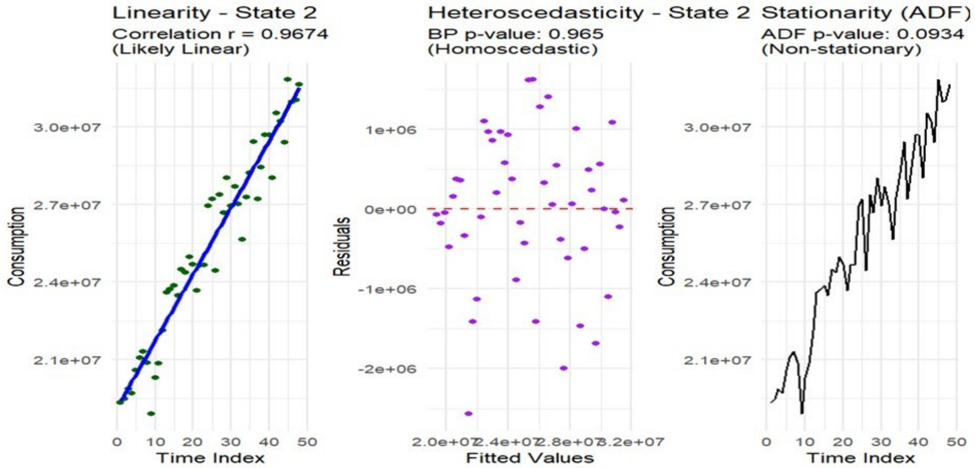


Figure 8. State Two Diagnosis and Statistical Test.

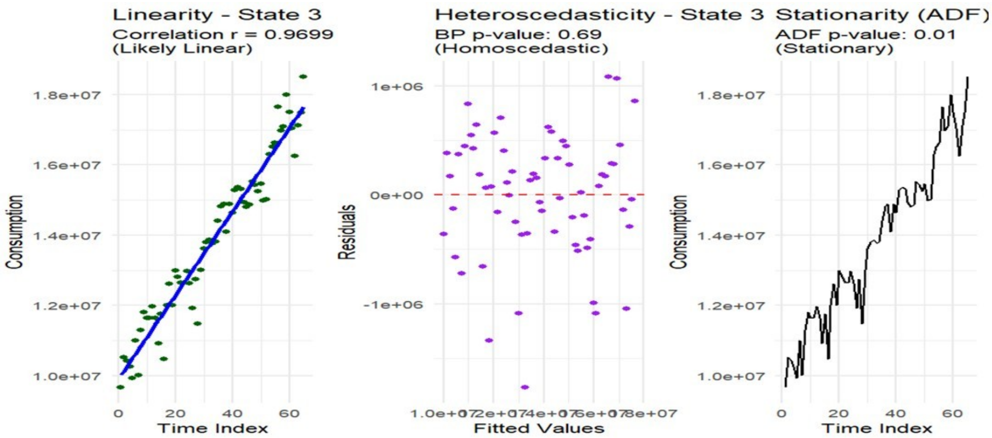


Figure 9. State Three Diagnosis and Statistical Test.

imately normally distributed—a desirable feature for a majority of linear and time series models. The BP test, with a value of 0.9650, confirms homoscedasticity and also reinforces residual stability of the model. While the ADF test shows a p-value of 0.0934, with mild non-stationarity, it does show that one might be able to achieve stationarity with little differencing. The value of the Durbin-Watson (DW) test of 1.6694 shows slight positive autocorrelation. The high value of 0.9674 for the correlation coefficient supports an extreme linear trend in State 2 and is proof that even at this regime, linear modelling is predictive.

State 3 as it shown in Figure 9, statistically behaves best for time series modelling. A p-value of 0.0723 through the Shapiro-Wilk test verifies normality of residuals. The BP test also yields a p-value of 0.6903, verifying homoscedasticity, as in the past states and as attested by [15], who used HMMs for predictive maintenance to verify stability of variance across regimes. Importantly, the ADF test has a p-value of 0.01, enabling us to reject the null hypothesis of a unit root, thereby establishing stationarity. The DW statistic of 1.3163 shows mild positive autocorrelation, which, albeit moderate, needs to be accommodated in model adjustment (augmenting AR terms). Finally, the correlation coefficient of 0.9699 reinforces a very high linear relationship, meaning trend-based models will be especially strong in such a state.

3.4. Theory Statement

"The provided time series $X^T = (X_1, X_2, X_3, \dots, X_T)$ violates the homoscedasticity assumption globally. However, by modelling the data using a Hidden Markov Model (HMM), we have an unobserved state sequence $C_t = (c_1, c_2, c_3, \dots, c_T)$ for which the residuals variance over each state $C_t = i$ is constant (homoscedastic)".

Hence, the optimal number of states m^* is that which minimizes heteroscedasticity over the states. This segmentation relies on the nonparametric HMM is capable of estimating regimes adaptively in a way to minimize within-state heteroscedasticity also this proves useful when working with electricity usage data, where load shifting, structural breaks, and seasonality have a tendency to render global assumptions invalid [18, 20, 23]. The original time series $X^T = (X_1, X_2, X_3, \dots, X_T)$ exhibits heteroscedasticity, meaning the variance of its residuals changes over time $Var(\epsilon_t) \neq Var(\epsilon_{t+k})$ for some $t \neq t+k$ by modelling this data with a Hidden Markov Model, we introduce a latent state sequence $C_t = (c_1, c_2, c_3, \dots, c_T)$, where each state $C_t = i$ corresponds to a segment of the data with a specific structure. The model defines Initial state distribution, $\delta = [\delta_1, \delta_2, \delta_3, \dots, \delta_m]$, Transition matrix, $\Gamma = [\gamma_{ij}] = \mathrm{P}(C_{t+1} = j \mid C_t = i)$ and Emission probabilities, $B = [b_i(x_t)] = \mathrm{P}(X_t = x_t \mid C_t = i)$.

The plot in Figure 10 demonstrates the rolling variance of

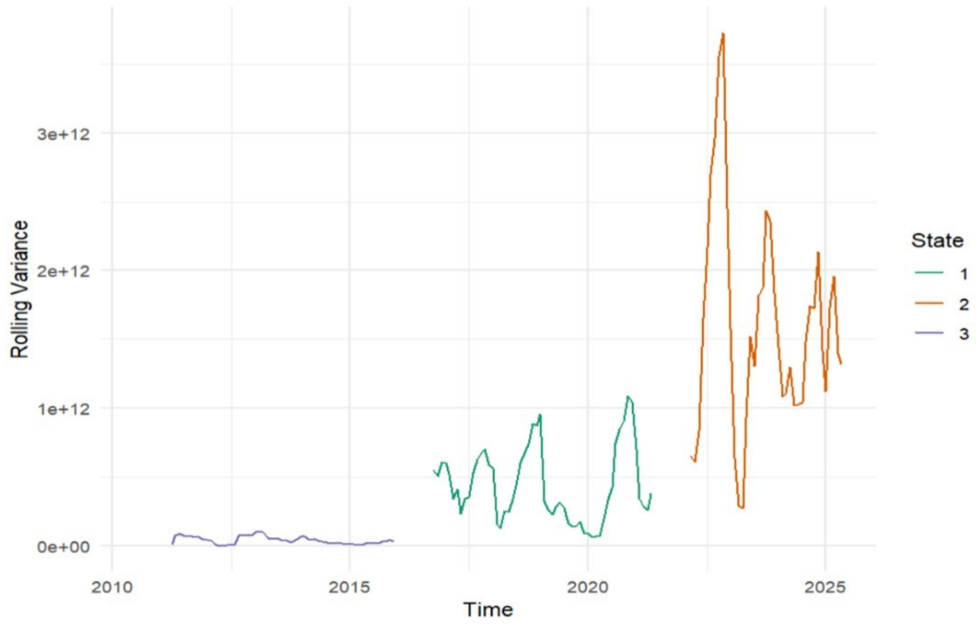


Figure 10. The Rolling Variance of Consumption Within each State (Window = 10).

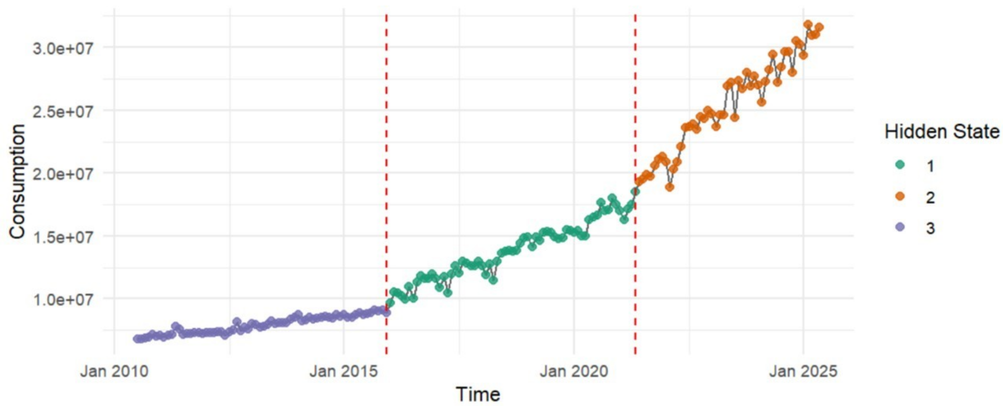


Figure 11. Electricity Consumption with 3-HMM Classification.

consumption over three states for 2010–2025, with a window of 10. Each state is represented by a different coloured line and reflects the way consumption variance fluctuates with time. The rolling variance for all states is fairly low and stable at the beginning but then exhibits sharp increases around the year 2020, particularly for State 2, with a sharp spike. This shows greater volatility in consumption for that state over the period. State 1 and State 3 show less extreme variation, with State 1 remaining very steady and State 3 showing small oscillations.

The plot overall shows sharp differences in consumption practice between the states, especially in the discerned peak, which implies underlying economic transformation or external factor influence over consumption behaviour.

The graph in Figure 11 illustrates electricity consumption from July 2010 to May 2025, categorized by a 3-state Hidden Markov Model. The model separates the data into three unknown states, distinguished by different colours green, orange, and purple. The green state appears to dominate most of the initial years, demonstrating relatively lower rates of consumption. This state is likely an indicator of a stable or reduced electricity demand phase. Statistically, it is State 3 with the lowest mean consumption ($7,922,075$) and lowest variation (SD $= 684,867$), which reflects a stable and low-demand phase for the entire period under observation. Hidden Markov models have extensively been used in energy consumption research to classify efficiently consumption regimes with an eye on structural time changes [17, 22]. With time, and particularly around 2015, there is an evident move towards the orange state, which reflects increasing consumption. This is in line with State 1, which exhibits a moderate mean level of consumption ($13,837,556$) and a larger spread (SD $= 2,323,326$), representing increasing and more heterogeneous electricity demand, which may be from economic growth, urbanisation, or broader access to energy infrastructure [20]. The purple state, also prominent, seems to be a peak or transitional phase. This aligns with State 2, having the highest mean consumption ($25,460,182$) and highest variability (SD $= 3,728,030$).

Figure 12. Electricity Consumption State 1 ($n = 65$).

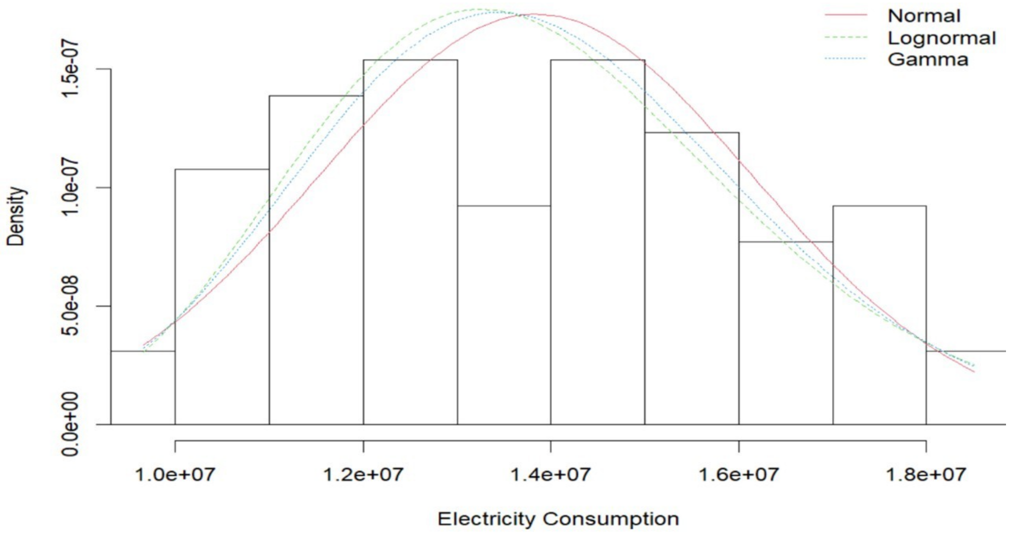


Figure 13. Density Comparison State One.

3.5. *Electricity Consumption at State One*

The graph in Figure 12 illustrates the pattern of electricity consumption by State 1 during the period from January 2016 to January 2021. During the period, there is an upward trend at a steep slope, indicating a consistent increase in electricity consumption. Use begins at approximately 10 million units at the start of 2016 and rises fairly consistently to approximately 18 million units at the start of 2021, although overall the trend is upwards, there are minor fluctuations along the way, which are indicative of seasonal patterns or extraneous effects.

Table 8. Probability Distributions of States One.

Dist.	Parameters	Log-Lik.	AIC	BIC
Normal	Mean = 13,837,556 SD = 2,323,326	-1044.530	2093.061	2097.409
Lognormal	Meanlog = 16.4287 SDlog = 0.1694	-1044.676	2093.351	2097.700
Gamma	Shape = 35.473 Rate = 2.5635e-06	-1044.421	2093.841	2097.490

To determine the best-fitting probability distribution of State 1 electricity consumption, three distributions were considered as candidates normal, lognormal, and gammaas it shown in Table 8. They were compared against one another based on their estimated parameters, log-likelihood values, and two popular model selection measures: Akaike Information Criterion and Bayesian Information Criterion. Both AIC and BIC punish the model complexity to avoid overfitting, and smaller values indicate an adequate balance between goodness-of-fit and model simplicity. Normal distribution assumes the data are symmetrically distributed around the mean and the probability density is entirely characterized by the mean $(13, 837, 556)$ and standard deviation $(2, 323, 326)$. log-likelihood for the model was -1044.530, which provided an AIC of 2093.061 and a BIC of 2097.409, both of which were the lowest compared to the other distributions. The results indicate a reasonable fit to State 1 observed data on the assumption of normality. Lognormal distribution, as is typical for positively skewed data, was meanlog 16.4287 and sdlog 0.1694. Its log-likelihood was -1044.676, slightly lower than normal models, so AIC (2093.351) and BIC (2097.700) were also higher. Despite the differences being small, these values suggest lognormal distribution is not fitting well enough to justify its higher complexity or asymmetry. Gamma distribution, which has the ability to handle skewed and non-

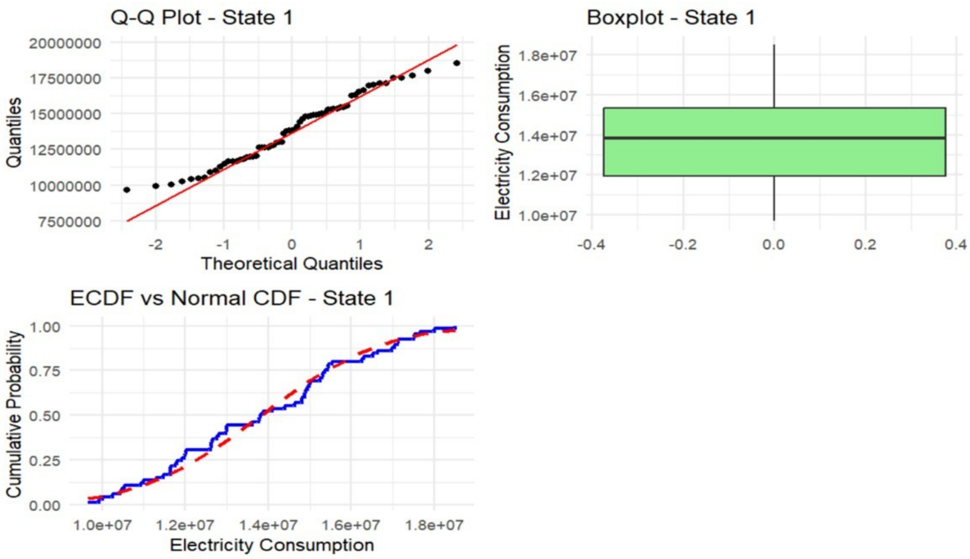


Figure 14. Normality Diagnosis State One.

negative data, was parameterized with 35.473 as shape and $2.5635e-06$ as rate. Despite being a facilitating model, gamma model returned a log-likelihood of -1044.421, just modestly superior to Lognormal's. Its associated AIC (2093.841) and BIC (2097.490) were still larger than its normal distribution counterparts, indicating that the added model complexity was not compensated by a better fit. Based on AIC and BIC values comparison through log-likelihood, Normal distribution is the best-fit model for State 1 electricity consumption data. Its lower AIC and BIC values show a better balance between model fit and parsimony. As data in state 1 is quite symmetrical and quite variable, the normality hypothesis is statistically sufficient and operationally sufficient for further analysis, forecasting and simulation in the Hidden Markov Model framework.

The "Density Comparison - State 1" graph in Figure 13 shows the scatter of electricity consumption in a given state, comparing three statistical distributions normal, lognormal, and gamma distributions. The x-axis is electricity consumption and the y-axis is density. Histogram bars give the empirical distribution, which is highest at the mid-range of consumption values. Of the overlaid distributions, the normal distribution (red line) is symmetric but not well-fitted to the histogram, indicating that it underestimates variation and cannot capture the right-skewness of the data. Conversely, the lognormal distribution (green dotted line) better captures actual data, tracking very closely the right-skew characteristic of consumption data, Similarly, the gamma distribution (blue dotted line) offers lasting flexibility in tracking the skewed pattern of the data, embracing various consumption attitudes.

Diagnostic plots, Q-Q plot, boxplot, and ECDF vs. normal CDF is it shown from Figure 14 below also provide further insights into the distributional nature of electricity consumption data. Q-Q plot (top left) plots the empirical quantiles of data points against theoretical quantiles of a normal distribution. Most of the data points follow the red reference line, particularly in the middle portion, indicating approximate normality.

The tail deviations are typically explained by non-normal consumption patterns induced by random behavioural or environmental determinants. The boxplot (top right) also indicates this. It displays a symmetric interquartile range (IQR) with the median at the centre of the box and no extreme outliers. The graphical symmetry thus confirms the Q-Q plot indication that electricity consumption distribution is approximately normal, especially in the middle quartiles. Boxplots do well in displaying non-normal characteristics such as skewness or outliers of which there are none that are overwhelming. The Empirical Cumulative Distribution Function (ECDF) vs. theoretical normal CDF (below) is a cumulative perspective. The concordance of ECDF (blue) and normal CDF (red) across most of the data range suggests a good fit of the normal distribution.

The Shapiro-Wilk test on State 1 electricity consumption data as it shown in Table 9 below gives the W statistic 0.96895 and the p-value 0.1016. At the conventional 5% level of significance, we cannot reject the null hypothesis of the Shapiro-Wilk test of normality of data. This is an indication that enough evidence does not support the establishment of non-normality in State 1 electricity consumption data. This is also clear from the diagnostic plots above. The Q-Q plot yielded a good overlap of empirical quantiles with quantiles of the normal distribution, especially in the middle region. The boxplot also yielded symmetric interquartile range with no outliers in the extremes favoring plausibility of normality. The ECDF vs. normal CDF plot also showed close overlap over the majority of the range of the distribution except for small deviations towards the extremes. Collectively, the p-value of over 0.05 from the Shapiro-Wilk test and graphical proof using the Q-Q plot and ECDF corroborate the finding that State 1's consumption data does not deviate significantly from normality [4, 10].

Table 9. State One Normality Test.

Test	W Statistic	p-value
Shapiro-Wilk	0.96895	0.1016

The electricity consumption data for State 1 consists of 65 observations with a mean (μ) of $13,837,556$ and a standard deviation (σ) of $2,323,326$. Assuming a normal distribution, the probabil-

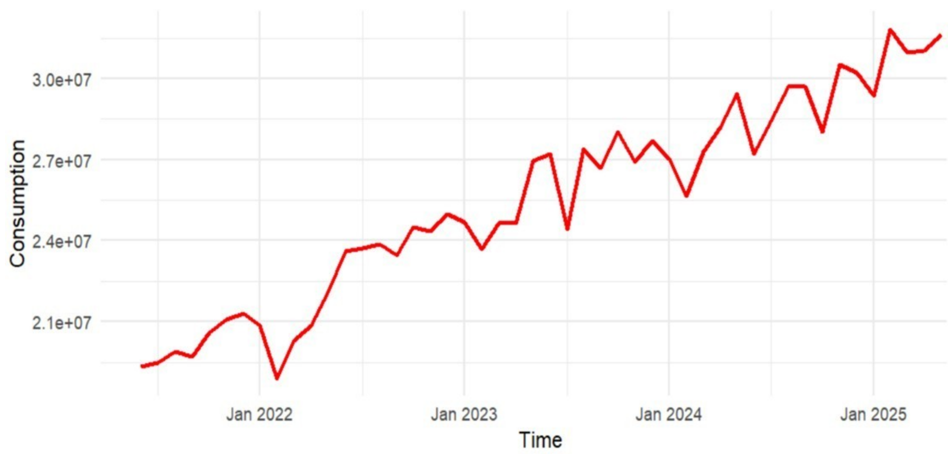


Figure 15. Electricity Consumption State 2 ($n = 48$).

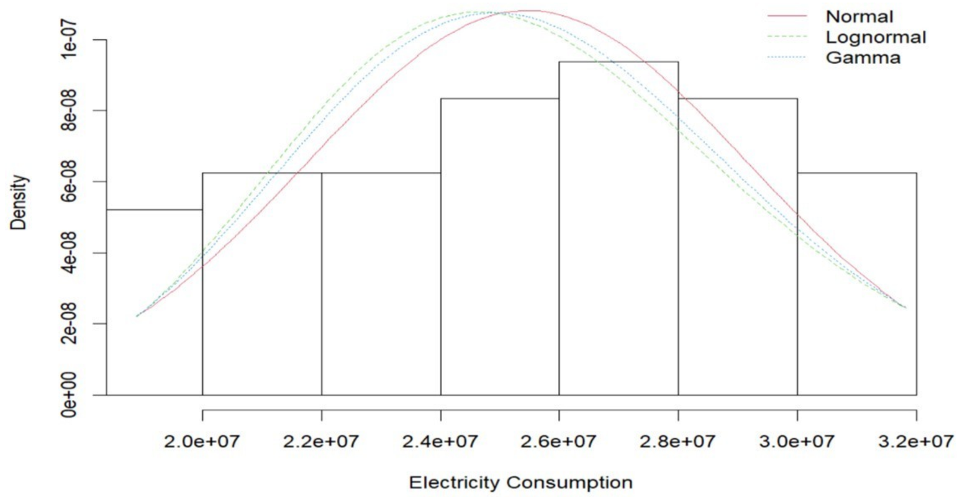


Figure 16. Density Comparison at State 2.

ity density function (PDF) for electricity consumption x_t is given by:

$$f(x_t, \mu, \sigma) = \frac{1}{\sqrt{2\pi \times 5.3958 \times 10^{12}}} \times \exp\left(-\frac{(x_t - 13,837,556)^2}{2 \times 5.3958 \times 10^{12}}\right), \quad x_t > 0. \qquad (1)$$

3.6. Electricity Consumption at State Two

The time series plot "Electricity Consumption - State 2" in Figure 15 below illustrates the trend in electricity consumption from January 2022 to January 2025, spanning 48 months of data. Time is quantified on the x-axis and the amount of electricity consumption on the y-axis. The red line plotted over the time series is a consistent and clear upward trend, starting with low consumption in early 2022 and rising incrementally over the observation period.

The "Density Comparison - State 2" graph in Figure 16 below illustrates relative goodness-of-fit of three statistical distributions normal, lognormal, and gamma distributions to the data on electricity usage. The histogram is right-skewed with a mode in the middle range, as is characteristic of usage data as seen in time series statistics on energy consumption [16]. Symmetric normal distribution curve (red line) is not able to explain this skewness adequately, as is found with results that normality assumptions tend to fail in their ability to model positively skewed consumption data. Also, the gamma distribution (blue dashed line), which has been promoted on the basis of the flexibility of its shape and scale parameters, also matches the empirical data very closely, in line with previous work in which gamma distributions provided good fits to skewed rain and energy use data. Lognormal and gamma models therefore provide good models of the underlying distribution of electricity consumption for State 2 that are superior to the Normal model. In selecting the best probabilistic model for electricity consumption in state 2, three of the most widely used continuous distributions (normal, lognormal, and gamma) were compared. Three measures were used to make the comparison, log-likelihood, Akaike Information Criterion, and Bayesian Information Criterion as it shown in Table 10 below. The lower values of AIC and BIC indicate that a model with a better trade-off between goodness of fit and parsimony is provided [13]. State 2's normal distribution had the mean 25,460,182 and the standard deviation 1,728,030. It provided the maximum log-likelihood value of -793.911, which was equal to minimum AIC (1591.821) and BIC (1595.563) among the three models. The values show that the normal distribution provides the most efficient and parsimonious model fit to state 2's data.

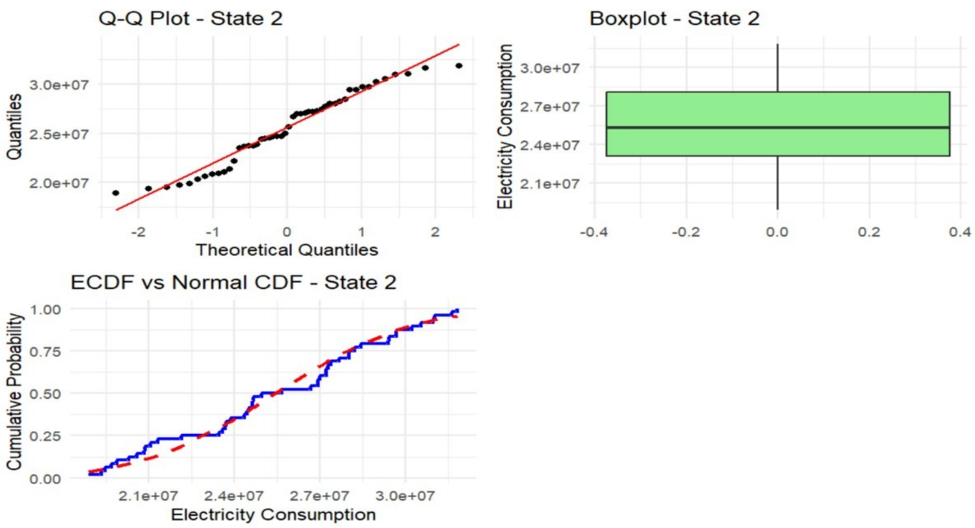


Figure 17. Normality Diagnosis at State Two.

This showed that when time series data are roughly symmetric and of moderate dispersion, the normal distribution will give reliable results in HMM segmentation. The lognormal model was also fitted with meanlog $= 17.0418$ and sdlog $= 0.1486$. The log-likelihood for the model was -794.602 and had a slightly increased AIC of 1593.205 and BIC of 1596.947 compared to the normal model. While the variations are small, they remain inclined towards the normal distribution.

The gamma distribution, with a shape of 46.2524 and rate of 1.8166×10^{-6}, possessed a log-likelihood of -794.261 and very slightly larger AIC (1592.521) and BIC (1596.264) than the normal distribution. Although its fit is better than that of the lognormal by a small amount, it still can't compete with the normal distribution. On relative AIC, BIC, and log-likelihood grounds, normal distribution captures best the electricity consumption data for state 2. It offers the most satisfactory accuracy-simplicity trade-off, implying that the data are well explained by a symmetrically distributed variable of moderate variance. The Q-Q plot (top

Table 10. Electricity Consumption in State Two.

Distribution	Parameters	Log-Likelihood	AIC	BIC
Normal	Mean $= 25,460,182$ SD $= 1,728,030$	-793.911	1591.821	1595.563
Lognormal	Meanlog $= 17.0418$ SDlog $= 0.1486$	-794.602	1593.205	1596.947
Gamma	Shape $= 46.2524$ Rate $= 1.8166 \times 10^{-6}$	-794.261	1592.521	1596.264

left) of observed electricity consumption quantiles versus a theoretical normal distribution as shown in Figure 17 below, shows that the majority of data points closely follow the reference line, particularly towards the middle of the distribution, and thus imply rough normality in most of the data. However, the tail deviations show minimal deviations from normality, a not unusual feature of environmental data because of extreme values or variability in measurements. The boxplot (top right) concludes the same through showing a symmetric distribution with the median near the midpoint of the interquartile range and no extreme outliers, which are typical signs of data being close to normal. The overlap of the empirical cumulative distribution function (ECDF) and the theoretical normal cumulative distribution function (CDF) (bottom) is a good one for most of the range of data, confirming normality for most consumption values, though the tail differences observed are in agreement with those in the Q-Q plot.

The Shapiro–Wilk test for State 3 electricity consumption is shown in Table 11 below, providing a W statistic of 0.95562 with a p-value of 0.06716. Since the p-value is greater than the conventional alpha level of $\alpha = 0.05$, we are not in a position to reject the null hypothesis of normality. This means that the data does not significantly deviate from a normal distribution, which complements the graphical tests such as Q–Q plots and ECDF comparisons.

Table 11. State Two Normality Test.

Test	W Statistic	p-value
Shapiro-Wilk	0.95562	0.06716

The electricity consumption data for State 2 is modelled using a normal distribution with a sample size of 48, a mean (μ) of 25,460,182 units, and a standard deviation (σ) of 1,728,030 units. The corresponding probability density function (PDF) is given by:

$$f(x_t, \mu, \sigma) = \frac{1}{\sqrt{2\pi \times 2.9871 \times 10^{12}}} \times \exp\left(-\frac{(x_t - 25{,}460{,}182)^2}{2 \times 2.9871 \times 10^{12}}\right), \quad x_t > 0. \tag{2}$$

3.7. Electricity Consumption at State Three

The graph of the time series titled "Electricity Consumption – State 3" as shown in Figure 18 reveals a clear upward trend in electricity consumption from January 2011 to January 2016, with 66 data points. The cyclic fluctuations observed in the consumption levels are characteristic of variations that occur due to

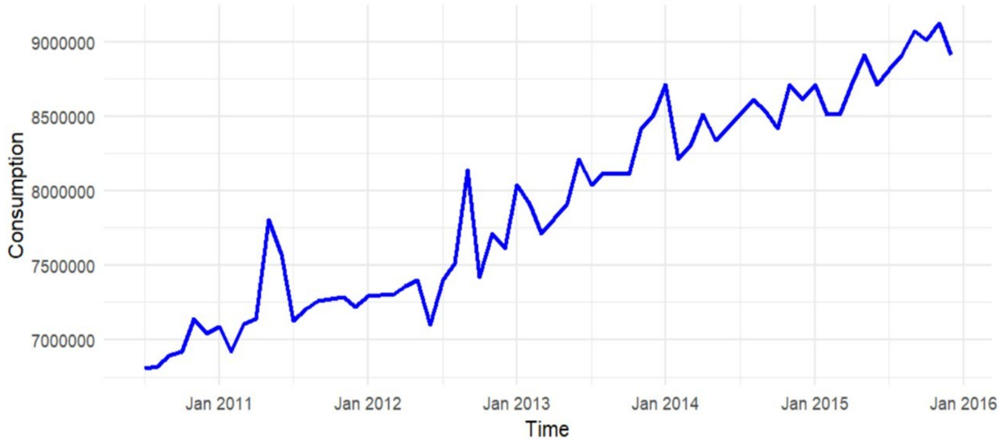


Figure 18. Electricity Consumption State Three ($n = 66$).

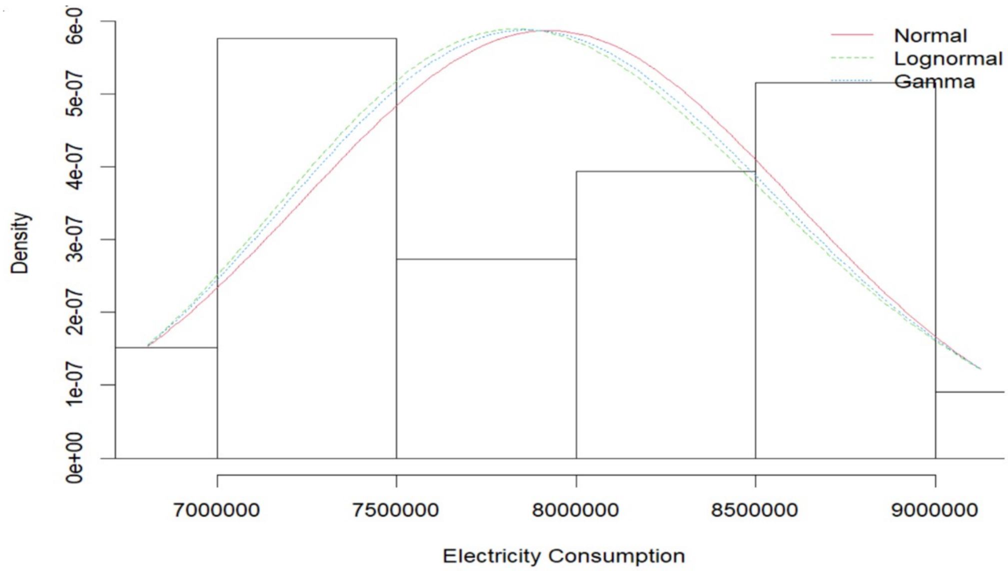


Figure 19. Density Comparison at State Three.

seasonality, as are likely within electricity consumption due to weather and seasonality changes in consumer behaviour.

The "Density Comparison – State 3" plot in Figure 19 below shows the potential to model data for electricity consumption according to different statistical distributions, which are Lognormal, Normal, and Gamma. The histogram indicates the right-skewness of the empirical data at 8 million units, which is to be expected in consumption data due to variation and positive skew of such readings. The Normal distribution's symmetric shape is not a good fit to describe such skewness, as evidence shows electricity consumption itself is non-normal.

For comparison of the most suitable probability distribution of State 3 electricity demand, three potential distributions (Normal, Lognormal, and Gamma) were fit with each of their own log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), as shown in Table 12. For model comparison, lower values for AIC and BIC are better models after accounting for fitand complexity.

The Normal model with mean 7,922,075 and variance 584,867 had the optimal value of log-likelihood (-979.987) out of the three models. Its AIC (1963.174) and BIC (1968.113) are also the lowest and it performs the best. This suggests that the consumption data of State 3 follows a symmetrical pattern of variability with relatively low spread, similar to what obtains in a Normal distribution.

Table 12. State Three Distribution Comparison.

Distribution	Parameters	Log-Likelihood	AIC	BIC
Normal	Mean $= 7,922,075$ SD $= 584,867$	-979.987	1963.174	1968.113
Lognormal	Meanlog $= 15.8815$ SDlog $= 0.0861$	-979.993	1963.986	1968.365
Gamma	Shape $= 135.417$ Rate $= 1.7094 \times 10^{-5}$	-979.946	1963.892	1968.272

Lognormal, with meanlog $= 15.8815$ and sdlog $= 0.0861$, had a very slightly lower log-likelihood (-979.993), but higher AIC (1963.986) and BIC (1968.365) than the Normal model. This indicates a marginally worse fit. The Gamma distribution with shape estimated as 135.417 and rate 1.7094×10^{-5} gave a log-likelihood of -979.946, marginally better than the other two. Its

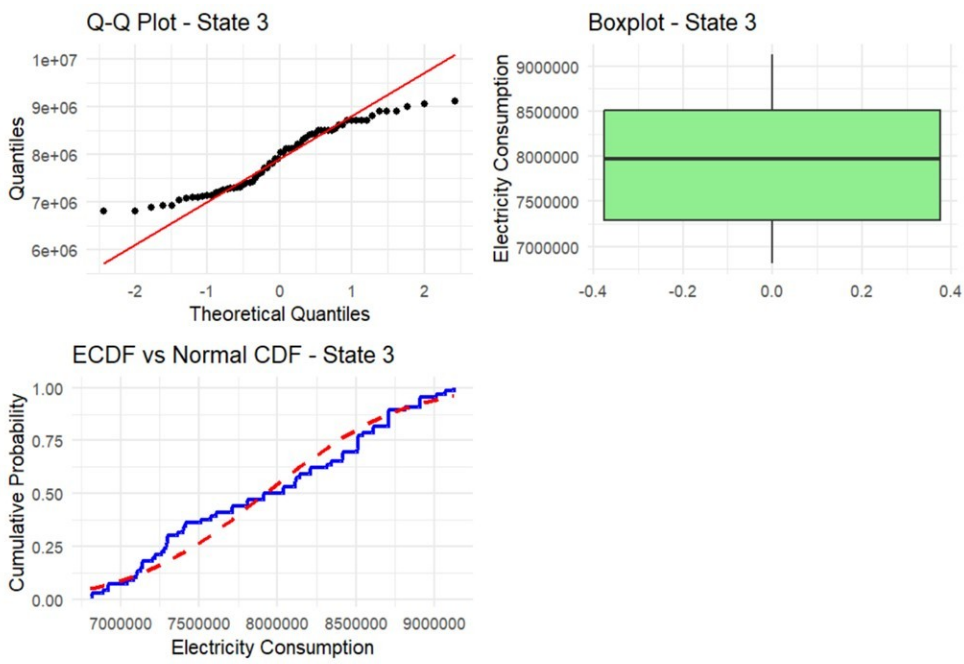


Figure 20. Normality Diagonosis at State Three.

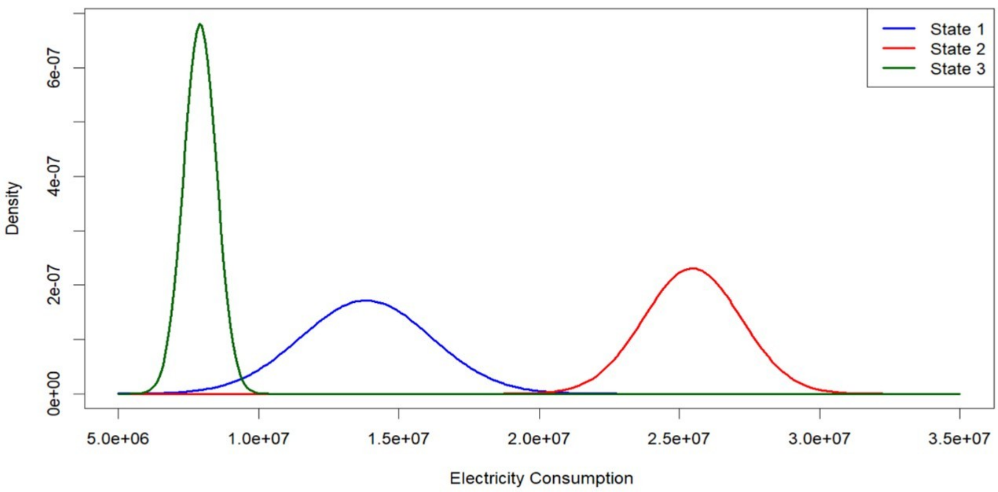


Figure 21. Normal distribution for HMM states.

AIC (1963.892) and BIC (1968.272) remain higher, however, than those of the Normal model.

The diagnostic plots for electricity consumption in State 3, as shown in Figure 20, produce a comprehensive assessment of the normality of data. The Q–Q plot shows the observed quantiles to track those of the standard normal distribution closely, particularly in the central region, which supports the approximate normality assumption; however, the tail discrepancies reveal the universal issue of fatter tails in the consumption data [10, 23].

Similarly, the symmetrical nature of the shape of the boxplot and the middle-centered median, unharmed by extreme outliers, reinforce the suggestion of normality. The close approximation of the empirical cumulative distribution function (ECDF) to the normal CDF also reinforces this concordance for most values, though tiny discrepancies at the tails of the distributions suggest some slight skewness or kurtosis. The Shapiro–Wilk test reaffirms this conclusion for State 3 data ($W = 0.95562$, $p = 0.06716$), demonstrating marginal acceptability of normality consistent with observed tail deviations from Q–Q plots.

Table 13. State Three Normality Test.

Test	W Statistic	p-value
Shapiro-Wilk	0.93659	0.06120

The Shapiro–Wilk test statistic, as shown in Table 13 of the electricity consumption data, is 0.93659 for the W statistic and 0.06120 for the p-value. At a 5% significance level, the p-value is larger than the threshold (0.05), indicating there is not sufficient evidence to reject the null hypothesis of normality. This suggests that the electricity consumption data distribution doesn't significantly differ from a Normal distribution. Although the W statistic is slightly less than 1, the result still verifies the assumption of near-normality, as suggested by earlier graphical examination such as Q–Q plots and ECDF plots.

The electricity consumption in State 3 is modelled using a normal distribution characterized by a sample size of 66, a mean (μ) of $7{,}922{,}075$, and a standard deviation (σ) of $584{,}867$. The resulting probability density function expresses the likelihood of observing a specific consumption level and is given by

$$f(x_t, \mu, \sigma) = \frac{1}{\sqrt{2\pi \times 3.4207 \times 10^{11}}} \times \exp\left(-\frac{(x_t - 7{,}922{,}075)^2}{1.06296 \times 10^2}\right), \quad x_t > 0. \tag{3}$$

The plot called "Normal Distributions for HMM States" in Figure 21 shows normal distribution curves for electricity usage for three different states: State 1, State 2, and State 3. Each is a density of observed levels of electricity usage for purposes of comparing usage patterns for the states.

State 1, represented by the blue curve, has a higher peak at a lower consumption level, indicative of a congested pattern of utilization. The thin nature of the curve represents lower variability in electricity use in this state, indicative of a more consistent demand profile as seen in local-scale regional energy research [11, 20]. In contrast, State 2, symbolized by the red curve, contains a higher peak and wider spread than State 1. This implies that the average level of electricity use is greater with greater variation.

4. Conclusion

Modelling the probability distributions of electricity consumption in every hidden state in a 3-state Hidden Markov Model (HMM) revealed significant results. It was tested for best-fit distributions between Normal, Lognormal, and Gamma in every hidden state. Log-likelihood values, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were the choice criteria, all based on the principle that lower values of AIC and BIC reflected better model fit with penalized complexity. For State 1 with mean electricity consumption of approximately 13.8 million units and standard deviation of 2.32 million, the best fitting was the Normal distribution with the highest log-likelihood ($-1{,}044.530$) and lowest AIC ($2{,}093.061$) and BIC ($2{,}097.409$) among the Lognormal and Gamma distributions. The Shapiro–Wilk test statistic value ($p = 0.1016$) confirmed the normality assumption, which was further complemented by Q–Q plots and empirical cumulative distribution functions (ECDFs) confirming symmetry and light tails.

Similarly, in State 2 with a larger mean consumption of about 25.5 million units and smaller relative variability (SD $=$ 1.73 million), the Normal distribution was superior to others. It provided the highest log-likelihood (-793.911), AIC ($1{,}591.821$), and BIC ($1{,}595.563$) values. Though weak autocorrelation was present in diagnostic plots, residuals were homoscedastic and roughly normal, in favor of the model. In State 3, the lowest state of consumption, with a mean of 7.92 million units and a standard deviation of $584{,}867$, the Normal distribution again best fit (log-likelihood $= -979.987$; AIC $= 1{,}1963.174$; BIC $= 1{,}1968.113$). The Shapiro–Wilk test ($p = 0.0612$) confirmed the normality assumption, and graphical diagnostics supported a symmetric and tightly clustered distribution. The Normal distribution fitted best in all three latent states. Results indicate that electricity consumption, when held constant for latent states, is best described by normally distributed data.

Author Contributions. **Eliasi M Jeremiah**: Conceptualization, methodology, formal analysis, supervision. **Ramkumar T Balan**: Software, validation, data curation, visualization. **Jairos K Shinzeh**: Investigation, resources, writing—review and editing. All authors discussed the results and actively participated in drafting, revising, and approving the final manuscript.

Acknowledgement. The author would like to thank the editors and reviewers for their insightful comments and helpful suggestions that significantly improved the quality of this manuscript.

Funding. The author received no funding for this study.

Conflict of interest. The authors declare no conflict of interest.

Data availability. Not applicable.

References

[1] R. Fachrizal, M. Shepero, D. van der Meer, J. Munkhammar, and J. Widén, "Smart charging of electric vehicles considering photovoltaic power production and electricity consumption: A review," *eTransportation*, vol. 4, p. 100056, 2020, doi:10.1016/j.etran.2020.100056.

[2] African Development Bank, "An outlook of Tanzania's energy demand, supply, and cost by 2030," African Development Bank (AfDB), Tech. Rep., 2022.

[3] B. J. Dye, "Unpacking authoritarian governance in electricity policy: Understanding progress, inconsistency and stagnation in tanzania," *Energy Research & Social Science*, vol. 80, p. 102209, 2021, doi:10.1016/j.erss.2021.102209.

[4] M. Haruna, B. Ilembo, and J. Lwaho, "Forecasting hydropower production in Tanzania using the SARIMA model," *Journal of Science, Innovation and Creativity*, vol. 4, no. 1, pp. 26–39, 2025, doi:10.58721/jsic.v4i1.958.

[5] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989, doi:10.1109/5.18626.

[6] W. Zucchini and I. L. MacDonald, *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2009.

[7] M. Aquino-Baleytó, V. Leos-Barajas, T. Adam, M. Hoyos-Padilla, O. Santana-Morales, F. Galván-Magaña, R. González-Armas, C. G. Lowe, J. T. Ketchum, and H. Villalobos, "Diving deeper into the underlying white shark behaviors at Guadalupe Island, Mexico," *Ecology and Evolution*, vol. 11, no. 21, pp. 14 932–14 949, 2021, doi:10.1002/ece3.8178.

[8] M. Chavan, S. Joshi, D. Patil, V. Kale, P. Kharat, and M. Kasar, "Time series analysis in electrical load forecasting," *Panamerican Mathematical Journal*, vol. 35, no. 1, pp. 210–219, nov 2024, doi:10.52783/pmj.v35.i1s.2307.

[9] J. Zhou, T. Peng, C. Zhang, and N. Sun, "Data pre-analysis and ensemble of various artificial neural networks for monthly streamflow forecasting," *Water*, vol. 10, no. 5, 2018, doi:10.3390/w10050628.

[10] D. Vora, N. Zade, P. Patwa, A. Kushwaha, G. Agrawal, and A. Sharma, "Temporal energy consumption exploration and time series forecasting of electricity dynamics," in *2024 IEEE Pune Section International Conference (PuneCon)*, 2024, pp. 1–6, doi:10.1109/PuneCon63413.2024.10895847.

[11] R. Glennie, T. Adam, V. Leos-Barajas, T. Michelot, T. Photopoulou, and B. T. McClintock, "Hidden markov models: Pitfalls and opportunities in ecology," *Methods in Ecology and Evolution*, vol. 14, no. 1, pp. 43–56, 2023, doi:10.1111/2041-210X.13801.

[12] A. Poritz, "Hidden markov models: a guided tour," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 7–13 vol.1, doi:10.1109/ICASSP.1988.196495.

[13] H. Shen, Z. Wang, X. Zhou, M. Lamantia, K. Yang, P. Chen, and J. Wang, "Electric vehicle velocity and energy consumption predictions using transformer and markov-chain monte carlo," *IEEE Transactions on Transportation Electrification*, vol. 8, no. 3, pp. 3836–3847, 2022, doi:10.1109/TTE.2022.3157652.

[14] H. Kaur and S. Ahuja, "Time series analysis and prediction of electricity consumption of health care institution using arima model," in *Proceedings of Sixth International Conference on Soft Computing for Problem Solving*, K. Deep,

J. C. Bansal, K. N. Das, A. K. Lal, H. Garg, A. K. Nagar, and M. Pant, Eds. Singapore: Springer Singapore, 2017, pp. 347–358.

[15] M. N. Azimi and S. F. Shahidzada, "A correcting note on forecasting conditional variance using ARIMA vs. GARCH model," *International Journal of Economics and Finance*, vol. 11, no. 5, pp. 145–153, 2019, doi:10.5539/ijef.v11n5p145.

[16] B. T. McClintock, B. Abrahms, R. B. Chandler, P. B. Conn, S. J. Converse, R. L. Emmet, B. Gardner, N. J. Hostetter, and D. S. Johnson, "An integrated path for spatial capture–recapture and animal movement modeling," *Ecology*, vol. 103, no. 10, p. e3473, 2022, doi:10.1002/ecy.3473.

[17] J. Lee and B. S. Hwang, "Energy demand pattern analysis in south korea using hidden markov model-based classification," *Asian Economic Journal*, vol. 38, no. 3, pp. 404–428, 2024, doi:10.1111/asej.12338.

[18] B. Hegde, Q. Ahmed, and G. Rizzoni, "Velocity and energy trajectory prediction of electrified powertrain for look ahead control," *Applied Energy*, vol. 279, p. 115903, 2020, doi:10.1016/j.apenergy.2020.115903.

[19] I. Ullah, R. Ahmad, and D. Kim, "A prediction mechanism of energy consumption in residential buildings using hidden markov model," *Energies*, vol. 11, no. 2, 2018, doi:10.3390/en11020358.

[20] J. Smith, "Technical skills and expertise in auditing: A comprehensive review," *Auditing & Accountability Journal*, vol. 35, no. 4, pp. 117–136, 2022.

[21] P. Twesigye, "Structural, governance, & regulatory incentives for improved utility performance: A comparative analysis of electric utilities in tanzania, kenya, and uganda," *Utilities Policy*, vol. 79, p. 101419, 2022, doi:10.1016/j.jup.2022.101419.

[22] H. M. Malik, N. Salem, and M. S. AlSabban, "A comparative study on the performance of hidden markov model in appliance modeling," in *2021 5th International Conference on Power and Energy Engineering (ICPEE)*, 2021, pp. 145–152, doi:10.1109/ICPEE54380.2021.9662546.

[23] M. Awad and R. Khanna, *Hidden Markov Model*. Berkeley, CA: Apress, 2015, pp. 81–104, doi:10.1007/978-1-4302-5990-9_5.

[24] M. Boo, F. Argüello, J. D. Bruguera, R. Doallo, and E. L. Zapata, "High-performance vlsi architecture for the viterbi algorithm," *IEEE Transactions on Communications*, vol. 45, no. 2, pp. 168–176, 1997, doi:10.1109/26.554365.