COMPARISON OF SOME ESTIMATION METHODS OF MISSING DATA IN HIDDEN MARKOV MODEL

ABSTRACT

This study compares four methods, Mean Imputation (MI), Median Imputation (MDI), Linear Interpolation (LI), and Kalman Filter Algorithm (KAL), for estimating missing values in time series data using Hidden Markov Models (HMM). The evaluation is based on accuracy measures: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The findings reveal that KAL outperforms other methods across all sample sizes under linear trend structures. On the other hand, MDI performs best under quadratic and exponential trend structures. HMMs were applied to the estimated series with MDI and KAL and compared with actual series models. The Akaike Information Criterion (AIC) values of the models for series with 12% missingness show minimal divergence from those of the actual series. This study underscores the importance of selecting suitable estimation methods tailored to specific trend structures in time series analysis.

Keywords:*State Space,Stochastic Process, Missing Values, Estimation, Variable.*

1 Introduction

1.1 Background to the study

A Hidden Markov Model (HMM) is a specific type of Markov chain in which the state is only partially observable. Before exploring HMMs, it's essential to understand one of their fundamental components: the Markov chain.

A Markov chain is a stochastic model that describes a sequence of potential events, where the probability of each event is determined solely by the state of the previous event. This concept falls under the broader category of Markov processes, named after the Russian mathematician A. Markov, who introduced it in the 19th century. A Markov process is defined by the Markov Property, which asserts that only the present state is necessary to forecast future events, as all pertinent past information is contained within it.

A Markov chain has the Markov property; let $X_n, n \in Z \geq 0$ be a **stochastic process**, taking values $i_1, i_2, i_3, \ldots, i_n$. If $X_n = i_k$, then we say that the process is in state i_k. In general, P_{ij} will mean the ***probability of transition*** from the current state i to the next state j.

In other words, the state at a particular point in time encapsulates all we need to know about the process's history to predict its future.

In essence, the state of a process at a specific moment in time encompasses all the information necessary to understand its past and predict its future. However, there are instances where a Markov process inadequately represents the patterns we seek to identify due to our inability to observe the true states of the system fully. To address this limitation, we can acknowledge the presence of hidden information, which is the premise of the hidden Markov Model (HMM).

A Hidden Markov Model is a doubly stochastic process featuring an underlying stochastic process that remains hidden (i.e., unobservable). It can only be inferred through another set of stochastic processes that yield the sequence of observed symbols. A fundamental description of a hidden Markov model can be articulated as follows:

N = number of states

T = number of observations emission parameter

$\theta i = 1 \ldots\ldots\ldots\ldots N$= for an observation associated with state i

$\varphi i = 1 \ldots\ldots\ldots N, j = 1 \ldots\ldots\ldots N$= probability of transition from state $i$to state j N-dimensional vector, composed of $\varphi i, j = 1 \ldots\ldots N$;

$\varphi i = 1 \ldots\ldots N$= the i-th row of the matrix $\varphi i = 1 \ldots\ldots N, j = 1 \ldots\ldots N$(sum of it is 1)

$x_t = 1 \ldots\ldots T$ = (hidden) state at time t

$y_t = 1 \ldots\ldots T$= observation at time t probability distribution

$F(y/\theta)$= of an observation parameterized on θ

HMMs did not gain much popularity until the early 1970s when Baum et al. successfully applied the technique to speech recognition by developing an efficient training algorithm for

HMMs. Still, since then, it has been applied differently by different writers and authors, e.g., A Hidden Markov Model inference approach to testing the Random Walk Hypothesis: Empirical evidence from the Nigerian Stock Market (Nkemnole. E., 2016). A hidden semi markov model with missing data and multiple observation sequences for mobility tracking (Shun-Zheng, Hisashi Kobayashi 2001), Parametric Hidden Markov Model for gesture recognition(Wilson et al., 1999), Segmentation of brain MR images through a hidden Markov random field mode land the expectation-maximization algorithm(Zhang et al., 2001), Folk music classification using Hidden Markov Models(Chai and Vercoe, 2001), Recognizing of humans from their gaits(Narayanan, 2003) and a host of many others.

The hidden Markov model (HMM) technique has become one of the most successful techniques in estimation and recognition (e.g. speech recognition), decoding in digital communication, and time series analysis. Hidden Markov Models are equivalently defined through a functional representation known as the **state space** model. The state-space model of an HMM certainly is one of the concepts of statistical model processing that has had a profound practical impact in recent years.

The state space model (Doucet and Johansen, 2009) of an HMM is represented by the following two equations:

(State equation) $x_t = f(x_{t-1}) + w_t$ (1)

(Observation equation) $y_t = g(x_t) + v_t$ (2)

where f and g are either linear or nonlinear functions, w_t and v_t are white noise processes. Models represented by (1) and (2) are referred to as state space models, and this includes a class of HMMs with non-linear Gaussian state space models, such as the stochastic volatility model and the bearings-only tracking model.

This research will investigate the hidden Markov model to determine the most appropriate technique for modeling time series data with missing values. These models can then be used to estimate the missing values. When one or more observations are missing it may be necessary to estimate the model and obtain estimates of the missing values. By including estimates of missing values, a better understanding of the nature of the data is possible with more accurate

forecasting. One of the key steps in time series analysis is to identify and correct obvious errors and fill in any missing observations, which will enable comprehensive analysis. In particular, different patterns and frequencies of missing values will be considered using many simulated data sets.

The problem of this research work statesthat in our society, we often have to analyze and make inferences using real data available for collection. Ideally, the data are carefully collected and have regular patterns with no outliers or missing values. In reality, this does not always happen, so an essential part of the initial examination of the data is to assess the quality of the data and to consider modifications where necessary. The treatment of missing data has been an issue in statistics for some time, but it has come to the forefront in recent years. This occurs because an observation may not be made at a particular time owing to faulty equipment, lost records, a natural disaster, or a mistake that cannot be rectified until later. A common problem frequently encountered is missing observations for time series data since the data are records taken through time. When one or more observations are missing, it may be necessary to estimate the model. There are various methods available for estimating missing values for time series data. However, a comparison of different methods for different types of data sets and positions for the missing data is lacking. This research has provided a comparison by using a variety of simulated data sets with missing values in different locations.

This research, therefore, aimed to model methods of missing data using the Hidden Markov Model (HMM). The specific objectives are to:

i. compare various methods of estimating missing data using sample sizes of 50, 200, and 1000.
ii. compute Root Mean Squared Error, Mean Absolute Error, and Mean Absolute Percentage Error for the different methods to determine the best method of estimating missing values.
iii. compute and compare the goodness of fit of estimated HMMs for the actual series and missingness series predicted.
iv. compute the estimates and draw a conclusion.

1.2 RESEARCH QUESTIONS

The following are relevant research questions answered in this work:

i. In what way(s) will the comparison analysis/approach be executed?
ii. What way(s) do we use to compute the RMSE, MAE, and MAPE for the different methods?
iii. How will the HMMs estimates be computed and compared for both series predicted?
iv. How will the estimates be analyzed and appropriate conclusions be reached?

2 Materials and Methods

This research focuses on the data source and the research methodology employed. The statistical model utilized includes the Hidden Markov Model (HMM) along with various methods for estimating missing values, such as Mean Imputation (MI), Median Imputation (MDI), linear interpolation (LI), and the Kalman Filter Algorithm (KAL).

In statistics, missing data, or missing values, refer to instances where no data value is available for a particular variable in an observation. The occurrence of missing data is quite common and can significantly impact the conclusions drawn from the dataset.

2.1 TYPES OF MISSING DATA

The types of missing data are as follows:

i. Missing Completely At Random (MCAR);

ii. Missing Not At Random (MNAR);

iii. Missing At Random (MAR);

2.1.1 Missing Completely At Random (MCAR):

When data is classified as missing completely at random (MCAR), it indicates that any values of other variables do not influence the absence of data. In this situation, the probability of missing data remains constant across observations; any individual data point is equally likely to be missing. The mechanism behind the missing data is not related to the underlying model, allowing us to disregard the missing values in our analysis.

The MCAR pattern arises when the missing values for a variable (x) are not dependent on any values from other measured variables, including variable (x). Consequently, the observed values represent a random subset of a complete dataset.

2.1.2 Missing Not At Random (MNAR):

Data are categorised as missing not at random (MNAR) when they do not follow the patterns of missingness considered random, such as missing completely at random (MCAR). In the case of MNAR data, underlying models likely explain the reasons for the missingness. If we understand these models, we can derive appropriate estimators for the model parameters that govern our data. The mechanism responsible for the missing data in this context is non-ignorable. Specifically, an MNAR pattern occurs when the missing values for a variable (x) are related to the values of that same variable (x).

2.1.3 Missing At Random (MAR):

Data are considered Missing At Random (MAR) when the likelihood of missing data on a variable (Y) does not depend on its value, provided that other variables in the design are accounted for. If the data meet at least the MAR criterion, the mechanism behind the missingness can be deemed ignorable. Specifically, the MAR pattern arises when the missing values of a variable (x) are associated with other measured variables. Still, the absence of data does not result from the variable (x) itself. Furthermore, the variables with missing values can be estimated using other available measures, such as a regression equation. Missing data can be treated as ignorable if the MCAR or MAR assumption holds.

2.2 SOURCE OF DATA

The nature of this study demanded the use of simulated data, which are derived from the Additive Model$X_t = M_t + S_t + e_t$.

2.3 SIMULATION SET-UP AND TECHNICAL STEPS FOR THE SELECTED STATISTICAL TESTS IN *R* ENVIRONMENT

The datasets were generated using the *R* statistical software package. Data are simulated from the Additive Model: $X_t = M_t + S_t + e_t$.

The trend-cycle component M_t used are;

i. Linear: $M_t = a + bt$ with $a = 1$ and $b = 2$

ii. Quadratic: $M_t = a + bt + ct^2$ with $a = 1$, $b = 2$ and $c = 3$

iii. Exponential: $M_t = be^{ct}$ with $b = 10$ and $c = 0.02$

It is assumed that $e_t\ N(0,1)$ for the additive model. $S_t, wheret = 1,2,\dots,12$ denote the seasonal indices, Table 1 presents the seasonal indices used for simulation.

Table 1: Seasonal Indices used for Simulation

$\boldsymbol{t}$	1	2	3	4	5	6	7	8	9	10	11	12
$\boldsymbol{S_t(Add)}$	1.1	1.2	1.1	1.0	1.0	1.0	0.9	0.9	0.9	0.9	1.1	1.1

Note: *$S_t(Add)$ denotes Seasonal indices for the Additive model*

To begin with, we will load several packages from the R library, including car (for simulating missing data), TestDataImputation (for mean and median imputation), interp (for linear interpolation and the Kalman filter algorithm), as well as HiddenMarkov and DepmixS4 (for implementing the Hidden Markov Model). Next, we will define the previously mentioned parameters. Following that, we will adjust our sample sizes as outlined below:

a) The small sample size consists of 12% missingness sets of 50;

b) The moderate sample size; consists of 12% missingness sets of 200;

c) The large sample size consists of 12% missingness sets of 1000.

The sample sizes mentioned above are varied across additive models with linear, quadratic, and exponential trends. This results in nine (9) series, each simulated with 12% missingness. The missing values in these series are estimated using four different methods: Mean Imputation (MI), Median Imputation (MDI), Linear Interpolation (LI), and Kalman Filter (KAL). These estimations will be compared against nine actual series. The corresponding R codes for Mean Imputation (MI), Median Imputation (MDI), Linear Interpolation (LI), and Kalman Filter (KAL) are available.

2.4 METHODS OF ESTIMATING MISSING VALUES

2.4.1 Mean Imputation

Mean imputation (MI) is a straightforward technique for handling missing data. It involves substituting the missing value with the mean of the observed values that precede the missing position(s). This is achieved by calculating the available values and dividing it by the number of observations before the missing data. This method preserves the sample size and is user-friendly (Eekhout et al., 2013).

$$MI = \hat{X}_{(i-1)s+j} = \frac{1}{(i-1)s+j-1}\left[X_1 + X_2 + X_3 + \cdots + X_{(i-1)s+j-1}\right] \quad \mathbf{3}$$

$$MI = \frac{1}{n}\sum_{t=1}^{n} X_t \quad \mathbf{4}$$

where $n = (i-1)s + j - 1$is the number of observations preceding the missingobservation(s).

2.4.2 Median Imputation

Series median (MDI) estimates the missing value with the Median of the remaining series.

Symbolically, the series median is given by

$$MDI = \hat{X}_{(i-1)s+j} = X_{\frac{N}{2}th} \quad \textbf{5}$$

Where $N = Totalnumberofobservationexcludingthemissingvalues$

2.4.3 Linear Interpolation

This method replaces missing values using a linear interpolation. It utilizes the last valid value before the missing value and the first value after the missing value for the interpolation. The linear interpolation (LI) for estimating missing values is given by

$$LI = \hat{X}_{(i-1)s+j} = \frac{1}{2}\left(X_{(i-1)s+j-1} + X_{(i-1)s+j+1}\right) \quad \textbf{6}$$

2.4.4 Kalman Filter

The Kalman filter (KAL) is a statistical algorithm that enables certain computations to be carried out for a model cast in a state space form. However, the smoothing algorithm is performed to obtain a more accurate estimate of missing values. LetY_{t-1}denote the set of past observations $\{y_1, \ldots, y_{t-1}\}$ and assume the conditional distribution of μ_t given $Y_{t-1}$is $N(\mu_t, p_t)$ where μ_t and p_t are assumed to have been determined. Hence, the celebrated Kalman filter equations for updating the missing values from time *t* to *t+1* are given by;

$$\mu_t = \mu_{t-1} + k_{t-1}v_{t-1},\ p_t = p_{t-1}(1 - k_{t-1}) + \sigma_\eta^2, \qquad k_{t-1} = p_{t-1}/f_{t-1},$$

$$v_{t-1} = y_{t-1} - \mu_{t-1}, \quad f_{t-1} = p_{t-1} + \sigma_\varepsilon^2 \quad \textbf{7}$$

For *t=1,2,…,n*, where v_{t-1} is the Kalman filter residual or prediction (signal) errors, f_{t-1} is its variance and k_{t-1} is the Kalman gain.

2.5 Comparison of Methods of Estimating Missing Values

Numerous measures are available for accessing the four methods (MI, MDI, LI and KAL) performances. We evaluate the deviation of $\hat{X}_{(i-1)s+j}$ from the actual, which can be calculated as $e_{(i-1)s+j} = X_{(i-1)s+j} - \hat{X}_{(i-1)s+j}$, to access the performance of the aforementioned methods. We compare the "Accuracy Measures" of the four methods; the method with the minimum Accuracy Measures is the best at that level. These "Accuracy Measures" are root mean squared error (RMSE), mean absolute error (MAE) and Mean Absolute Percentage Error (MAPE), which are defined as follows.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2} \quad \textbf{8}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|(A_t - F_t)| \quad \textbf{9}$$

$$MAPE = \left[\frac{1}{m_0}\sum_{k=1}^{m_0}\left|\frac{e_k}{X_k}\right|\right] \times 100 \quad \textbf{10}$$

Where A_t is the actual value in time *t*, and F_t is the forecast value in time *t*. we considered one missing value at a time for different, $m_0 < n$ position, $n > 1$.

After that, each of the best method series at different sample sizes and trend structures is modelled using HMM alongside the actual series. The goodness of fit of the models is then compared.

2.5.1 BAUM-WELCH ALGORITHM

The parameters of a Hidden Markov Model (HMM) are estimated using the Baum-Welch algorithm. This algorithm utilizes the well-established Expectation-Maximization (EM) technique to derive the maximum likelihood estimates of the parameters of a Markov model, based on a set of observed feature vectors. The EM algorithm comprises two steps: the E-step and the M-step. In the E-step, the conditional expectation is computed.

E-step: the expected likelihood, $Q(\theta \vee \theta^{(K)})$

$Q(\theta|\theta^{(K)}) = E_{g(k)}(\log f(x|\theta^{i})|y,\theta)$

is computed, where $\theta^{(k)}$ is the current parameter estimate.

M-step: In the M-step, a new parameter estimate $\theta^{(k+1)}$ is obtained by maximizing$Q$).

The E-step and M-step are repeated until some stooping criterion is met, such as

$|\theta^{n+1} - \theta^{n}| < Q$, for some specified Q, obtaining suitable initial parameters inclusive.

3 Results and Discussion

3.1 Data presentation

The data utilized in this study are simulated. The simulated data used were derived from the Additive Model:$X_t = M_t + S_t + e_t, where t = 1,2,3,\dots,9$. Nine series were considered from the additive model; the series were partitioned into three groups according to their sample sizes, i.e., 50, 200, and 1000. Each group comprises three series that follow three different trend structures, namely linear, Quadratic, and Exponential (see the appendix for the simulated series plots).

3.2 COMPARATIVE ANALYSES OF SOME STATISTICAL METHODS FOR ESTIMATING MISSING VALUES

As mentioned earlier, this study compares four statistical methods using their loss functions or accuracy measures (RMSE, MAE, and MAPE), with the one with the least being considered the best.

3.2.1 Evaluation of the Methods when Sample Size is Small

Table 2 summarises the accuracy measures for the four methods of estimating missing values when the sample size is small for the selected trend (Linear, Quadratic, and Exponential). The results in Table 2 indicate that KAL recorded two (2) lowest values of the accuracy measures (MAE and MAPE) out of the three measures considered for the linear trend. Also, it indicates that MDI recorded two (2) lowest values of the accuracy measures (MAE and MAPE) out of the three accuracy measures considered for the quadratic and exponential trends. This implies that for a small sample size, KAL performed best for linear trend structures, and MDI performed best for quadratic and exponential trend structures.

Table 2: Summary result of estimation of missing value

Trend Component	Accuracy Measures	Estimation Method			
		MI	**MDI**	**LI**	**KAL**
Linear	RMSE	**2.649548**	2.650424	2.771986	2.672517
	MAE	1.0007	1.0007	0.9872792	**0.959303**
	MAPE	0.007527	0.007416	0.008269	**0.006784**
	Accuracy Score	1/3	0	0	**2/3**
Quadratic	RMSE	**2319.233**	2562.877	3787.697	2577.596
	MAE	452.1502	**403.5013**	890.2353	571.0032
	MAPE	0.000439	**0.000252**	0.001352	0.000686
	Accuracy Score	1/3	**2/3**	0	0
Exponential	RMSE	127650.9	**3.753932**	1895660.8	141424.21
	MAE	31267.95	**0.8245**	268319.3	34198.6
	MAPE	36.70758	**0.000603**	140.8656	45.02647
	Accuracy Score	0	**3/3**	0	0

Source: Researchers' compilations

3.2.2 Evaluation of the Methods When Sample Size is Moderate

Table 3 summarises the accuracy measures for the four methods of estimating missing values when the sample size is moderate for the selected trend (Linear, Quadratic, and Exponential). The results in Table 3 indicate that KAL recorded two (2) lowest values of the accuracy measures (RMSE and MAE) out of the three measures considered for the linear trend. The results also indicate a conflicting performance for MDI, LI, and KAL, as they recorded the lowest value in MAPE, MAE, and RMSE, respectively. This implies that MDI, LI, and KAL best estimate missing values when the sample size is moderate, and the trend follows a quadratic structure. In addition, MDI recorded the lowest values of the accuracy measures (RMSE, MAE, and MAPE) for the exponential trend. This implies that MDI outperformed others when the sample is small and the trend follows an exponential structure.

Table 3: Summary result of estimation of missing value

Trend Component	Accuracy Measures	Estimation Method			
		MI	**MDI**	**LI**	**KAL**
Linear	RMSE	1.553115	1.759002	2.240325	**1.545315**
	MAE	0.331571	0.3662652	0.4371246	**0.328248**
	MAPE	0.000793	**0.000622**	0.00079	0.000789
	Accuracy Score	0	**1/3**	0	**2/3**
Quadratic	RMSE	3778.192	4161.102	3982.469	**3661.596**
	MAE	1101.555	1160.667	**987.9734**	1050.958
	MAPE	0.00484	**0.003565**	0.003978	0.005102
	Accuracy Score	0	**1/3**	**1/3**	**1/3**
Exponential	RMSE	2534654	**2512328.8**	2720963.5	2546924.3
	MAE	306517.6	**275066.47**	426979.97	314754.62
	MAPE	360.5943	**0.001388**	1097.168	281.9716
	Accuracy Score	0	3/3	0	0

Source: Researchers' compilations

3.2.3 Evaluation of the Methods When Sample Size is Large

Table 4 summarises the accuracy measures for the four methods of estimating missing values when the sample size is large for the selected trend (Linear, Quadratic, and Exponential). Similarly to the reports in Table 2, our results (Table 4) indicate that KAL recorded two (2) lowest values of the accuracy measures (RMSE and MAE) out of the three accuracy measures considered for the linear trend. Also, it indicates that MDI recorded two (2) lowest values of the accuracy measures (MAE and MAPE) out of the three accuracy measures considered for the quadratic and exponential trends. This implies that for large sample sizes, KAL performed best

for linear trend structures, and MDI performed best for quadratic and exponential trend structures.

Table 4: Summary result of estimation of missing value

Trend Component	Accuracy Measures	Estimation Method			
		MI	**MDI**	**LI**	**KAL**
Linear	RMSE	5717924	10.95917	2.983935	**2.35911**
	MAE	1667748	9.043338	0.8872693	**0.7748485**
	MAPE	4227.582	**0.009937**	0.017085	0.011757
	Accuracy Score	0	1/3	0	**2/3**
Quadratic	RMSE	**3055.605**	3136.429	3695.043	3099.308
	MAE	976.3247	**973.6958**	1088.255	984.0337
	MAPE	0.317189	**0.26505**	0.473352	0.331552
	Accuracy Score	1/3	**2/3**	0	0
Exponential	RMSE	**1524381**	1524916.9	8762248.5	1570153.9
	MAE	200130.3	**173719**	2279966.1	234634.2
	MAPE	281.448	**0.001336**	102.4718	802.0694
	Accuracy Score	1/3	**2/3**	0	0

Source: Researchers' compilations

3.2.4 **Hidden Markov Model (HMM) Estimations**

HMMs (for actual data and estimated missing values series) were applied to the simulated data at each level of trend components for all the sample sizes considered: small (50), moderate (200), and large (1000). The best-performing methods (MDI, LI, and KAL) at different comparison levels (see Table 2 to Table 4) were adopted to estimate the missing values at different levels based on their performances.

Table 5: Parameters Estimation Summary of HMMs

Sample Size	Trend	Model	No. of States	No. of iteration	LogLikelihood( ), AIC[ ]	Initial Probability)	Model Parameters $\lambda = (A, B)$
Small	Linear	Actual	3	55	(-650.3) [4.567]	(0.5, 0.5, 0)	(0.075, 2.2E-04)
		KAL	3	92	(-685.6) [5.874]	(0.5, 0.5, 0)	(0.07, 0.08)
	Quadratic	Actual	3	44	(-670.6) [4.70]	(1, 0, 0)	(0.62, 0.04)
		MDI	3	59	(-674.2) [4.75]	(1, 0, 0)	(0.68, 0.003)
	Exponential	Actual	3	47	(-686.2) [4.57]	(1, 0, 0)	(0.29, 0.23)
		MDI	3	40	(-697.2) [4.66]	(1, 0, 0)	(0.08, 0.34)
Moderate	Linear	Actual	3	82	(-793.0) [6.09]	(0.5, 0.5, 0)	(0.007, 3.4E-04)
		KAL	3	85	(-790.07) [6.13]	(0.5, 0.5, 0)	(0.12, 2.3E-04)
	Quadratic	Actual	3	79	(-858.2) [6.70]	(1, 0, 0)	(0.24, 0.04)
		MDI	3	87	(-861.0) [6.71]	(1, 0, 0)	(0.48, 0.054)
		LI	3	96	(-863.9) [6.69]	(1, 0, 0)	(0.45, 0.035)
		KAL	3	102	(-866.8) [6.68]	(1, 0, 0)	(0.34, 0.04)
	Exponential	Actual	3	76	(-872.5) [6.72]	(1, 0, 0)	(0.38, 0.021)
		MDI	3	84	(-885.7) [6.78]	(1, 0, 0)	(0.45, 0.025)
Large	Linear	Actual	3	222	(-1789.9) [6.75]	(0.5, 0.5, 0)	(0.05, 0.004)
		KAL	3	234	(-1793.1) [6.86]	(0.5, 0.5, 0)	(0.07, 4.1E-4)
	Quadratic	Actual	3	245	(-1756.1) [6.78]	(1, 0, 0)	(0.12, 0.05)
		MDI	3	250	(-1876.1) [6.98]	(1, 0, 0)	(0.23, 0.006)
	Exponential	Actual	3	234	(-1955.1) [6.08]	(1, 0, 0)	(0.26, 0.036)
		MDI	3	254	(-1986.7) [6.28]	(1, 0, 0)	(0.36, 0.045)

Source: Researchers' compilations

Furthermore, HMMs were fitted for the series estimated by these methods (MDI, LI, and KAL) and for the actual series. The updated parameters estimations were obtained using the Baum-Welch algorithm; Table 5 presents the model's estimations summary. Log-likelihood and AIC were also estimated to evaluate the model's goodness-of-fit. From Table 5 above, the estimated AICs and parameter $\lambda = (A, B)$ of all HMMs fitted for the "missing values series estimated" show no or less divergence from the actual series models. Hence, MDI, LI, and KAL methods of

estimating missing values are best employed in different scenarios, as highlighted in sections 3.2.1 to 3.2.3.

4 Conclusions

The study's aim and objectives have been principally accomplished. The analyses show the comparative results of the missing value estimations under three different trend structures for small, moderate, and large sample sizes.

In the analyses, the methods of Mean Imputation (MI), Median Imputation (MDI), Linear Interpolation (LI), and Kalman filter algorithm (KAL) were tested under three different trend structures, namely Linear, Quadratic and Exponential for small, moderate and large sample sizes. The results for the four tests when the sample size is small (50) depict each method's potency as follows: 66.7% and 33.3% potencies for Kalman filter algorithm (KAL) and Mean Imputation (MI) respectively under Linear trend structure; however LI and MDI recorded zero potencies; 66.7% and 33.3% potencies for Median Imputation (MDI) and Mean Imputation respectively under Quadratic trend structure however LI and KAL record zero potencies; and 100% for Median Imputation (MDI) under Exponential trend structure whereas others (MI, LI and KAL) recorded zero.

In addition, the results for the four methods of missing values estimation when sample size is moderate (200) depict each method's potency as follows: 66.7% and 33.3% for KAL and MDI respectively under Linear trend structure while MI and LI recorded zero; MDI, LI and KAL recorded potency of 33.3% each under Quadratic trend structure while MI recorded zero; and 100% for Median Imputation (MDI) under Exponential trend structure whereas others (MI, LI, and KAL) recorded zero.

Lastly, the results for the four methods of estimating missing values when the sample size is large (1000) depict each method's potency as follows: 66.7% and 33.3% for KAL and MDI respectively, under linear trend structure while MI and LI recorded zero, 66.7% and 33.7% for MDI and MI respectively under Quadratic and Exponential trend structures while LI and KAL recorded zero.

The following recommendations are made to support further research in this field of stochastic statistics:

i. Kalman filter algorithm should be employed to estimate missing values at any level of sample sizes under linear trend structures.

ii. Median imputation should be employed at any level of sample sizes under quadratic and exponential trend structures.

Disclaimer (Artificial intelligence)

Option 1:

I hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

References

[1] Acock, A .C. (2005) "Working with missing values", Journal of Marriage and the Family, 67:1012-28.

[2] Allison, P.D. (2001). Missing Data, Thousand Oaks, C.A: Sage Publications.

[3] Alshamaa, D., Chkeir, A., Mourad-Chehade, F., Honeine, P.(2019): *A Hidden Markov Model for Indoor Trajectory Tracking of Elderly People. Andrew C. Harvey (2001): Forecasting, Structural Time Series Models and the Kalman Filter.*

[4] Barbara Engelhardt Scribes: Ani Mohan, Lucas Spangher, Shiwen Zhao, Xiaoyang Zhuang (2013) Lecture notes on Hidden Markov Models and State Space Models.

[5] Bert Huang (2015): *Hidden Markov Models* (YouTube Video). Available at: https://www.youtube.com/watch?v=9yl4XGp5OEg&t=11s Published: 3rd November, 2015.

[6] Chris Chatfield (2003): Analysis of Time Series an Introduction.

[7] Curtis F. Gerald & Patrick O. Wheatley 1994): Applied Numerical Analysis.

[8] David Sheung Fung (2006), Methods for the Estimation of Missing values in time series analysis.

[9] Dhok, S., Bamnote, G. (2012): *Credit Card Fraud Detection Using Hidden Markov Model* International Journal of Advanced Research in Computer Science Volume 3, No. 3, May-June 2012.

[10] Eivind Damsleth (1979): A method to find the optimal linear combination of the forecast and back forecast for missing values in a time series.

[11] G. Gardner A. C Harvey and G.D.A Phillips (1980): Algorithm of Likelihood Function for A Stationary Autoregressive Moving Average (ARMA).

[12] Gomez et al (1995): The bootstrap Method of inputting Missing Natural Resource Inventory Data.

[13] Henry Maltby *et al* (2018): *Markov Chains.* Available at https://brilliant.org/wiki/markov-chains/#markov-chain.

[14] Hung-Wen Yeh, Wenyaw Chan & Elaine Symanski (2012) Journal on Intermittent Missing Observations in Discrete-Time Hidden Markov Models.

[15] James D. Hamilton (1994): Time series Analysis.

[16] Kalman (1960): Articles to missing values in time series state space modeling.

[17] Mathematicalmonk (2011): *(ML 14.1) Markov models – motivating examples* (YouTube Video). Available at: https://www.youtube.com/watch?v=7KGdE2AK_MQ Published: 6th July, 2011.

[17] Mathematicalmonk (2011): *(ML 14.2) Markov chains (discrete time) (part 1)* (YouTube Video). Available at:https://www.youtube.com/watch?v=WUjt98HcHlk&t=2s Published: 6th July, 2011.

[18] Mathematicalmonk (2011): *(ML 14.3) Markov chains (discrete time) (part 2)* YouTube Video). Available at:https://www.youtube.com/watch?v=j6OUj9tleVM&t=13s Published: 6th July, 2011.

[19] Mathematicalmonk (2011): *(ML 14.4) Hidden Markov models (HMMs) (part 1)* (YouTube Video). Available at:https://www.youtube.com/watch?v=TPRoLreU9lA&t=70s Published: 7th July, 2011.

[20] Mathematicalmonk (2011): *(ML 14.5) Hidden Markov models (HMMs) (part 2)* (YouTube Video). Available at:https://www.youtube.com/watch?v=M_IIW0VYMEA&t=1s Published: 7th July, 2011.

[21] Mathematicalmonk (2011): *(ML 14.6) Forward-Backward algorithm for HMMs* (YouTube Video). Available at:https://www.youtube.com/watch?v=7zDARfKVm7s Published: 7th July, 2011.

[22] Mathematicalmonk (2011): *(ML 14.7) Forward algorithm (part 1)* (YouTube Video). Available at:https://www.youtube.com/watch?v=M7afek1nEKM&t=9s Published: 7th July, 2011.

[23] Mathematicalmonk (2011): *(ML 14.9) Backward algorithm* (YouTube Video). Available at:https://www.youtube.com/watch?v=jwYuki9GgJo&t=38s Published: 7th July, 2011.

[24] Michael A.Wincek and Gregory (1986): An explicit procedure to obtain the exact maximum likelihood estimates of the parameters in a regression model with ARMA time series errors with possibly non-consecutive data.

[25] Nkemnole, E., Abass, O., Kasumu, R. (2013): *Parameter Estimation of a class of hidden Markov model with Diagnostics,Journal of modern applied statistical methods,* Volume 12, Issue 1, Article 21

[26] Nkemnole, E. (2016): *A Hidden Markov Model inference approach to testing the Random Walk Hypothesis: Empirical evidence from the Nigerian Stock Market.* Journal of Economic and Financial Sciences. 9. 696-713. 10.4102/jef.v9i3.66.

[27] Osvaldo Ferreiro (1987): The Estimation of Missing Observation in Stationary Time Series for Autoregressive Moving Average Models.

[28] Peter J. Brockwell & Richard A. Davis (1991): Time series: Theory and Methods

[29] Quant Education (2014): *Introduction to Markov Chains*.(YouTube Video). Available at: https://www.youtube.com/watch?v=F23AgV1Zk-E Published: 4th April, 2014.

[30] Stamp, M. (2018): *A Revealing introduction to Hidden Markov model.Department of Computer science, San Jose State University*.

[31] Volker Tresp and Reimar Hofman (1998). Techniques for Nonlinear Time series Prediction with Missing and Noisy Data.

[32] Wojciech Salabun et al, Int. J. Computer Technology & Applications, Vol. 5 (4),1384-1391

[33] Wikipedia (2018): *Hidden Markov model.* Available at: https://en.wikipedia.org/wiki/Hidden_Markov_model Published: 4th September, 2018

[34] Wikipedia (2017): *Markov chain.* Available at: https://en.wikipedia.org/wiki/Markov_chain Published: February, 2017

[35] Wikipedia (2018): *Markov model.* Available at: https://en.wikipedia.org/wiki/Markov_model published: 17th June, 2018

[36] Wilson, D., Bobick, A. (1999): *Parametric Hidden Markov Models for Gesture Recognition. Pattern Analysis and Machine Intelligence*, IEEE Transactions on. 21. 884 - 900. 10.1109/34.790429.

[37] Yu Luo, Limin Du (2003) carried out their study on A Hidden Markov Model-Based Missing Data Imputation Approach

[38] Zhang, Y., Brady, M., Smith, S., (2001): S*egmentation of brain MR images through a hidden Markov random field mode land the expectation-maximization algorithm,* IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 20, NO. 1

[39]Agbailu, A. O., Seno, A., & Clement, O. O. (2020). Kalman filter algorithm versus other methods of estimating missing values: time series evidence. *Studies*, *4*(2), 1-9.