Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v16n3p764

An application of Poisson hidden Markov model to forecast endometrial cancer cases in Odisha, India

By Swain, Tripathy, Jena

15 December 2023

Electronic Journal of Applied Statistical Analysis
Vol. 16, Issue 3, 2023, 764-783
DOI: 10.1285/i20705948v16n3p764

An application of Poisson hidden Markov model to forecast endometrial cancer cases in Odisha, India

Prafulla Kumar Swain[a], Manas Ranjan Tripathy*[b], and Diptismita Jena[b]

[a]*Department of Statistics, Utkal University, Bhubaneswar, Odisha, India - 751004*
[b]*Department of Statistics, Ravenshaw University, Cuttack, Odisha, India - 753003*

15 December 2023

The time series model is often used for forecasting but is inadequate when the data series is unbounded and over-dispersed in nature. Moreover, if the observations are serially dependent, then a Markov-dependent mixture model, i.e., the Poisson Hidden Markov model, can be used. The objective of this study was to apply the Poisson Hidden Markov model to forecast the monthly new endometrial cancer cases. The monthly total number of registered endometrial cancer cases has been collected from a local cancer hospital in Odisha from January 2017 to December 2021. In this paper, we have applied the Poisson hidden Markov model to forecast the number of endometrial cancer cases for the next 12 months. The three-state Poisson hidden Markov model was found to be the best-fitted model and was used to forecast the endometrial cancer cases for the next 12 months. The results showed that the number of endometrial cancer cases is most likely to be in state 2 in January and in state 3 for the rest of the months in 2022. The monthly forecasted mean of endometrial cancer cases varies between 34 [95% CI: 31.253-37.109] and 38 [95% CI: 34.920-41.032] for the year 2022. Based on the evidence, this study reveals that the average monthly endometrial cancer cases will increase in the future. From the findings, it is also suggested that the three-state hidden Markov model can be used to fit and forecast the distribution of the number of endometrial cancer cases.

*Corresponding author: manasranjantripathy402@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

keywords: Mixture model, Hidden Markov model, Poisson hidden Markov model, Endometrial cancer, Forecasting.

1 Introduction

Gynaecological cancers are among the most common cancers in women and hence an important public health issue. Due to the lack of cancer awareness, variable pathology, and dearth of proper screening facilities in developing countries such as India, most women report at advanced stages, adversely affecting prognosis and clinical outcomes. Cancer can affect the uterus which is lined with a special tissue called the endometrium. When cancer grows in this lining, it is called endometrial cancer (EC). Most uterine cancers are EC. If left untreated, EC can spread to the bladder or rectum, or it can spread to the vagina, fallopian tubes, ovaries, and more distant organs. Fortunately, EC grows slowly and, with regular check-ups, is usually found before spreading too far. Hence it is very important to detect and treat EC cases in early stages.

In recent years, a high morbidity rate of EC cases has been experienced in Odisha (Pattnaik et al., 2020). This may be due to the fact that by the time the patients sought medical condition, the cancer had already reached an advanced stage. It is very challenging for healthcare providers to treat such a large number of patients when most of them are suffering at such an advance stage. Therefore, it is very important for oncologists to have a prior idea of the number of patients they need to serve so as to provide them with relatively better treatment and equip the infrastructure adequately and be prepared to handle the patient influx.

A Markov model deals with a stochastic process for randomly changing systems or values with the assumption that the future states do not depend on its past states. Often, the change of states in a data sequence may not be known or hidden, but they can be inferred later. A hidden Markov model (HMM) can be used to reveal these hidden states. Further HMMs are used to analyze, weakly dependent time series datasets (Laverty et al., 2002) and are flexible for both univariate and multivariate discrete-valued data series.

An HMM consist of two parts, the first part is an unobserved or hidden finite state of a Markov chain and the second part is the observed value of a stochastic process. Hence, an HMM generates the observations, which depends upon the state of an underlying and unobserved Markov process. When HMM is influenced by a Poisson process, we may have a mixed effect model known as Poisson hidden Markov model (PHMM) which is constituted by a discrete Markov model and a Poisson regression model (Zucchini and MacDonald, 2009). A detailed background for using different models which deals with discrete valued count data in different situations is given below.

When the data series is unbounded in nature, it is inappropriate to use the standard time series model i.e., autoregressive moving average (ARMA) model, to describe its characteristics (Zucchini and MacDonald, 2009). The Poisson distribution can be used to describe the unbounded count data with a low probability of occurrences and it is characterized by equal mean and variance. However, this may not always be true in practice. Sometimes, the distributional variance is greater than its mean, indicating an

over-dispersed distribution. In such cases, an independent mixture of Poisson model can be used as it deals with the unobserved heterogeneity in the population (George and Thomas, 2018).

If the count observations change due to random factors, then the assumption of independence no longer holds and an independent mixture of Poisson model cannot be used. In those situations, the Markov dependent mixture model (i.e., PHMM) has been suggested (Paroli et al., 1994). Thus, PHMMs are extensions of mixture models and are generally used for modelling data with overdispersion and serial dependence (Zucchini and MacDonald, 2009). It is worth noting that a Poisson process is the special case of PHMM when the Markov chain has only one state.

HMMs have applications in various fields such as biomedical signal processing (Cohen, 1998), speech recognition (Juang and Rabiner, 1991; Rabiner, 1989), and DNA sequencing (Churchill, 1989). Sarvi et al. (2017) predicted the air pollution indicator PM2.5 using HMM. It is also used to study count data applications such as predicting the number of traffic accident cases (Laverty et al., 2002), death counts (George and Thomas, 2018), and the number of insurance claims (Paroli et al., 1994). Nowadays, HMMs are commonly used for disease mapping (Green and Richardson, 2012) and studying infection behavior (Cooper and Lipsitch, 2004).

In this context, the serial dependence may arise through some unobserved disease progression process, particularly in advanced stages, and the HMM would seem to be more appropriate for explaining the number of EC patients. Although several applications of HMMs have been experienced in different areas, they are not yet common in the medical field, especially for forecasting the number of patients expected to come to the hospital.

Therefore, the main aim of this paper is to demonstrate the use of PHMM to identify the hidden states and forecast the number of EC cases for the next 12 months based on the number of EC cases reported on monthly basis from January 2017 to December 2021.

The rest of this paper is organized as follows: Section 2 includes the methodology and data sources. In Section 3, independent mixtures of Poisson distributions and PHMMs with different states are fitted, and the performance of each model is evaluated. The discussion and conclusion are presented in Sections 4 and 5, respectively.

2 Materials and Methods

2.1 Mixture Model (MM)

A mixture distribution consists of several component distributions. Let us consider Y_t be a random variable with the observed number of EC cases (say y_t) at different time $t(t = 1, 2, 3, \ldots, n)$. Let $\delta_1, \delta_2, \delta_3, \ldots, \delta_m$ (s.t. $\sum_{i=1}^{m} \delta_i = 1$) be the assigned probabilities of m component distributions with the probability function $p_1, p_2, p_3, ..., p_m$ respectively.

Now, if Y_t follows a mixture distribution, then it has the probability function as follows;

$$P(Y_t = y_t) = \sum_{i=1}^{m} \delta_i \, p_i(y_t) \tag{1}$$

The k^{th} moment of this distribution is as follows:

$$E({Y_t}^k) = \sum_{i=1}^{m} \delta_i \, E({Y_{t_i}}^k) \tag{2}$$

The parameter estimation of the mixture distribution is done by maximum likelihood method of estimation (MLE). The likelihood function of the mixture model having m component distributions is given by;

$$L(\theta_1, \theta_2, ..., \theta_m | y_1, y_2, ..., y_n) = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i \, p_i(y_j, \theta_i) \tag{3}$$

Where; $\theta_1, \theta_2, ..., \theta_m$ are the parameter vectors of the respective component distributions.

Here we consider that each EC patient count is generated by one of m Poisson distributions with mean $\lambda_1, \lambda_2, ..., \lambda_m$, where the choice of mean (λ_i) is selected with probability δ_i $(i = 1, 2, ..., m)$. Then the proposed model in Eq. (1) is converted into Poisson independent mixture model and can be defined as;

$$P(Y_t = y_t) = \sum_{i=1}^{m} \delta_i \frac{e^{-\lambda_i} \lambda_i^{y_t}}{(y_t!)} \tag{4}$$

Then the likelihood function of Poisson independent mixture model is;

$$L(\lambda_1, \lambda_2, ..., \lambda_m; \delta_1, \delta_2, \ldots, \delta_m | y_1, y_2, ..., y_n) = \prod_{t=1}^{n} \sum_{i=1}^{m} \delta_i \, \frac{e^{-\lambda_i} \lambda_i^{y_t}}{(y_t!)} \tag{5}$$

Where; $\delta_i, \lambda_i > 0$ for $i = 1, 2, \ldots, m$ with the constraint $\sum_{i=1}^{m} \delta_i = 1$. Here, each component specified by one parameter, hence in total $2m$-1 independent parameters are to be estimated.

2.2 Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is usually considered a type of dependent mixture model. The distribution in this model generates observations dependent on the state of an underlaying and unobserved Markov process. Let Y_t be the observed number of EC patients during the time interval (t-1, t] and C_t be the unobserved parameter process satisfying the Markov property. Then the distribution of Y_t depends only on current state C_t and not on previous states or observations. If $Y^{(t)}$ and $C^{(t)}$ represents the histories from time 1 to t, then the HMM can be expressed as;

$$P\Big(C_t | C^{(t-1)}\Big) = P\Big(C_t | C_{t-1}\Big); \ \textbf{where} \ \text{t=2,3, ...} \tag{6}$$

$$P\Big(Y_t | Y^{(t-1)}, C^{(t)}\Big) = P\Big(Y_t | C_t\Big); \ \text{where} \ \text{t} \in \text{N} \tag{7}$$

Where; $C^{(t)}$=(C_1, C_2, ..., C_t), $C^{(t-1)}$=(C_1, C_2, ..., C_{t-1}) and $Y^{(t-1)}$=(Y_1, Y_2, ..., Y_{t-1}). In case when the Markov chain $\{C_t\}$ has m states then we call $\{Y_t\}$ as the m-state HMM.

Now, the probability function of Y_t given that the Markov chain is in state i at time t defined as;

$$p_i(y_t) = P\Big(Y_t = y | C_t = i\Big) \tag{8}$$

Let us define u_i as the probability that the process is in the state i at time t i.e., $u_i(t) = P(C_t = i)$ for i=1,2,...,n. Then we have;

$$\begin{aligned} P(Y_t = y_t) &= \sum_{i=1}^{m} P(C_t = i) P(Y_t = y | C_t = i) \\ &= \sum_{i=1}^{m} u_i(t) p_i(y_t) \\ &= \Big\{u_1(t), u_2(t), ..., u_m(t)\Big\} \begin{pmatrix} p_1(y_t) & ... & 0 \\ ... & ... & ... \\ 0 & ... & p_m(y_t) \end{pmatrix} \begin{pmatrix} 1 \\ . \\ . \\ 1 \end{pmatrix} \\ &= u(t)\, P(y_t)\, 1^{'} \end{aligned} \tag{9}$$

Where; $P(y_t)$ be the diagonal matrix with i^{th} diagonal element $p_i(y_t)$, u(t) be the initial distribution of the Markov chain and $1^{'} = (1 \ldots 1)^T$.

Introducing the transition probability matrix $\Gamma = (\gamma_{ij})$ of the Markov chain, we have the distribution at $t+1$ to be $u(t+1) = u(t)\Gamma \Rightarrow u(t) = u(1)\Gamma^{t-1}$. Then the Eq. (9) can be rewritten as;

$$P(Y_t = y_t) = u(1)\Gamma^{t-1} P(y_t) 1^1 \tag{10}$$

The above Eq. (10) holds if the Markov chain is homogeneous but not necessarily stationary. But, when we assume that the Markov chain is stationary, with its stationary distribution δ, then $\delta\Gamma^{t-1} = \delta$ for all t $\in$ N and

$$P(Y_t = y_t) = \delta\, P(y_t) 1^{'} \tag{11}$$

Hence the mean and variance can be defined as;

$$\begin{aligned} Mean = E(Y_t) &= \sum_{i=1}^{m} P\big(C_t = i\big)\, E\big(Y_t | C_t = i\big) \\ &= \sum_{i=1}^{m} u_i(t)\, E\big(Y_t | C_t = i\big) \\ &= \sum_{i=1}^{m} \delta_i\, E\big(Y_t | C_t = i\big) \quad \text{(in stationary case)} \end{aligned} \tag{12}$$

$$Var(Y_t) = E(Y_t^2) - (EY_t)^2 \tag{13}$$

If $E(Y_t|C_t = i) = \lambda_i$ then $E(Y_t) = \sum_{i=1}^{m} \delta_i \lambda_i$ and $E(Y_t^2) = \sum_{i=1}^{m} \delta_i(\lambda_i^2 + \lambda_i)$.

Let $y_1, y_2, y_3, \ldots.., y_n$ be the sequence of observations generated through a HMM, δ be the initial distribution of a stationary distribution and Γ be a transition probability matrix then the likelihood function of the HMM is;

$$L = \delta\, \Gamma\, P(y_1)\, \Gamma\, P(y_2) ... \Gamma\, P(y_n)\, 1^{'} \tag{14}$$

The parameter estimation of the HMM can be done by numerical maximization of the likelihood function.

2.3 Poisson Hidden Markov Model (PHMM)

A Poisson Hidden Markov Model (PHMM) is made up of a mixture of two random processes i.e., a Poisson process and a Markov process. These kinds of models are suitable to handle count-based time series data. Generally, in PHMM the k-state Markov process is hidden at each time step and one doesn't know exactly about the Markov regime or the Markov process. Then, the effect of the possible existence of each regime on the mean value is predicted by the Poisson model.

Let us consider Y_t to be a random variable with the observed number of EC cases (say y_t) at different time(s) t $(t = 1, 2, 3, \ldots, n)$ obeys a Poisson process. Therefore Y_t $(t = 1, 2, \ldots, n)$ are n independent and Poisson distributed random variables each possibly with different mean λ_t and the model becomes;

$$P(Y_t = y_t) = P(Y_t = y_t|\lambda_t) = \frac{e^{-\lambda_t} \lambda_t^{y_t}}{y_t!} \tag{15}$$

Now, to mix a discrete Markov process into a Poisson process, consider a Markov process is in a state i $(i = 1, 2, \ldots, m)$. However, we don't know in which state the Markov process is at time t, we only assume that it is influenced by the Poisson process. Then, an expression for the predicted Poisson probability of observing y_t at time t, under-laying Markov process at state i becomes;

$$P(Y_t = y_t|\lambda_{ti}, C_t = i) = \frac{e^{-\lambda_{ti}} \lambda_{ti}^{y_t}}{y_t!} \tag{16}$$

Taking a Poisson probability mass function and assuming an m state Markov process at each time t; we will get a set of m probabilities of observing y_t (each one conditional upon the Markov process being in state) as follows;

$$\begin{aligned} P\big(Y_t = y_t|\lambda_t\big) \quad = \quad & \Big\{P(Y_t = y_t|\lambda_{t1}, C_t = 1), P(Y_t = y_t|\lambda_{t2}, C_t = 2), ... \\ & ..., P(Y_t = y_t|\lambda_{tm}, C_t = m)\Big\} \end{aligned} \tag{17}$$

Hence, using the law of total probability we get the probabilities predicted by the PHMM of observing y_t at time t as follows;

$$P(Y_t = y_t) = \sum_{i=1}^{m} P(C_t = i)P(Y_t = y|C_t = i) = \sum_{i=1}^{m} \delta_{ti} \frac{e^{-\lambda_{ti}} \lambda_{ti}^{y_t}}{y_t!} \tag{18}$$

In light of Eq. (9) and (10) we can rewrite Eq. (18) as;

$$P(Y_t = y_t) = u(1)\Gamma^{t-1}P(y_t)1^{'} \tag{19}$$

Now, assuming that the Markov chain is stationary, with its stationary distribution δ and $\delta\Gamma^{t-1} = \delta$ for all t $\in$ N, then ;

$$P(Y_t = y_t) = \delta P(y_t)1^{'} \tag{20}$$

Thus, the likelihood function of the PHMM becomes;

$$L = \delta\Gamma P(y_1)\Gamma P(y_2)...\Gamma P(y_n)1^{'} \tag{21}$$

Here, P(y_t) is the diagonal matrix with i^{th} diagonal element $p_i(y_t) = \frac{e^{-\lambda_{ti}} \lambda_{ti}^{y_t}}{y_t!}$, $\Gamma = (\gamma_{ij})$ is the transition probability matrix of the Markov chain, $u(1)\Gamma^{t-1}$ is the distribution at time t and $1^{'} = (1...1)^T$.

The parameters of the PHMM viz. δ, λ and Γ can be estimated by numerical maximization of the likelihood function (Eq. 21) subject to the constraints; λ_i, δ_i, $\gamma_{ij} \geq 0$, $\sum_{i=1}^{m} \delta_i = 1$ and $\sum_{i=1}^{m} \gamma_{ij} = 1$ for all $i, j = 1, 2, \ldots, m$. Reader may refer to Zucchini and MacDonald (2009) for the equations of mean and variance of the parameter estimates of the PHMM.

2.4 Forecasting, Decoding and State prediction

For discrete-valued observations the forecasting distribution can be obtained by using the following equation;

$$P\Big(T_{T+h} = y_T | Y^{(T)} = y^{(T)}\Big) = \phi_T \, \Gamma^h \, P(y_t) \, 1^{'} \tag{22}$$

Where; $\phi_T = \alpha_T / \alpha_T 1^{'}$ and $\alpha_T = \alpha_{T-1}\gamma P(y_t)$.

Generally decoding is of two types i.e., local decoding and global decoding. Local decoding involves determining the most likely state separately for each time t ($t = 1, 2, 3, \ldots, T$) by maximizing the following conditional probability function;

$$\begin{aligned} P\Big(C_t = i | Y^{(T)} = y^{(T)}\Big) &= \frac{\alpha_t(i) \, \beta_t(i)}{L_T} \\ &= \frac{P\big(C_t = i, Y^{(T)} = y^{(T)}\big)}{P\big(Y^{(T)} = y^{(T)}\big)}; \;\; i = 1, 2, .., m \end{aligned} \tag{23}$$

Where; $\alpha_t(i)$ and $\beta_t(i)$ are the vectors of forward and backward probability respectively, $\alpha_t(i) \; \beta_t(i) = P(C_t = i, Y^{(T)} = y^{(T)})$ and $L_T = P(Y^{(T)} = y^{(T)})$.

Global decoding involves in determining the most likely sequence of states c_1, c_2, ..., c_T which maximizes the joint probability;

$$P\Big(C^{(T)}, Y^{(T)}\Big) = \delta_{c_1} \prod_{t=2}^{T} \gamma_{c_{t-1},\, c_t} \prod_{t=1}^{T} P_{c_t}(y_t) \tag{24}$$

State prediction defined as the conditional distribution of the state C_t for t>T and the equation for state prediction is given by;

$$P\Big(C_{t+h} = i | Y^{(T)} = y^{(T)}\Big) = \frac{\alpha_T \Gamma^h e_i^{'}}{L_T} = \phi_T \Gamma^h e_i^{'} \tag{25}$$

Where, $\phi_T = \alpha_T / \alpha_T 1^{'}$, $\alpha_T = \alpha_{T-1} \Gamma P(y_t)$ and $e_i = (0, ..., 0, 1, 0, ..., 0)$ has one in i^{th} position only. Here, α_T is the vector of forward probabilities, in case of discrete state-dependent distributions, the elements of α_T being made up of products of probabilities, become progressively smaller as t increases, and are eventually rounded to zero.

2.5 Model Selection Criteria

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) values are calculated for model selection and are defined as;

$$AIC = 2 \times p - 2 \times log(L) \tag{26}$$

$$BIC = p \times log(n) - 2 \times log(L) \tag{27}$$

Where; L is the likelihood value, p is the number of parameters and n is the number of observations of the fitted model. The model with the lowest AIC and BIC values is considered as the best fit model.

2.6 Data Source

Total number of registered EC cases has been collected from Acharya Harihar Post Graduate Institute of Cancer (AHPGIC), Cuttack, Odisha, on monthly basis from January 2017 to December 2021. Hence, the data contains the frequency of EC cases for 60 consecutive months. AHPGIC is a nodal cancer hospital catering to patients from eastern India.

3 Analysis and Results

Table 1 shows the monthly series of detected EC cases from January 2017 to December 2021 along with its mean and standard deviations. The maximum number of cases (i.e., 60) was detected in March and December, 2018 and the minimum number of cases (i.e., 8) was detected in July, 2020. Figure 1, also illustrates clear fluctuations among the monthly EC cases. A Relatively lower number of EC cases was reported during the

Table 1: Monthly EC cases from January 2017 to December 2021

Months	2017	2018	2019	2020	2021	Monthly avg. cases($\bar{x}$)	S.D. (σ)	$\bar{x} \pm \sigma$
January	25	17	47	59	44	38	15	(23, 54)
February	31	48	43	47	53	44	7	(37, 52)
March	48	60	38	42	54	48	8	(40, 56)
April	28	45	34	14	48	34	12	(22, 46)
May	35	42	36	14	21	30	10	(19, 40)
June	28	38	54	28	25	35	11	(24, 45)
July	42	43	43	8	37	35	13	(21, 48)
August	24	39	40	28	44	35	8	(27, 43)
September	35	50	56	31	41	43	9	(33, 52)
October	12	43	48	24	41	34	13	(20, 47)
November	44	29	49	37	41	40	7	(33, 47)
December	48	60	52	35	30	45	11	(34, 56)
Yearly avg. cases($\bar{x}$)	33	43	45	31	40			
S.D. (σ)	11	12	7	15	10			
$\bar{x} \pm \sigma$	(22, 44)	(31, 55)	(38, 52)	(16, 45)	(30, 50)			

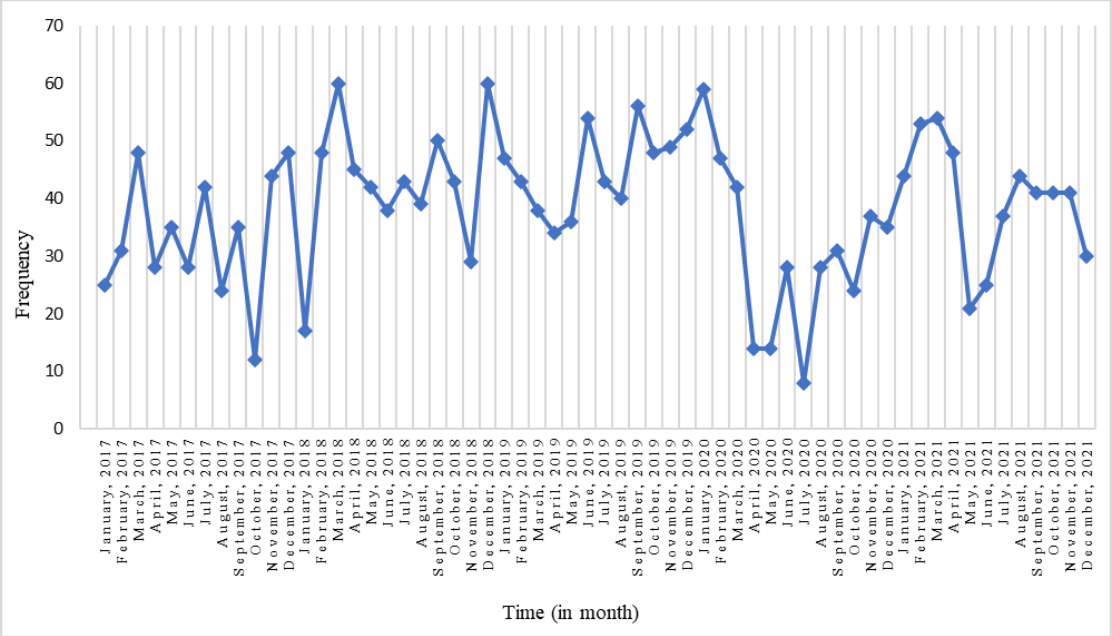

Figure 1: Time series plot of total EC cases in 60 consecutive months (January 2017 to December 2021)

period from April to July 2020. It might be due to emerging Covid-19 cases and effective shutdown and lockdown in that period.

For this data, variance = 149.209, is approximately four times larger than its mean = 38.333. This clearly shows a strong over dispersion relative to Poisson distribution. So, a single Poisson distribution may not be an appropriate choice to represent the data. Hence, we have tried Poisson independent mixture models to describe the data.

Table 2 shows the output of Poisson Independent Mixture models fitted to the monthly EC cases. From Table 2, it is noticed that there is a clear improvement in the log-likelihood values by adding an additional component up to three components in the model. Adding of fourth component shows little improvement in the log-likelihood value, rather it is insufficient to justify by looking at its AIC and BIC value. In the Table 2, the three component (m=3) Poisson independent mixture model fits well for the data as it has the lowest AIC and BIC value among all other considered models. Also, the model gives stationary distribution δ=(0.081,0.271,0.647) and the component dependent mean vector λ=(13.102,29.134,45.361).

Table 2: Outcome of Poisson Independent Mixture models (MMs) fitted to monthly EC cases

Number of components (m)	i	δ_i	λ_i	- log L	AIC	BIC	Mean	Variance
m=1	1	1.000	38.333	251.518	505.036	507.130	38.333	38.333
m=2	1	0.259	22.223	239.713	485.426	491.709	38.333	129.006
	2	0.741	43.962					
m=3	1	0.081	13.102	231.325	**472.651**	**483.122**	38.333	145.067
	2	0.271	29.134					
	3	0.647	45.361					
m=4	1	0.080	13.046	231.164	476.3286	490.989	38.333	148.460
	2	0.246	28.507					
	3	0.505	43.223					
	4	0.170	49.982					
Observations							38.333	149.209

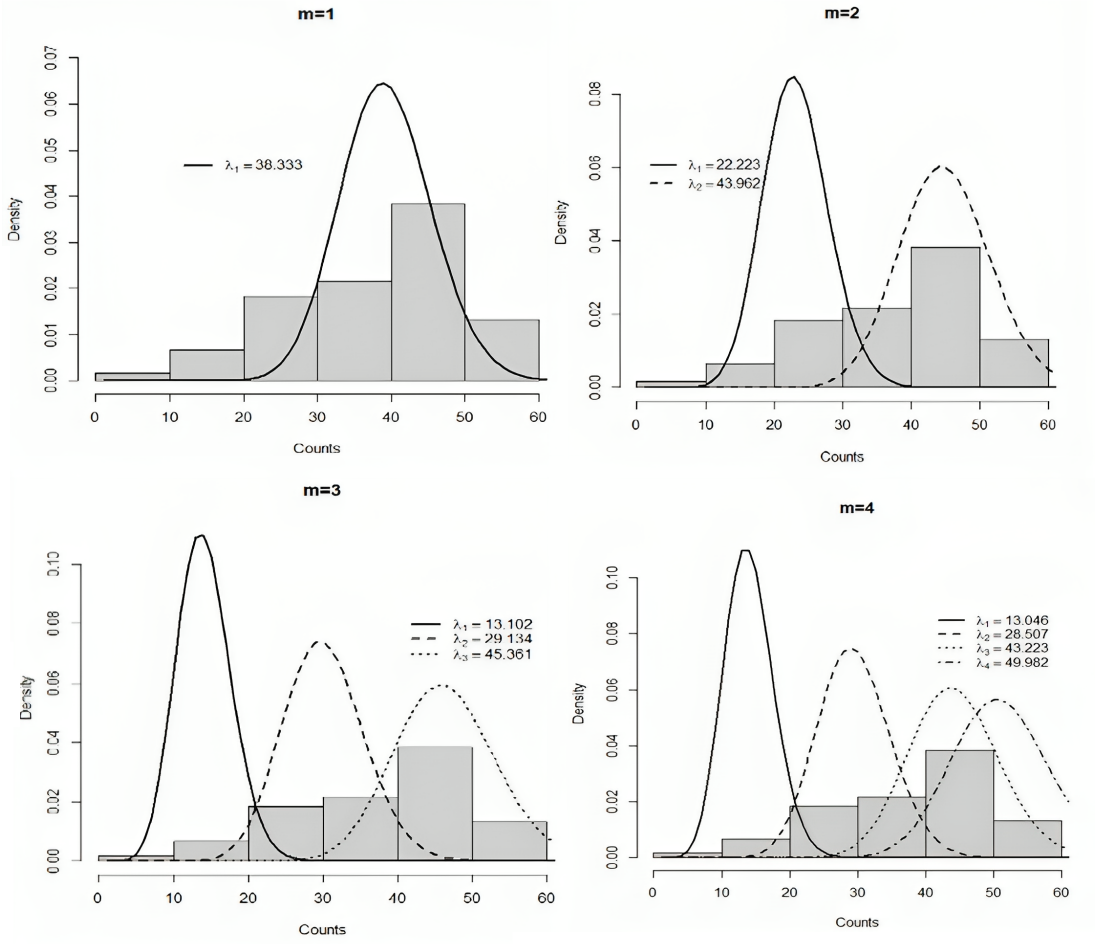

Figure 2: Histogram of EC patients compared with mixture of 1, 2, 3 and 4 Poisson distributions

Figure 2, shows a histogram of the observed number of EC cases and the four Poisson independent mixture models fitted to the data. It is clear from the figure that a Poisson mixture distribution fits the data much better than a single Poisson distribution. Visually, the three and four component model seem adequate. The better fit of these mixture models is also verified by their variances presented in Table 2.

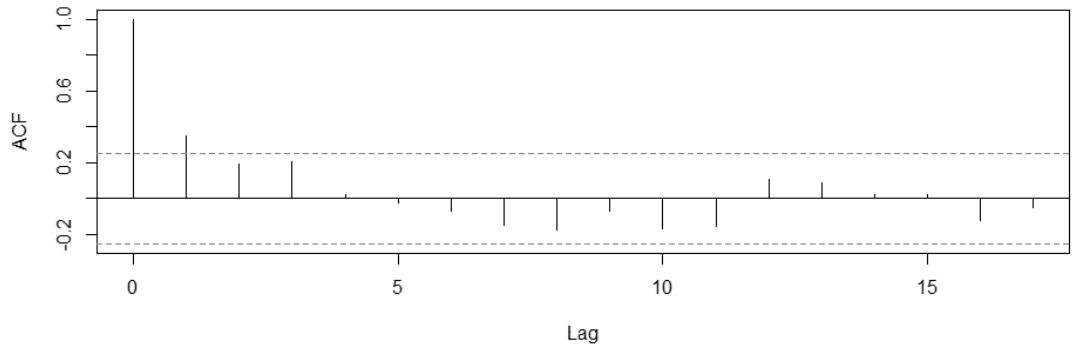

Figure 3: Sample auto-correlation function (ACF) plot for the monthly frequency of EC patients

The basic limitation of an independent mixture model is that it doesn't have serial dependance in its observations. However, in Figure 3 the sample auto correlation function (ACF) plot clearly indicates that the observations in the data set are serially dependent. Hence, one simple way to allow serial dependence in the observations is to relax the serially independent assumption of the parameter process. Thus, the resulting model of the observations becomes Markov switching model or Markov dependence model or PHMM. Therefore, next we have considered the HMMs and fitted PHMMs with one, two, three and four states to the data set.

To fit a PHMM, we first need to estimate its different parameters viz. δ,Γ and λ by using maximum likelihood estimation technique. The estimates of transition probability matrix Γ for different states were obtained by R-software and are given below in the Table 3. In Table 3 the transition probability matrix Γ for the 2 state PHMM shows that, the probability of moving from state 1 to state 2 in a single step is 0.424 and the reverse probability is 0.181. The probability that state(s) 1 and 2 remain in their original state(s) are 0.576 and 0.819 respectively. The transition probability matrix Γ for 3 and 4 state PHMM can be understood in a similar way. The corresponding estimates of δ, λ and AIC, BIC values of all fitted PHMM given in Table 4.

Table 3: Estimated transition probability matrix Γ for different PHMMs

PHMM (m=2)	PHMM (m=3)	PHMM (m=4)
$\begin{pmatrix} 0.576 & 0.424 \\ 0.181 & 0.819 \end{pmatrix}$	$\begin{pmatrix} 0.188 & 0.462 & 0.350 \\ 0.110 & 0.636 & 0.254 \\ 0.062 & 0.130 & 0.808 \end{pmatrix}$	$\begin{pmatrix} 0.162 & 0.000 & 0.283 & 0.555 \\ 0.072 & 0.499 & 0.066 & 0.363 \\ 0.106 & 0.000 & 0.894 & 0.000 \\ 0.124 & 0.876 & 0.000 & 0.000 \end{pmatrix}$

Table 4: Outcome of Poisson Hidden Markov Models (PHMMs) fitted to monthly EC cases

Number of States (m)	i	δ_i	λ_i	- log L	AIC	BIC	Mean	Variance
m=1	1	1.000	38.333	290.938	583.877	585.971	38.333	38.333
m=2	1	0.299	23.295	237.672	483.344	491.721	38.113	131.788
	2	0.701	44.435					
m=3	1	0.088	13.377	227.079	**472.159**	**491.008**	37.964	146.145
	2	0.323	30.401					
	3	0.589	45.805					
m=4	1	0.160	14.456	223.112	478.224	511.733	37.280	143.551
	2	0.106	26.968					
	3	0.280	37.662					
	4	0.454	45.967					
Observations							38.333	149.209

Table 4 shows the output of PHMMs fitted to the monthly number of EC cases. This table shows a clear improvement in the likelihood values by adding an additional state. Adding a fourth state shows little improvement on the likelihood value, but increases its AIC and BIC values which restricts us from adding an additional state in the model. Hence, PHMM with three states has the lowest AIC and BIC value(s) among all the fitted models and can be considered as the best fit model for explaining the number of EC cases. The model gives stationary distribution δ=(0.088, 0.323, 0.589) and the state dependent mean vector λ=(13.377, 30.401, 45.805).

Table 5: Estimated most likely state (local decoding) and sequence of states (global decoding) of the three state PHMM from the observed data set

Year	Decoding	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2017	Local	2	2	3	2	2	2	2	2	2	1	3	3
	Global	2	2	3	2	2	2	2	2	2	1	3	3
2018	Local	1	3	3	3	3	3	3	3	3	3	2	3
	Global	1	3	3	3	3	3	3	3	3	3	2	3
2019	Local	3	3	3	3	3	3	3	3	3	3	3	3
	Global	3	3	3	3	3	3	3	3	3	3	3	3
2020	Local	3	3	3	1	1	2	1	2	2	2	2	2
	Global	3	3	3	1	1	2	1	2	2	2	2	2
2021	Local	3	3	3	3	2	2	**2**	3	3	3	3	2
	Global	3	3	3	3	2	2	**3**	3	3	3	3	2

Table 5 shows, the prediction of most likely, state (local decoding) and sequence of states (global decoding) given the observed data series of the three state PHMM. These hidden states are obtained by Viterbi algorithm (Viterbi, 1967) and are very similar in both local and global decoding. However, they differ only in July 2021. The values of these decoding(s) help us in identifying the correct states, at the time of substantive interpretation of the states (Zucchini and MacDonald, 2009).

Table 6: Different states of 3 state PHMM associated with the number of EC cases

Range of EC cases	8 to 17	17 to 20	21 to 31	32 to 42	43 to 60
State(s) of PHMM	1	1 or 2	2	2 or 3	3

After carefully examining the Viterbi outcomes of the different states of a three-state PHMM in relation to the number of EC cases, we have reached a conclusion regarding the relationship between the range and state of the PHMM. The results are shown in Table 6. According to the Table 6, state 1 corresponds to less than 17 EC cases, state 2 corresponds to the EC cases between 21 to 31 and state 3 corresponds to more than 43 EC cases. However, if the number of EC cases is between 17 to 20 it may be labeled as either state 1 or 2. Similarly, if the number of EC cases is between 32 to 42 it may be considered as either state 2 or 3.

Table 7: Forecasted state, mean, variance and 95% CI of EC cases for the next 12 months using three state PHMM

	Jan	Feb	Mar	Apr	May	Jun
m=1	0.101	0.101	0.096	0.092	0.091	0.090
m=2	0.542	0.438	0.385	0.357	0.341	0.333
m=3	0.357	0.461	0.519	0.551	0.568	0.577
Mean	34.181	35.783	36.761	37.323	37.602	37.757
Variance	133.876	145.479	147.280	146.824	146.908	146.644
95%	(31.253,	(32.731,	(33.691,	(34.257,	(34.535,	(34.693,
CI	37.109)	38.835)	39.833)	40.389)	40.669)	40.822)
	Jul	Aug	Sep	Oct	Nov	Dec
m=1	0.089	0.089	0.089	0.089	0.089	0.088
m=2	0.328	0.326	0.324	0.323	0.323	0.323
m=3	0.583	0.585	0.587	0.588	0.588	0.589
Mean	37.867	37.897	37.928	37.944	37.944	37.976
Variance	146.265	146.309	146.352	146.373	146.373	145.862
95%	(34.807,	(34.837,	(34.867,	(34.882,	(34.882,	(34.920,
CI	40.927)	40.958)	40.989)	41.005)	41.005)	41.032)

Table 7 hows the probability that the markov chain will lie in the specified state (m=1,2 or 3) in the given months of the year 2022. The state prediction result for the next 12 months (i.e., January to December, 2022) shows that the frequency of EC cases will most likely lie in state 2 (between 21 to 31 cases) with probability 0.542 in january, in state 3 (between 43 to 60 cases) with a probability 0.461 in february and in state 3 (between 43 to 60 cases) for all other months in 2022 with probability >0.5.

The monthly forecast mean varies from 34.181 $\approx$ 34 to 37.976 $\approx$ 38 EC cases. The forecast variance ranges between 133.876 and 147.280 which is less than the observed variance from the data. The forecasted 95% CI[1] for the EC cases in January becomes (31.253, 37.109), in February becomes (32.731, 38.835) and so on. Overall, 95% CI of forecasted monthly EC cases varies between 31.235 $\approx$ 31 and 41.032 $\approx$ 41.

[1]CI: Confidence Interval

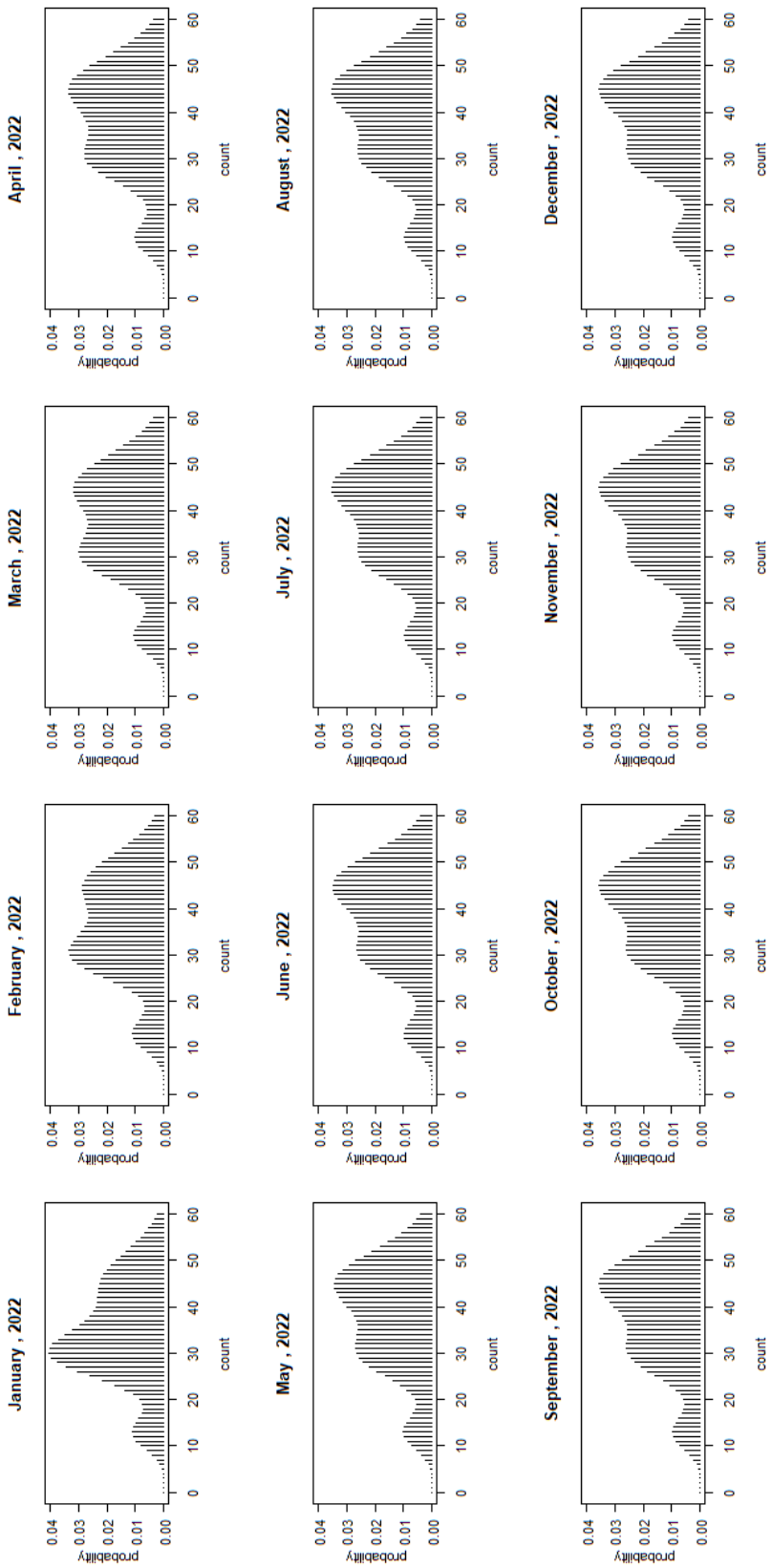

Figure 4: The mode forecast distribution curves for the next 12 months for the EC cases by three state PHMM

Figure 4 shows the mode forecast distribution curves of EC cases for the 12 months ahead. According to the forecast distribution curves it is most likely that around 30 EC cases will be reported in January. Similarly, it is most likely that between 30 to 40 EC cases will be registered in February. For the other remaining months 40 to 50 EC cases likely to be registered. These results justify our findings in the Table 7.

4 Discussion

It is very important to have an idea of the number of expected patients in future time to make a better plan for the diagnosis and course of treatment of their critical condition and make required infrastructural changes. In the literature review, we found that a limited number of studies have been conducted on the applications of HMM to forecast their states and subsequently the number of count cases. To the best of our knowledge, this is the first type of study which forecasts the number of EC cases for the state Odisha. Therefore, this study helps to identify the hidden states of the data that follow PHMM and forecast the number of EC cases. Further, the PHMM captures different time series components through its modelling of hidden states, transition probabilities, and emission probabilities.

In this study, we have taken the number of EC cases consecutively 60 for months from 2017 to 2021 to forecast the number of EC cases for the next 12 months. The time series data set under consideration was found to be an unbounded and over dispersed sequence. Usually, in such cases, a mixture modelling approach is found to be a good alternative to ARIMA (Zucchini and MacDonald, 2009). We have fitted the independent mixture of Poisson model for different components as it takes unobserved heterogeneity into account. Based on the lowest AIC and BIC value, the three component Poisson independent mixture model fits well for the data.It has been seen that the state specific PHMM was able to forecast the number of death counts due to dengue fever (George and Thomas, 2018), monthly cases of Hepatitis-B (Jose et al., 2020), monthly congenital hypothyroidism cases by Sadeghifar et al. (2016).

The ACF plot showed the number of EC cases were serially dependent. Hence, we have considered the Markov dependent mixture model and fitted PHMM with different states to the number of EC cases. Based on the lowest AIC and BIC value, a three state PHMM was found to be the best fitted model to explain the distribution of the number of EC cases. In a similar study conducted by Sadeghifar et al. (2016), congenital hypothyroidism cases followed a two state PHMM, and the model further used for forecasting the number of cases for next 24 months. Azizah et al. (2019), used a two-state PHMM using the Bayesian approach to study the Seismicity Activity Rates of earthquakes.

After being assured of the validity of the usage of PHMM in our case, we then used PHMM to forecast the number of EC cases for next 12 months and the EC cases are most likely to lie in state 2 for January and in state 3 for rest of the months. The monthly forecast mean of number of EC cases varies between 34 to 38 for the year 2022. However, for the first four forecasted months viz. January, February, March and April,

the EC cases increase approximately by a single unit viz. 34, 35, 36 and 37 respectively. For the next forecasted months, the number of EC cases are nearly invariant. The Forecast distribution curves show the increasing pattern of EC cases in future months. The interval forecast states that the monthly EC cases vary between 31 to 41.

Previous studies stated that using HMM to forecast stock price is explainable and has solid statistical foundation. Some hybrid systems have been developed using AI paradigms with HMM to further improve accuracy and efficiency of the forecasts (Hassan and Nath, 2005). Some studies also applied HMMs for other purposes than forecasting. In a study conducted by Paroli et al. (2000), PHMMs were used in non-life insurance. Results suggested that PHMMs are a more general approach compared to the Poisson distribution and Poisson process to model the claim number in non-life insurances.

As far as the limitation is concerned, in this study we have focused only on the frequency of registered EC cases per month without considering any other effective factor. In addition to this, for this study we have taken only 60 data points from a single cancer hospital. Hence, this study can be generalized more accurately by taking other effective factors into account and taking a comparatively larger sample.

5 Conclusion

This study reveals that the average EC cases will increase in future months. Based on this analysis, it is also suggested that, the three-state HMM can be used to fit the distribution of the EC cases. Apart from potentially giving better fits to data, mechanistic models can provide insight into underlying processes and allow investigators to make inferences about key parameters. From the statistical application point of view, we conclude that the PHMMs are more realistic in practice than the classical Poisson processes, as they consider the unobservable underlying environmental effect, which affects the counts.

To our knowledge, this represents the first application of Poisson hidden Markov models to forecast the number of endometrial cancer cases. These estimates can be reliable if women's death rate and EC specific death rate are constant for the year 2022. The outcomes of this study will contribute towards providing better treatment care and patient management, and in turn reduce the mortality of EC patients. This work may be extended by considering a PHMM setup with different covariates associated with EC patents.

Acknowledgement

We would like to thank Dr. Smruti Sudha Pattnaik, Department of Gynaecology Oncology, AHPGIC, Cuttack, Odisha for providing the data for this study and Khushi Agrawal for overall review and language correction of the manuscript. Further, we would also like to express our sincere gratitude to the reviewers of EJASA journal for their thoughtful and insightful comments on our manuscript. Their suggestions have significantly improved the clarity and rigor of our work.

References

Azizah, N., Astutik, S., et al. (2019). Two-state poisson hidden markov models for analysis of seismicity activity rates in west nusa tenggara. In *IOP Conference Series: Materials Science and Engineering*, volume 546, page 052015. IOP Publishing.

Churchill, G. A. (1989). Stochastic models for heterogeneous dna sequences. *Bulletin of mathematical biology*, 51(1):79–94.

Cohen, A. (1998). Hidden markov models in biomedical signal processing. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286)*, volume 3, pages 1145–1150. IEEE.

Cooper, B. and Lipsitch, M. (2004). The analysis of hospital infection data using hidden markov models. *Biostatistics*, 5(2):223–237.

George, J. and Thomas, S. (2018). Modeling of dengue fever death counts using hidden markov model. *Environmental Statistics*, 8(9):1–11.

Green, P. J. and Richardson, S. (2012). Hidden markov models and disease mapping. *Journal of the American statistical association*, 97(460):1055–1070.

Hassan, M. R. and Nath, B. (2005). Stock market forecasting using hidden markov model: a new approach. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pages 192–196. IEEE.

Jose, A., Sebastian, T., and S.George (2020). Modelling time series of count data using hidden markov models. *Stochastic Modeling and Applications*, 24(1):25–33.

Juang, B. H. and Rabiner, L. R. (1991). Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272.

Laverty, W. H., Miket, M. J., and Kelly, I. W. (2002). Application of hidden markov models on residuals: an example using canadian traffic accident data. *Perceptual and motor skills*, 94(3_suppl):1151–1156.

Paroli, R., Redaelli, G., and Spezia, L. (1994). Poisson hidden markov models for time series of overdispersed insurance counts. pages 461–474.

Paroli, R., Spezia, L., Redaelli, G. L. M., et al. (2000). Hidden markov models for time series of overdispersed insurances counts. In *Proceedings of the XXXI International ASTIN Colloquium*, pages 461–474. Proceedings of the XXXI International ASTIN Colloquium.

Pattnaik, S. S., Mohapatra, J. J., Parija, J., Sarangi, L., Das, R., Rout, N., Giri, S. K., Samantray, S., Panda, N., Devi, P., Das, P. K., Nayak, B. L., Mohapatra, M. R., Padhi, A. K., and Panda, S. (2020). A case series of poorly differentiated carcinoma endometrium ihc confirmed reported at ahrcc cuttack, state odisha , country india. *International Journal of Scientific and Research Publications*, 10(3):189–201.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Sadeghifar, M., Seyed-Tabib, M., Haji-Maghsoudi, S., Noemani, K., and Aalipur-

Byrgany, F. (2016). The application of poisson hidden markov model to forecasting new cases of congenital hypothyroidism in khuzestan province. *Journal of Biostatistics and Epidemiology*, 2(1):14–19.

Sarvi, F., Nadali, A., Khodadost, M., Moghaddam, M. K., and Sadeghifar, M. (2017). Application of poisson hidden markov model to predict number of pm2.5 exceedance days in tehran during 2016-2017. *Avicenna Journal of Environmental Health Engineering*, 4(1):58031–58031.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.

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