EasyChair Preprint

№ 2106

A Machine Learning Approach to Biodiversity Time Series Analysis

Rajarshi Paul and Th. Shanta Kumar

December 8, 2019

2nd International Conference on information systems & management science (ISMS) 2019

A Machine Learning Approach to Biodiversity Time Series Analysis

Rajarshi Paul[a], Th. Shanta Kumar[b]

[a] *Research Scholar, GIMT, Guwahati-781017, India*
[b] *Associate Professor, Department of CSE, GIMT, Guwahati-781017, India*

Abstract: : In this paper we have accessed the ecological changes of resources through time. A brief concept of time series and a case study of the observation of dragon flies in Kerala region have been studied. A machine learning approach of processing the time series data, some forecasting results like migration location, population growth, future presence on a particular dragonfly species and the prediction approach is highlighted in this paper.

1. Introduction

People are very much interested in the questions like what will happen with our metrics in the next day, month, year or decade. Suppose what will be the sale of Hyundai cars in the next quarter of the year, how much bird migration will be from Siberia next year, how many people will suffer from dengue or encephalitis next year etc. Underlying principle of time series prediction lies in the estimation of unknown value of unknown variable. Data recorded periodically at certain interval of time is known as time series. If we technically talk about Time series we have time(t) as an independent variable and a (y_t) as dependent variable. The output of a time series model is the prediction of the value of y at a time t($\hat{y}_t$). Prediction is just a specific value such as rainfall of the month of June, 2021, whether a party will come to power or not in the next election, kind of insects that can be found inside deep Amazon forest etc. Forecast is a prediction which has some statistical measure with a confidence level of certainty which can be defined in percentage.

Nomenclature

Biodiversity
Biodiversity time series
Dragonfly distribution
Dragonfly migration
Discrete transitive auto-regressive model
Discrete functional variability time series
Irregular time series
LSTM encoder-decoder-predictor
Open data
Entomology
Conservation
Public participation in scientific research, Kerala, India

2. Biodiversity and Time Series

A key scientific challenge is to quantify and forecast temporal change in biodiversity attributable to both natural and anthropogenic causes[2]. Forecasting biodiversity change is essential for developing successful policies to mitigate biodiversity loss and for addressing basic ecological issues, such as the relationship between diversity and ecosystem function, the linkage between diversity and stability and the detection of ecological tipping points relation to the existence of alternative stable states[2]. Because most biodiversity studies are observational rather than experimental—particularly at large scales, we argue that temporal relationships between biodiversity, ecosystem services and hypothesized driver variables are among the strongest possible evidence for causal links. Moreover, temporal studies of biodiversity are essential for forecasting future change in community structure and ecosystem function[2].

3. Scientific approach

Time series bio diversity data may have the following patterns:-
Trend: Trends can be linear or non-linear component. Trends exhibits either increasing behaviour or decreasing behaviour with respect to time
Seasonal: Seasonal pattern is either linear or non-linear pattern that repeats at particular intervals of time.
Cyclic: Cyclic pattern persists for a longer period of time and have a wave form.
Random/Noise: A component or a phenomenon that does not follow any specific pattern is a noise.

3.1. Stationarity

In time series analysis stationarity means constant statistical properties like mean, variances, co-variance et. over a period of time. We can say that the measurement of y of those values remains same at a time p(y_t) and also at other point in time. To perform time series analysis the dataset should be stationary to get a good analytical result.

3.2. Auto-Regressive model

Auto regressive or AR model suggests that the value of the variable y in y-axis at time t depends on the value of y at time t-1 i.e. it depends on the previous value. If y depends on more than one previous values then it is denoted by p parameters[3]

$$AR(p): y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, \ldots\ldots y_{t-p}) \quad (1)$$

where p is the number of past values.

We can write the above expression as[3],

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \ldots\ldots + \beta_p y_{t-p} + \varepsilon_t \quad (2)$$

Again there can be a variability in the functional dependency between past r variables in bio diversity time series data. We can mention it as discrete variability of transitive dependency or discrete transitive auto-regressive model, denoted by $AR_{Transitive}(p)$,which can be expressed as,

$$AR_{Transitive}(p) = f(y_{t-1}, y_{t-2} = f(y_{t-3}, y_{t-4}), y_{t-3} = f(y_{t-4}, y_{t-5}, \ldots y_{t-r}), \ldots y_{t-p}) \quad (3)$$

Again if the time series is an irregular time series i.e. there is no fixed interval between observations and assuming that it is a collection of some continuous time series with variable set of data and also the previous time series affects the next one in different ways(i.e. with different functionalities) then we can mention it as discrete functional variability time series and can express as,

$$AR_{discrete}(p) = f(y_{t-1}, y_{t-2} = f_1(y_{t-3}, y_{t-4}), y_{t-3} = f_2(y_{t-4}, \ldots y_{t-k}), \ldots y_{t-p}) \quad (4)$$

where p is the number of past values to consider.

3.3. Moving Average Model

In case of moving average model considers white noise error terms. It can be expressed as[3].

$$y_t = \beta_0 + \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \ldots\ldots \varphi_q \varepsilon_{t-q} \quad (5)$$

Where, ε_t, ε_{t-1}, ε_{t-2} etc. are white noise error terms.

3.4. ARMA

Auto regressive moving average model is a combination of the autoregressive model and the moving average model which uses both the past and the white errors and predicts the future time series. Mathematical expression of ARMA is as follows[3],

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \ldots + \varphi_r \varepsilon_{t-q} \quad (6)$$

3.5. ARIMA

Autoregressive Integrated Moving Average model is most popularly used in Time series forecasting. In this model we try to achieve stationarity. It is a generalization of ARMA. The ARIMA model considers the following 3 parameters,

p: No. of previous orders that should be allowed

d: differentiation degree to be considered

q: Moving average order

4. A case study

Dataset: The dataset which is considered here is the time series observation of dragonflies from in and around Thumboor, observed by Rison Thumboor. This dataset is uploaded to Indian Biodiversity Portal. Most of the knowledge produced on biodiversity remains fragmented and is not available in any standardized structure impeding for instance, our ability to perform robust analyses on species distribution patterns[1]. However, generating accurate data on biodiversity over large spatial and temporal scales is an extremely difficult challenge to meet without a multi-institution and/or multi-actor collaborative schemes[1]. Dataset contains scientific name of the dragonfly species whose photo is captured, it's sex, date of capture, capture location in latitude and longitude. It is an irregular time series data. Since the dataset contains the name of the dragonfly species, it's captured photo, date of the photos in a discrete interval of time, it's location etc. One analysis that can be possible is the count of dragonfly monthly or daily or yearly basis. Again the same can be done for a particular dragonfly species also and put those data to design an ARMA model. Location data like latitude and longitude gives a precise presence of the species and its aggregation from the dataset gives us a population idea. Although this aggregation data can be taken as a sample to predict the actual population. From the photographer's enthusiasm and the quantity of data recorded we can have an idea of the presence of a particular species at different locations. Data recorded by an enthusiastic photographer continuously engaged in the job without lapse can provide us a good platform for prediction of species migration.

A particular dragonfly species record needs to be aggregated first for processing based on the number of observations monthly or yearly. But aggregating records monthly or even yearly results in irregularly spaced observations in time as shown in the following two tables for two dragonfly species "Acisoma panorpoides" and "Brachythemis contaminata". **Table 1** lists the actual time series data.

4.1. Tables

Table 1- Dragonfly species and it's date of observation

Dragonfly species	Observation date
Acisoma panorpoides	2014-08-16
	2014-08-20
	2015-07-30
	2015-08-09
	2017-11-21
	2018-01-23
	2018-10-25
	2016-10-26
	2017-09-10
	2017-11-12
	2017-12-04
	2017-12-08
	2017-12-19
	2017-12-24
	2018-10-07
Brachythemis contaminata	2012-01-25
	2014-08-16
	2015-08-09
	2015-08-13
	2018-01-02
	2015-10-24
	2015-11-06
	2015-11-22
	2016-03-01
	2016-12-18
	2017-04-02
	2017-04-09
	2017-10-23
	2017-11-12
	2017-12-08
	2017-12-18
	2018-09-15
	2018-09-16
	2018-10-07
	2018-10-27

Table 2-Aggregation of dragonfly observations year and monthwise

Dragonfly species	Year	Month	Count
Acisoma panorpoides	2014	8	2
	2015	7	1
		8	1
	2016	10	1
	2017	9	1
		11	2
		12	4
	2018	1	1
		10	2
Brachythemis contaminata	2012	1	1
	2014	8	1
	2015	8	2
		10	1
		11	2
	2016	3	1
		12	1
	2017	4	2
		10	1
		11	1
		12	2
	2018	1	1
		9	2
		10	2

Table 3-Aggregation of dragonfly observations year wise

Dragonfly species	Year	Count
Acisoma panorpoides	2014	2
	2015	2
	2016	1
	2017	7
	2018	3
Brachythemis contaminata	2012	1
	2014	1
	2015	5
	2016	2
	2017	6
	2018	5

For the above time series data we can determine dragonfly population change or density change with a simple differentiation term, but we must make it regular by any method such as interpolation.

$$\left[\frac{\partial y_t}{\partial t}\right]_{species=const.\&loc=const|} \neq const. \quad (7)$$

where, ∂y_t = change of total aggregated population count in an irregular time series, ∂t = amount of change in time. Note that species and location are kept constant and y_t is the species presence or density along temporal axis.

Now, let $[\partial y_t/\partial L]_{t=constant}$ be the population of a species w.r.t. location, keeping time constant. Let us now obtain this value from different time serieses generated for different locations for a particular species, and $[\partial y_t/\partial t]_{loc=constant}$ be the change of population w.r.t. time keeping location constant from any t-series. The migration between two locations L_1 and L_2 can be explained with an differentiation term after subtracting the mortality factor $[(D_{L1}/P_{L1}\times 10^n + D_{L2}/P_{L2}\times 10^n)\,\partial t]$,

$$\left(\left[\frac{\partial y_t}{\partial t}\right]_{location=L_1|} - \left[\frac{\partial y_t}{\partial t}\right]_{location=L2}\right) - \left(\frac{D_{L1}}{P_{L1}}\times 10^n + \frac{D_{L2}}{P_{L2}}\times 10^n\right)\partial t \approx const \quad (8)$$

Where, D_x is the death occurring within a given time period at location 'x', P_x is the size of population or density within which the death occur at location x, 10^n is a conversion factor such as 10^5 means mortality rate per 100000 dragonflies at a location 'x'.

4.2. Observations

The dataset considered in this paper is a sparsely sampled dataset. We can observe randomness in the series if we consider the yearly count. There is also year gaps as observed in the time series i.e. it is sparsely sampled and irregular time series. For our study we can consider a generalizations of stochastic process. Let the probability space considered as $(\omega, \mathcal{A}, P)$ where $\omega\in\Omega$ the sample space of possible outcomes of an random experiment, $\mathcal{A}$ is σ-space or subsets of Ω, P is the probability function or probability measure on $(\omega, \mathcal{A})$. An integer valued stochastic process is a family of random variables $\{X_\gamma,\gamma\in\Gamma\}$ defined on $\Omega \times \Gamma$ taking values in $\mathbb{Z}^+$ i.e. set of positive integers. Thus the random variables of the family (measurable for every $\gamma\in\Gamma$) are functions of the form,

$$X(\gamma,\omega): \Gamma \times \Omega \mapsto \mathbb{Z}^+.$$

For $\Gamma\subset\mathbb{Z}^+$, we have a discrete-time process and for $\Gamma\subset\mathbb{R}$ we have continuous time process. Our bio diversity dataset consists an irregular time series after month wise or year wise aggregation of observations and we can consider it as a discrete time process. We can denote discrete time stochastic process as $X=\{X_t,t\in T\}$, where T is <month, year> touple or <year>.We also denote X_t as X(t). For a fixed value of ω, say $\tilde{\omega}$ we have $\{X(t,\tilde{\omega}), t\in T\}$ where $\tilde{\omega}$ can be the event of population count below a threshold level.

To predict population migration as well as future population count the data set have to be smoothened first and need to do some preprocessing operations. We can use an interpolation technique to smoothen the curve. We can take date, month-year or year as independent variable. We can estimate/guess observations for any unknown intermediate {date}, {month-year} or {year}. Thus we can make the discrete time series to a continuous time series. But interpolation is considered as an numerical technique to guess the value of any unknown independent variable. Estimating the future population or population out of scope of the timeline is of course an extrapolation problem.

5. Machine Learning approach to handle bio-diversity t-series

Recurrent Neural Network is a kind of neural network which can predict the next item of observations depending upon the previous items observed in the sequence. In RNNs the hidden layers acts as an internal storage for storing earlier stages of observed sequence .But the drawback of the generic RNNs is that these networks remember only a few earlier steps in the sequence and thus are not suitable for remembering longer sequences of data. This problem is addressed in LSTM(Long Short Term Memory). "Memory line" is introduced in LSTM network. Earlier sequence or trend or pattern is memorized using some gates along with a memory line designed with a typical LSTM. LSTMs are special kind of RNNs with additional features to memorize the sequence of data. So, it is one of the suitable approaches for analyzing long time-series data by deep learning approach. The best idea is to handle a bio-diversity series with randomness associated with the observations. Pattern of growth of insect species may repeat over a season or month or even over a decade and may observe the same random pattern in future with little bit of variation as the climate remains little bit same for long duration of time and seasonality persists. Although the variation of growth rate as well as the mortality rate of insets are very random but their growth rate exhibits almost the same pattern seasonally, if no other external factors say like natural calamity affects the growth rate. If randomness is associated with a certain process and there is also some noise associated and impacting the process then we can use a machine learning approach like LSTM autoencoder-decoder to get insight detail of the dragonfly time series. One interesting aspect of time series of insects is to forecast future population sequence on the temporal axis.

5.1. Comparison between ARIMA and LSTM in timeseries analysis

ARIMA model can be chosen if the timeseries data is non-stationary. ARIMA is a type of model to capture temporal patterns in a time series data. As we know that ARIMA is composed of three key aspects AR(auto regressive)- dependencies between an observation and the old observations, I (integrated)- making the time series stationary by measuring the differences of observations at different time, and MA(moving average)- dependency between observations and the residual error terms when MA model is used to the old observations:

A simple AR model of order p, AR(p) is,

$$y_t = c + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t \qquad (9)$$

Where, y_t is the stationary variable, c is a constant, p is the order of the model and each β_i are autocorrelation co-efficients at lags of 1,2,3,....p. The residuals are Gaussian white noise series with mean zero and variance σ_ε^2. Similarly an Moving average model , MA of order q, i.e. MA(q) is,

$$y_t = \mu + \sum_{i=0}^{q} \phi_i \varepsilon_{t-i} \qquad (10)$$

where μ is $E(y_t)$ i.e. expectation of y_t and assumed to be zero, ϕ_i are the weights applied to the current and previous values of a random variable or stochastic term and $\phi_0=1$. And ε_t is assumed to be Gaussian white noise series with mean zero and variance σ_ε^2. We combine equation (9) and (10) model to form the ARIMA model of order (p,q) as:

$$y_t = c + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t + \sum_{i=0}^{q} \phi_i \varepsilon_{t-i} \qquad (11)$$

Where, $\beta_i \neq 0$, $\phi_i \neq 0$, and $\sigma_\varepsilon^2 > 0$. ARIMA is capable of dealing with non-stationary time series data because of its “integrate” step. The “integrate” component involves differencing the time series to convert a non-stationary time series into a stationary. The general form of seasonal ARIMA model is ARIMA(p, d, q)×(P,D,Q)S, where p is the non-seasonal AR order, d is the non-seasonal differencing, q is the non-seasonal MA order, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA order, and S is the time span of repeating seasonal pattern, respectively.

There is an alternative approach of using deep learning algorithms for processing time series data. LSTM is one among them where we can preserve the features of training data for a longer duration of time. Both the algorithms are used for time series forecasting and both are based on single forecasting technique i.e. to predict the next data point for each dataset. The approach is based on training sets containing one more observation then the previous one, i.e. look ahead view of the data. There are some variations of this approach:

a) Estimating single set of training data and then computing one step forecasting on the remaining set s.

b) One step forecasting for the next multiple steps.

c) Refitting the model at each iteration before each forecasting.

Our study in this paper is to use LSTM network for irregular times series analysis and forecasting.

5.2. Autoencoders

An auto encoder is a neural network model that seeks to learn a compressed representation of an input. Auto encoders are basically unsupervised machine learning approach as no labeling of input data is done by human beings but we can mention it as self supervised learning approach. Training is a part of encoder to recreate the input. In auto encoder the input and output data are same. In general the first part of an autoencoder is the encoder "**f**" to encode the sequence '**x**' of H-dimensional space to L-dimensional space, where L<H, i.e. e(x)=u. The second part is the decoder which attempts to reconstruct the original input sequence of dimension H from the lower dimensional representation L, i.e. d(u)=d(e(x))=r(x)≈x. Where r(x) is the reconstruction phase.The learning phase of the autoencoder is to how to compress the data to a lower dimensional space. Objective of this kind of model is to reconstruct the output with minimum amount of information loss. Once it is trained we can compress the test data by the encoder part. The vector x is replaced with y_t in the **fig 1**.

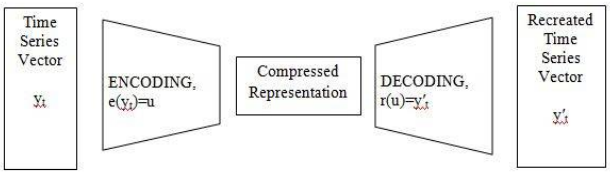

Fig. 1 - An general auto encoder model for an irregular time series.

Training an encoder with missing values of an irregular time series:-

The missing values of an time series is not given to an encoder it is trained to reconstruct the uncorrupted parts of the time series only. The corrupted or assumed or guessed data can be fed to the network, however, the error only for the reconstruction of uncorrupted data points is evaluated and used in the backpropagation. This processes works on the assumption that missing points have different series distributions. If a node value is not used in the backpropagation in the recent iteration then the next set will have missing values in different position. Thus weights of the network are not updated in the current iteration phase but in the next phase. Binary bias is added to monitor the quality of reconstruction.

5.3. LSTM Autoencoder on a time series

The dragonfly observations are not continuous. After aggregation of the observations monthly or yearly as shown in **table 2** and **table 3** we can generate a discrete irregular time series of observations y_t along temporal direction. We can consider irregular time series having random nature as a sequence. Bio diversity sequence prediction is a challenging problem because machine learning algorithms and neural networks work with fixed length of input. But in real scenario a sequence can vary in length. So, the two challenging problem with sequences are:

a) Sequence data can vary in length.
b) Temporal ordering of the observations can make it difficult to extract the features suitable to use for supervised approach.

Although we are not adopting the methodology to train an encoder phase with missing values but in this paper we will first interpolate the irregular time series with nearest neighbor interpolation technique(data smoothing) to fill the gaps and then will use an LSTM autoencoder for time series generation. We can even skip the interpolation step and directly feed the sequence. We are not to going to predict the null values, missing values or noises in the sequence here neither going to demonstrate that we can use this model to adopt variable length bio diversity time series here. Although the bio diversity time series data are ordered in time, our objective is not just to compress and regenerate the sequence but to see how the model is predicting or regenerating the time series data:

A flow diagram of time series generation is mentioned in **fig 2**:

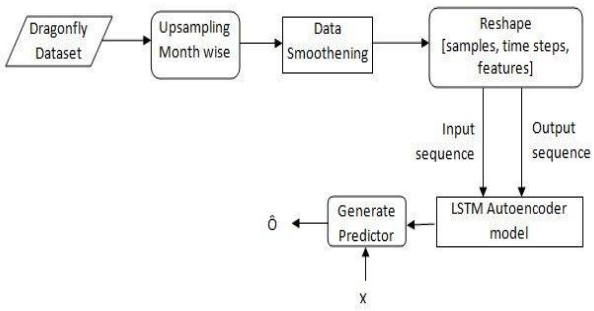

Fig. 2 - Time series generation process

The LSTM auto encoder model is shown in **fig 3** below,

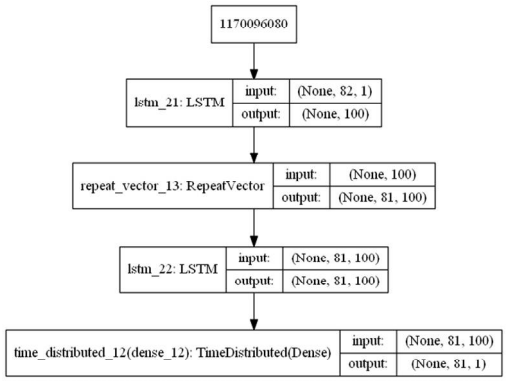

Fig. 3 - LSTM autoencoder model for timeseries generation

6. Results and Experiments

The discrete time series data carrying aggregated observations of a dragonfly species "Brachythemis contaminata" shown in **fig 4**.

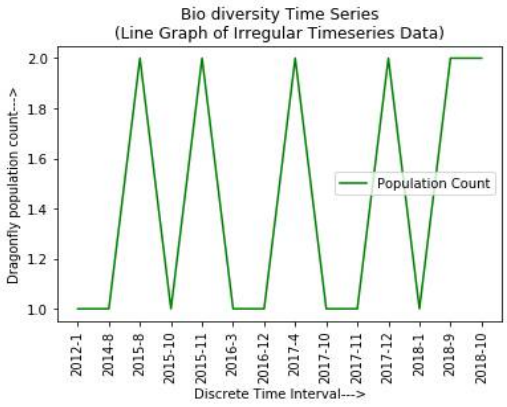

Fig. 4 - Line graph of dragonfly observations.

Polynomial interpolation from 2012 to 2018 gives us negative values. Which is an invalid number for population count. Therefore this type of interpolation methods like spline, polynomial, cubic etc. are not suitable for us. The time series graph looks similar to like **fig 5** below. The red line is the mean axis.

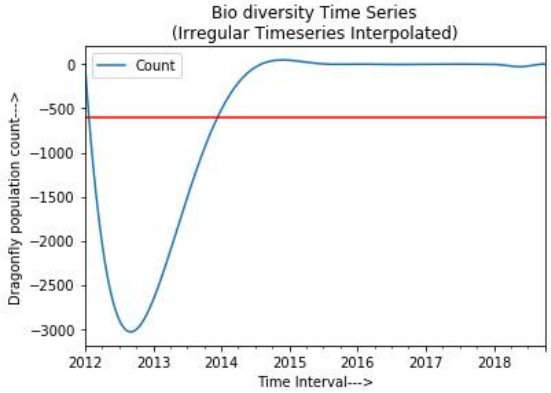

Fig. 5 - Polynomial interpolation of dragonfly observations.

If we go for nearest neighbor interpolation method we can fill up the timeline(x-axis) of the timeseries carrying integer value along the y-axis.The following figure, **fig 6** depicts the point,

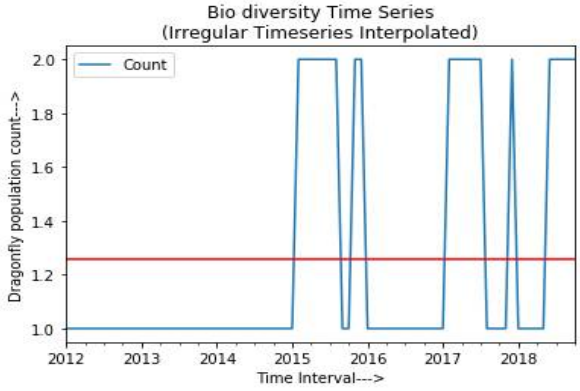

Fig. 6 -Nearest neighbor interpolation of dragonfly observations

Acknowledgements

We acknowledge the Indian bio diversity web portal for providing the dataset on dragon fly distribution in and around Thumboor for our time series analysis work on species biodiversity. From the photographer's enthusiasm and the quantity of data recorded we can have an idea of the presence of a particular species at different locations.

References

Vattakaven, T., Hanraads, G. M. Rohit., Balasubramanian. D, Rejou-Mechain.M, Muthusankar G, Ramesh. R. B,& Prabhakar, R. (2016). *Biodiversity Portal: An integrated, interactive and participatory biodiversity informatics platform.,(e10279)*, p4, Biodivers Data J.

Magurran, E., Anne., Buckland, T. S., Chao, A., Chazdon L. R, Colwell K. R, Curtis, T., Gaston J. K, Gotelli J. N, Kosnik A., M., McGill, B., McCune L. J., Morlon H., Mumby J. P., Ovreas L. (2012). *Quantifying temporal change in biodiversity: challenges and opportunities* (3rd ed.) (280: 20121931). New York: MacMillan.

Brockwell, J. P., Davis, A. R. (2002). *Introduction to Time Series and Forecasting* (pp. 281–304).(ISBN 0-387-95351-5).

Huaiyu, W., Shengnan, G.,Kang,Y.,Xiaohui L.,Youfang L., (2019). *Knowledge Based System:CTS-LSTM:LSTM based neural networks for correlated time series prediction* (pp. 105239), https://doi.org/10.1016/j.knosys.2019.105239.

Siami N.,S.,Tavakoli N.,Namin,S.,A., (2018). *A Comparison of ARIMA and LSTM in Forecasting Time Series* (ISBN: 978-1-5386-6805-4),18397816, DOI:10.1109/ICMLA.2018.00227.

Menke,W.,Menke. J.(2016). *Environmental Data Analysis with Matlab, Filling the missing data* (pp. 223-237) (https://doi.org/10.1016/B978-0-12-804488-9.00010-0), Elsevier Inc.

Menke,W.,Menke. J.(2017). *Geographical Models with Mathematica, Statistical and Probability Models for Given Relationships Between Societies and the Natural Environment* (pp. 19-72) (https://doi.org/10.1016/B978-1-78548-225-0.50003-8), Elsevier Inc.

Magurran,A., E., Baillie Stephen,R., Buckland S. T., Dick,M.,J., Elston D., A., Scott E., Marian, Smith, R., I., Somerfield Paul,J.,Watt, A.,D. (2010). *Long-term datasets in bio diversity research and monitoring: accessing change in ecological communities through time* (pp.574-582),Vol 25,Issue 10. Treads in Ecology & Evolution, Elsevier B. V.

Banks,H.,T.,Banks. E.,J.,Joyner L.S., Stark, D.J.(2008).*Mathematical and Computer Modelling. Dynamic models for insect morality due to exposure to insecticides*(pp. 316-332),Vol 48, Issue 1-2, (https://doi.org/10.1016/j.mcm.2007.10.005), Elsevier Inc.

Lacus, Stefano. M.(2008). *Simulation and Inference for Stochastic Differential Equations With R Examples* (1st ed.) (ISBN 978-0-387-75838-1), Springer-Verlag New York.

Saha,A.,McRAE,L.,Dodd,K.,C., Gasden, H., Hare M., K., Lukoschek,V., Bohm,M.(2018). *Tracking Global Population Trends: Population Time Series Data and a Living Planet Index for Reptiles*, Vol. 52, No 3, (259–268), Journal of Herpetology, 52(3), 259–268.