International Journal of Engineering & Technology, 7 (2.32) (2018) 240-244

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Reservoir Computing for Healthcare Analytics

Shantanu S Pathak[1], D Rajeswara Rao[2]

[1,2]Koneru Laxmanai Education Foundation (K L University) Guntur, AP, India
**Corresponding author E-mail: shantanuspathak@gmail.com*

Abstract

In this data age tools for sophisticated generation and handling of data are at epitome of usage. Data varying in both space and time poses a breed of challenges. Challenges they possess for forecasting can be well handled by Reservoir computing based neural networks. Challenges like class imbalance, missing values, locality effect are discussed here. Additionally, popular statistical techniques for forecasting such data are discussed. Results show how Reservoir Computing based technique outper-forms traditional neural networks.

***Keywords**: Reservoir Computing, Spatio-Temporal Data, Echo state networks, Data Analytics*

1. Introduction

In this data age tools for sophisticated generation and handling of data are at epitome of usage. Soft computing, machine learning, and cognitive sciences are developing towards analysis of such data. Also other systems like nano-technology, IoT, Embedded Systems are producing enormous data. Out of varieties of data, data varying with both time and space is special. It has typical characteristics like stability, seasonality, locality [1]. It needs special handling. In this work attempt is made to explore a spatio-temporal atmospheric data. Unique challenges, statistical test, and solutions are presented here. This work contributes by specifying steps towards such analysis, various conclusions drawn and potential applications of various techniques on such data. Role of Reservoir computing in such data handling is highlighted here. In following section literature survey is presented, then dataset description, and observations are put forth. Subsequently, challenges offered by this dataset and potential solutions are presented.

2. Literature Survey

In this section various state of the art techniques used for handling Time-Series data are discussed. Here, Statistical as well as machine learning perspectives are given.

Auto Regressive Moving Average (ARMA)- This model has two parts, one handling lag based instances of time series and other one handling error terms. Work [2] discusses parameter values of this model in details, while [3] works on automatic selection of these parameters. On other hand long term predictions are achieved in [4]. Degree of lag is specified by p term and degree of averaging of error is specified by q.

$$\mathrm{X}_t = c + e_t + \sum_{i=1}^{p} \sigma_i \mathrm{X}_{t-i} + \sum_{i=1}^{q} \Theta_i \varepsilon_{t-i}$$

Acceptable values of pand q can be taken by iterative process. Initial guess can be based on partial autocorrelation functions.

Auto Regressive Integrated Moving Average (ARIMA)- This is generalization of ARMA with capability to transform series in stationary format. This capability is introduced by integration module. Other parts are same as ARMA model. Work [5] applies ARIMA on European dataset for West Nile Virus. On other hand Zhang in [6] applies ARIMA along with neural networks. Combination of SVM and ARIMA is applied in work [7].

Degree of non-stationarity is denoted by d. Here first time series will be applied with differencing for making it stationary. Then resultant time series is given to ARMA for forecasting. Differencing is applied by,

$$y' = y_t - y_{t-m}$$

Recurrent Neural Network (RNN)- These are neural networks designed to handle recurring patterns. They are Reservoir computing based. Patterns which repeat themselves over time or space can be easily detected by them. This is achieved by maintaining information regarding past patterns [8]. Even nonlinear time series can be modeled using these networks [9, 10]. Recently even pre-trained models are available to be directly be used for time series predictions [11]. In RNN output is dependent on input and hidden state maintained by the network [12]. The current hidden state is dependent on previous hidden state. So, RNNs systematically model sequential dependence of input. Assume, at time t consider yt as output, for input x_t and current hidden state mt . wxh represents weights assigned to connection between input and hidden state, whh represent weights between hidden states h_t and h_{t-1}. who represents weights for hidden state to output connection. W represents all types of weights. In simplest form these dependencies can be expressed as,

$$y_t = f(h_t, W)$$

$$h_t = g(h_{t-1}, x_t, W)$$

2.1 Data Analytics

In this section various facets of data are discussed. The dataset considered here is West Nile Virus dataset.

Missing Values In time series there are peculiar missing values dealing with no observations on a particular date within the rage of observation period [13]. As such dates fall under observation period they cannot be neglected or dropped. Such neglect or dropping will lead to improper time series analysis. In such case there can be multiple strategies which can be applied to this situation. The dates with no observation could be added with zero or NA observation as applicable. In some cases these can be filled with mean value. But putting mean values will distort curvature of time series and may lead to wrong analysis.

Leap Year Analysis In observation period there may be cases of leap years. There are statistical methods to detect such years. As these years will contain extra observations, analysis should be inclusive. Also special attention need to be given in finding mean or mode vales as the period is varying. In current data such case was handled and data analysis presented is based on proper values. In literature methods for handling such variations are presented in work [14].

Stationarity A stationarity series is distribution of samples independent of time. First step for successful handling of time series is to make it stationarity that is to remove any trends or seasonal patterns. Stationarity of a series can be tested using summary statistics or using Dickey-Fuller test [15].

$$\Delta y_t = \delta y_{t-1} + u_t$$

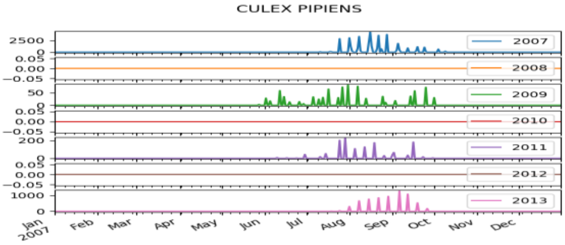
In summary if mean and variance of various parts of series are roughly in same range then series can be stable. In terms of Dicky-fuller test if p-value (expressed as d in formula) is under threshold, then time series is stationary.

Segregation of Multiple Time Series In provided data, there can be multiple embedded time series. First segregation of each time series is necessary to find valid patterns individually or cumulatively. If not segregated then they may lead to patterns.

2.2 Spatio-Temporal Data Analytics

Data about geographic spread of an entity or item over a period of time can be termed as spatio-temporal data [16]. Work [17] makes a detailed analysis of how geographic data varying over time can be presented and analysed. Wide variety of literature like [18, 19] focus on object detection using spatio-temporal analysis. Spatio temporal data has typical characteristics as listed below.

Importance by Time- Significance of this data decays over time. Patterns have significant effect only in loci of time.



Importance by Space- Events separated by space can be easily termes as independent. Certain events within specific periphery can easily be correlated. So, space plays major role in interdependence of variables.

Variations by Time and Space- When both features combined together, it provides different perspectives. Analyzing and using such patterns leads to enrichment of data for decision making.

3. Dataset Description

3.1 Basic Statistics

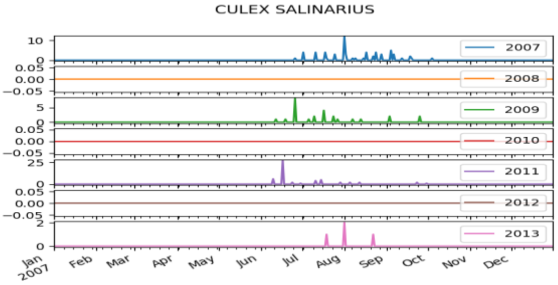
Data is taken from kaggle competition [20, 21]. It contains weather parameters, location of mosquitos' nets and insecticide sprays from year May-2007 to September-2013. The data is recorded for 95 days, for 7 varieties of mosquito species in 136 traps across Chicago. So, there are total 952 potential time series each spanning six years of time. Details are given in Table 1.Total number of observations are 10,506. This suggests that there are missing observations as on a particular day a species of mosquito was not observed in a trap. So, all such values can be safely assumed to be zero.

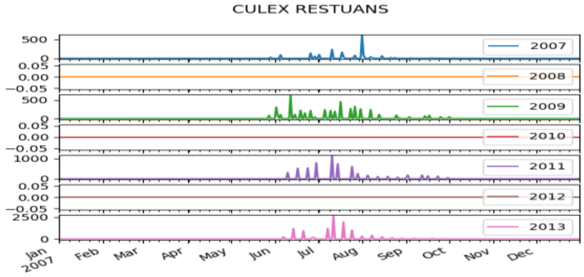
Table 1: Basic Statistics of Dataset

Data Head	Description	Value
Mosquito	Traps Period of observation (Days)	2316
	No of Positive Mosquito Observations	10506
	No of Positive Virus Observations	551
	Total Days of Positive Mosquito Observations	95
	Total Days of Positive Virus Observations	53
	No of Traps (Locations) for Observation	136
	No of Species of Mosquitoes Observed	7
	Potential No of Time Series	952
Pesticide Spray	Total Days of Spray	10
	Total Spray Records (Location wise)	14836
Weather	No of Parameters	19
	No of Days of records	1472
	No of Stations data	2

3.2 Cumulative Trends with Time

Trends in Number of Mosquitoes Species wise count of mosquitoes is presented in fig1(a,b,c,d). These graphs show how there is sudden increase in number of mosquitoes in month of June - October. All other times they are not active. Also, this occurrence varies species wise which can be seen here. Seasonal trend is clear with these graphs. Observations are made over 2,316 days and only 95 days have positive observations of mosquitoes out of which only 53 days have WNV virus detected.





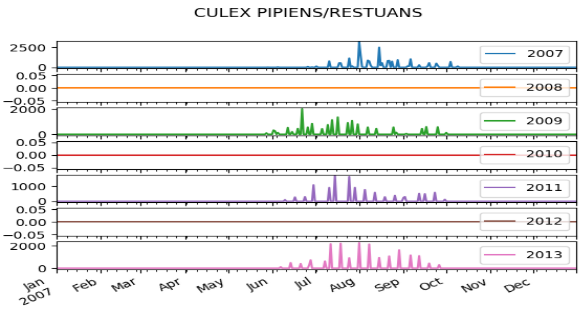


Fig. 1. **(a,b,c,d)**Number of Mosquitos(2007-2013)

So, all observations except these days are null or zero. This makes time series very challenging. Also, spraying efforts were taken only for 10 days in second half of this observation period. So, finding significant spraying pattern or effect pattern is a hilarious task.

Seasonal Mean Number of Mosquitoes The seasonal trend can be further made clear by mean number of mosquitoes found between June and October. Here in figure 2(a,b) show, species wise peak values differ but trend is mostly common. Also, there is gradual decrease towards end of the season. Mean analysis shows how two species Culex pipiens and restauns dominate overall observations over all years and over all blocks as carriers of WNV.

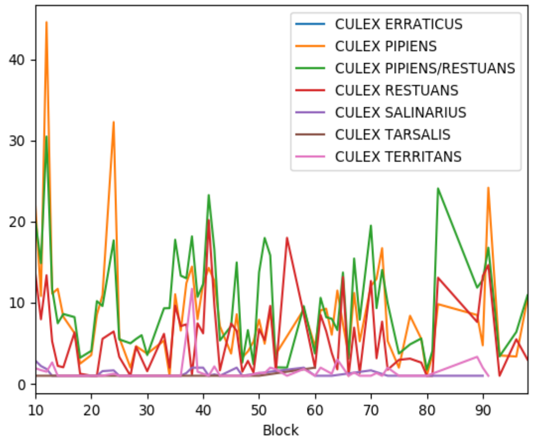
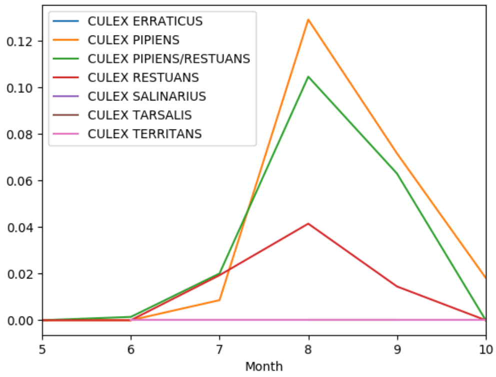


Fig. 2. (a,b) Mean No of Mosquitos

Trends in West Nile Virus occurrence Through monthly mean plot in figure 3 (a,b) it is clear that this virus trend also lies in June to October. It is a seasonal trend. Block wise plot of WNV shows, Culex Pipines breed are most likely to carry this virus. Culex Erraticus and culex satalnus is not found to carry WNV.



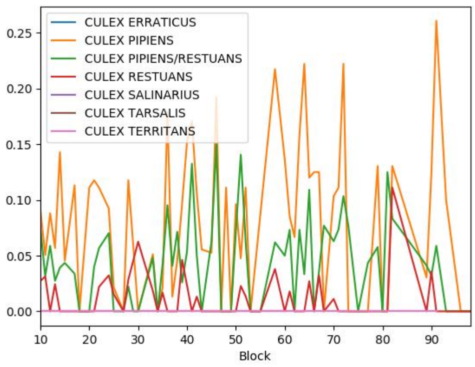
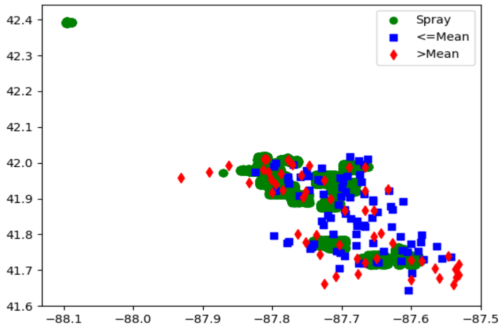


Fig. 3. (a,b) Count of WNV Occurence

•**Species Wise analysis of Mosquitoes** Mean value of count analysis shows varied pattern in species monthly population. Culex pipiens are peak in august and restauns peak in august. Also, some data is unidentified about species identification between ciles pipies / restauns. As there is significant amount of such data, it cannot be directly converted or removed or merged. So, it is kept under separate heading. Culex Erraticus and Culex Tarsalis are very rarely occurring species. These observations can safely be neglected.

3.3 Locationwise Analysis

In specific areas high number of mosquitoes and WNV was observed. Figure 4 (a,b) show details of the spread. Red spots indicate high number of mosquitoes in first figure and high occurrence of WNV virus. It can be easily observed that high population of mosquitoes and WNV virus co-occur in space. Further it can seen that no particular block or location is house for mosquitoes. Their population is distributed over complete space. Spraying efforts are seen in same areas where mosquitoes have high population. This means spraying was done on right target. Additionally, spraying efforts in 2013 has shown few changes in pattern of occurrence and population of mosquitoes.



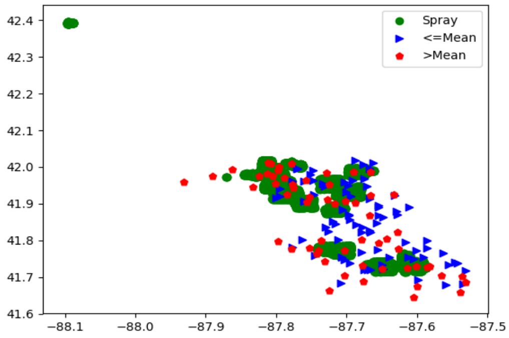


Fig. 4. (a,b) Latitude & Longitudewise spread of mosquitos and spray activities

3.4 Weather Analysis

Weather plays major role in breeding and overall life span of mosquitoes. Fig 1 shows analysis of every parameter of weather over complete span of observation. Observations in this dataset are collected from two fixed weather stations. Weather parameters are strongly related with occurrence of West Nile virus. So, all parameters of weather conditions are taken in consideration for predictions.

4. Challenges

In this section challenges posed by discussed dataset are presented. Also, general solutions for that challenge from literature are pointed. Additionally steps taken on this dataset regarding the challenge are discussed.

Size of Data In presented data there are total 136 locations of observations for 7 different species of mosquitoes. As there is sufficient segregation between locations, so population at each location is dependent only on mosquito population at that location separated by time. This results in 952 potential independent time series. Here each time series is observed for 2,316 days. So potential number of independent observations is 22,04,832. This is significant size of observations to be considered. Adding weather data to each observation will make each time series multi-dimensional. So, such size can be a potential challenge to be handled by solution makers. With apropriate resources this challenge can easily be handled.

Extreme Imbalance In presented data ratio of positive observation to negative observation is negligible. Even in cumulative manner when species are considered together, there are only 551 records with positive virus observed compared to 10,506 mosquito existence observations. If total mosquito occurrences (10,506) are compared to all the observations over complete time span (22,04,832), this also reflects extreme imbalance. Such imbalance can be handled by applying balancing techniques, or by considering cumulative statistics only. Balancing techniques can be under sampling or over sampling [22, 23]. Here, under sampling over spread of time is applied than random under sampling.

Missing Values Here there are only 10K instances available out of total expected instances of 2.2 million. This a huge part of time series is missing. Work [13] presents systematic handling of missing values in time series. Additionally, method of imputation can be applied as shown in work [24]. Here, for this dataset, missing values are treated as zeros. This is based on assumption that no observation indicates absence of mosquito so the virus.

Multi-Variate Time Series In this dataset multiple time series are presented. Each time series is assumed to have association with weather observations. So each time-series is having multi-variate model. This further takes challenge to next level. Work [25] present Bayesian solution to this problem, while [26] provides neural network solution. Appropriate system for forecasting can be chosen by researchers to provide acceptable detection and exclusion.

Locality Data behavior is expected as per the locality of occurrence. Mosquitoes separated by space a expected have independent behavior. In this dataset locations recorded have different accuracy. So, assumption made about locality may not hold in few cases. Although this may have very small effect but if modeled may lead to better predictions.

5. Results

On given dataset traditional Neural Network(NN) and Reservoir Computing based Recurrent Neural Network (RNN) is trained. Prediction for the presence of virus on a given day with perticular weather condition is a binary prediction problem. Following figures show comparison of performance of NN vs RNN. Performance is measured for validation dataset, reserved for testing. Results clearly show how RNN outperforms traditional NN.

6. Conclusion

Spatio-temporal datasets pose unique challenges for researchers in forecasting domain. Here, one case study dataset is discussed in details with data analytics, general methods of forecasting, challenges and reservoir computing based solutions. Challenges like missing data handling, class imbalance problem are discussed here. Results show, reservoir computing based Recurrent Neural Network outperforms tradiional methods for forecasting.

References

[1] J. D. Hamilton, Time series analysis, vol. 2. Princeton university press Princeton, 1994.

[2] K. Astrom and T. Soderstrom, “Uniqueness of the maximum likelihood estimates of the parameters of an arma model,” IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 769–773, 1974.

[3] K. H. Chon and R. J. Cohen, “Linear and nonlinear arma model parameter estimation using an artificial neural network,” IEEE Transactions on Biomedical Engineering, vol. 44, no. 3, pp. 168–174, 1997.

[4] Y. Shen, J. Guo, X. Liu, Q. Kong, L. Guo, and W. Li, “Long-term prediction of polar motion using a combined ssa and arma model,” Journal of Geodesy, vol. 92, no. 3, pp. 333–343, 2018.

[5] T. A. Groen, G. L’ambert, R. Bellini, A. Chaskopoulou, D. Petric, M. Zgomba, L. Marrama, and D. J. Bicout, “Ecology of west nile virus across four european countries: empirical modelling of the culex pipiens abundance dynamics as a function of weather,” Parasites & vectors, vol. 10, no. 1, p. 524, 2017.

[6] G. P. Zhang, “Time series forecasting using a hybrid arima and neural network model,” Neurocomputing, vol. 50, pp. 159–175, 2003.

[7] P.-F. Pai and C.-S. Lin, “A hybrid arima and support vector machines model in stock price forecasting,” Omega, vol. 33, no. 6, pp. 497–505, 2005.

[8] J. T. Connor, R. D. Martin, and L. E. Atlas, “Recurrent neural networks and robust time series prediction,” IEEE transactions on neural networks, vol. 5, no. 2, pp. 240–254, 1994.

[9] E. Egrioglu, U. Yolcu, C. H. Aladag, and E. Bas, “Recurrent multiplicative neuron model artificial neural network for non-linear time series forecasting,” Neural Processing Letters, vol. 41, no. 2, pp. 249–258, 2015.

[10] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, “Recurrent neural networks for multivariate time series with missing values,” arXiv preprint arXiv:1606.01865, 2016.

[11] P. Malhotra, V. TV, L. Vig, P. Agarwal, and G. Shroff, “Timenet: Pretrained deep recurrent neural network for time series classification,” arXiv preprint arXiv:1706.08838, 2017.

[12] B. A. Pearlmutter, “Learning state space trajectories in recurrent neural networks,” Neural Computation, vol. 1, no. 2, pp. 263–269, 1989.

[13] J. Honaker and G. King, “What to do about missing values in timeseries cross-section data,” American Journal of Political Science, vol. 54, no. 2, pp. 561–581, 2010.

[14] W. R. Bell and S. C. Hillmer, “Modeling time series with calendar variation,” Journal of the American statistical Association, vol. 78, no. 383, pp. 526–534, 1983.

[15] D. A. Dickey and W. A. Fuller, “Distribution of the estimators for autoregressive time series with a unit root,” Journal of the American statistical association, vol. 74, no. 366a, pp. 427–431, 1979.

[16] H. Huang, “Spatio-temporal data analysis,” Handbook Of Medical Statistics, p. 215, 2017.

[17] D. J. Peuquet and N. Duan, “An event-based spatiotemporal data model (estdm) for temporal analysis of geographical data,” International journal of geographical information systems, vol. 9, no. 1, pp. 7–24, 1995.

[18] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3361–3368, IEEE, 2011.

[19] Naftel and S. Khalid, “Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space,” Multimedia Systems, vol. 12, no. 3, pp. 227–238, 2006.

[20] Kaggle.com, “PredictWest Nile Virus.” https://www.kaggle.com/c/predict-west-nile-virus, 2008.

[21] C. of Chicago, “Health data and Reports.” https://www.cityofchicago.org/city/en/depts/cdph/provdrs/health_data_and_reports.html.

[22] X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory undersampling for class-imbalance learning,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 2, pp. 539–550, 2009.

[23] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” Intelligent data analysis, vol. 6, no. 5, pp. 429–449, 2002.

[24] P. K. Hopke, C. Liu, and D. B. Rubin, “Multiple imputation for multivariate data with missing and below-threshold measurements: Timeseries concentrations of pollutants in the arctic,” Biometrics, vol. 57, no. 1, pp. 22–33, 2001.

[25] G. Koop, D. Korobilis, et al., “Bayesian multivariate time series methods for empirical macroeconomics,” Foundations and Trends® in Econometrics, vol. 3, no. 4, pp. 267–358, 2010.

[26] K. Chakraborty, K. Mehrotra, C. K. Mohan, and S. Ranka, “Forecasting the behavior of multivariate time series using neural networks,” Neural networks, vol. 5, no. 6, pp. 961–970, 1992.