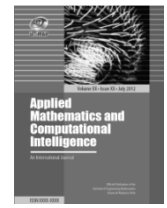
App. Math. and Comp. Intel., Vol. 2 (1) (2013) 137–148

http://amci.unimap.edu.my






# Modeling of prediction system: An application of the nearest neighbor approach to chaotic data

**N. Z. A. Hamid[a,b,*] and M. S. M. Noorani[b]**

[a]*Department of Mathematics, Faculty of Science and Mathematics,*
*Universiti Pendidikan Sultan Idris*
*35900, Tanjung Malim, Perak, Malaysia*

[b]*School of Mathematical Sciences, Faculty of Science and Technology,*
*Universiti Kebangsaan Malaysia,*
*43600, Bangi, Selangor, Malaysia*

Received: 20 February 2012; Revised: 12 February 2013, Accepted: 25 February 2013

**Abstract:** *This paper is about modeling of chaotic systems via nearest neighbor approach. This approach holds the principle that future data can be predicted using past data information. Here, all the past data known as neighbors. There are various prediction models that have been developed through this approach. In this paper, the zeroth-order approximation method (ZOAM) and improved ZOAM, namely the k-nearest neighbor approximation (KNNAM) and weighted distance approximation method (WDAM) were used. In ZOAM, only one nearest neighbor is used to predict future data while KNNAM uses more than one nearest neighbor and WDAM add the distance element for prediction process. These models were used to predict one of the chaotic data, Logistic map. 3008 Logistic map data has been produced, in which the first 3000 data were used to train the model while the rest is used to test the performance of the model. Correlation coefficient and average absolute error are used to view the performance of the model. The prediction results by the three models are in excellent agreement with the real data. This shows that the nearest neighbor approach works well to predict the chaotic data. Unfortunately, increasing the number of nearest neighbors from ZOAM to KNNAM not managed to improve prediction performance. However, the added element of the distance is a great idea for improving prediction performance. Overall, WDAM is the best model to predict the chaotic data compared to ZOAM and KNNAM.*

**Keywords:** *Chaos theory, chaotic data, nearest neighbour approach, zeroth-order approximation method, k-nearest neighbor approximation method, weighted distance approximation method, prediction, Logistic map.*

**PACS:** *05.45.-a, 05.45.Gg, 05.45.Tp.*

* Corresponding Author: nor_zila@yahoo.com (N. Z. A. Hamid)

## 1 Introduction

This study is an extension of the study by [1] and [2] who have proven that Malaysia rainfall data are chaotic in behavior. Since the data are chaotic, we plan to use chaos theory to do prediction of rainfall data. However, there is a lot of thought and belief that rainfall data may contain noise ([3], [4], etc.), which can affect the prediction process. So, in this study the prediction method from chaos theory will be applied to synthetic data in advance to see its effectiveness. In addition, tests on synthetic data can reduce costs and save time because they are free from noise disturbance. If the prediction results are good, so, in the future, research will be done on rainfall data using the same methods.

In this paper, the zeroth-order approximation methods (ZOAM) and improved ZOAM, the k-nearest neighbor approximation method (KNNAM) and weighted distance approximation method (WDAM) was constructed to predict the chaotic data via nearest neighbor approach, NNA. NNA is an approach that uses past data (neighbors) to predict future data. ZOAM uses only one nearest neighbor while KNNAM uses more than one nearest neighbor and WDAM adds distance element for prediction process. These models were used to predict one of the synthetic data, Logistic map.

### *1.1 Prediction Models*

#### *1.1.1 Chaos Theory*

In chaos theory, all types of systems are considered nonlinear. A linear system is a system whose evolution is a linear process. All systems that are not linear are called nonlinear systems. In these systems, the change in a variable at an initial time can lead to a change in the same or a different variable at a later time that is not proportional to the change at the initial time.

With respect to time domain, nonlinear systems were divided into two categories; discrete and continuous. The general forms of systems are $x(t+1) = f(x(t))$ for discrete time and $dx/dt = f(x)$ for continuous time. A discrete time system is also called a map. Equations of this form are called difference equations. Some examples of discrete system are Henon, Ikeda and Logistic map. On the other hand, a continuous time system is called a flow. Equations of this form are called ordinary differential equations. Examples of continuous system are Lorenz and Rossler flow. In this study, chaos theory was implemented into one of discrete system, Logistic map [5]. Logistic map equation is

$$x(t+1) = r * x(t) * (1 - x(t)) \tag{1}$$

where the behavior of the system was controlled by parameter $r$ and initial condition $x(1)$. The changes of $r$ and $x(1)$ may impact the system's behavior.

The impacts of varying $r$ are as follow. Lets generate 50 data of the Logistic map with $x(1) = 0.8$. As illustrated in Fig. 1, for $0 < r < 1$, the system is converging to 0 as $t \to \infty$. For $1 < r < 3$ the system becomes stable at one value as $t \to \infty$. Furthermore, for $3 < r < 1 + \sqrt{6}$, the system is oscillating between 2 values as $t \to \infty$. Furthermore, for $r > 1 + \sqrt{6}$ system is oscillating between 4 or 8 or 16 or $2^n$ values as $t \to \infty$. Moreover, it will have infinite oscillation points and there are no specific behaviors that can describe the system. Hence, this type of behavior is called chaos and the data of the system are known as chaotic. So, in this study on chaotic data, we have to choose $r > 1 + \sqrt{6}$.

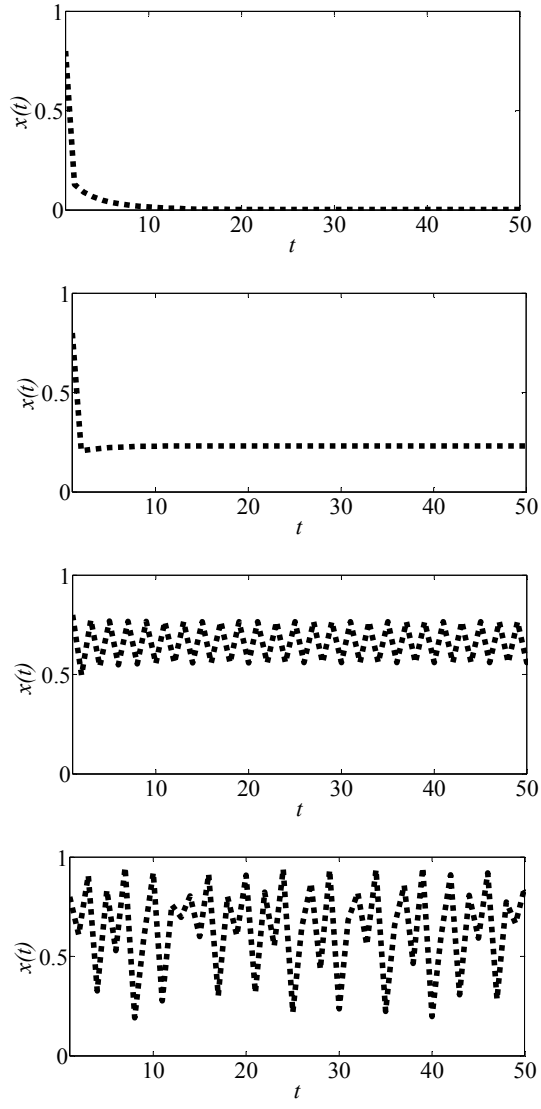



Figure 1: Impacts of varying $r$ on Logistic map.

From top, $r = 0.8$, $r = 1.3$, $r = 3.1$ and $r = 3.8$.

In theory, if the data are proven chaotic, then, they exhibit some properties [6]: (i) they are governed by one or more control parameters; (ii) a small change in which can cause the chaos to appear or disappear; (iii) their governing equations are nonlinear; (iv) they exhibit sensitive dependence on initial conditions and hence they are unpredictable in the long run. (i), (ii) and (iii) properties have been described above where the Logistic map was controlled by parameter $r$ and we can observe that a small change in $r$ values can cause the chaos to appear or disappear and the equation is obviously nonlinear.

In order to examine (iv) let us consider the Logistic map again. But, we vary the initial condition, $x(1) = 0.8$ and $x(1) = 0.81$. As in Fig. 2, after three iterations, for small difference of $x(1)$, the value of data as $t \to \infty$ are getting different and far from each other. Different $x(1)$ produce different outcome as time change. As time increase, the system yields widely diverging

outcomes. This is what it called sensitive dependence on initial conditions. This sensitivity makes the outcomes diverge and puts an effective limit on the ability to predict the behavior of chaotic systems over long periods of time. Hence, only short-term prediction can be made.

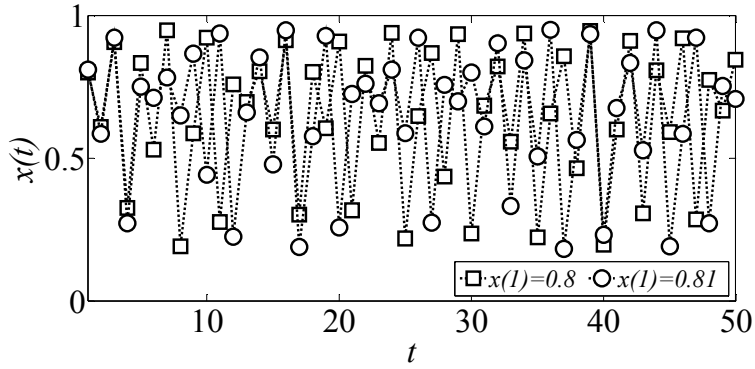



Figure 2: An illustration of sensitive dependence on initial conditions.

*1.1.2 Prediction Tools*

As to date, prediction tools in chaos theory were divided into two types, namely local and global approximation method. Via global method, all of the data in system were utilized to predict future data while local method only uses the nearby data (nearest neighbors) to the last known data in system in order to make predictions. Via global approximation method, we have to find and choose an appropriate functional form that represents the whole data. On the other hand, via local approximation method, the future data are been predicted base on nearest neighbor only. That's why, local method is also known as nearest neighbor approach, NNA. In general, global method provides good approximations if the real data are well behaved and not very complicated and we can fit a function to the whole data [7]. Since not all chaotic data (e.g. real data) can be represented well as a function, so, we decided to use the local method (or NNA) as the predictor.

NNA employed the principal that the future data are been predicted using information based on past data (neighbors). Local approximation method or NNA was introduced by [8] who applied it to the short term predictions of Mackey-glass, Rayleigh-Benard and Taylor-Couette flow. [3], [4], [9], [10], [11] and many more researchers employed this approach in various hydrology systems. Recently, in 2010, with a slightly modification, [12] utilized the approach to predict daily discharge in Russian. Moreover, in 2011, [13] modified it to predict storm in North Sea. However, dealing with real data is quite hard. [3], [4], [14-16] and many more highlighted the presence of noise in their studies. They found that NNA is employed base on the assumption that the data are noise-free. But it is well known that data from natural processes and experiments contain some amount of noise, such as the measurement error. When such a method, developed for noise-free data, is applied to noisy data, such as the rainfall data, it may not be possible to obtain very accurate results. Since we are still at preliminary stage on implementing NNA in Malaysia, we test it on noise-free data first. If this study is success, then we will proceed to apply the concept of NNA to real system (with noise reduction) in our future research.

*1.2 Research Objectives*

In this study, three prediction models developed from NNA will be tested on the chaotic data. ZOAM uses only one nearest neighbor while KNNAM uses more than one nearest neighbor. So, from ZOAM to KNNAM, we want to study whether the growing number of nearest neighbors is

used, the better the prediction. WDAM add distance element for prediction process. This is different from KNNAM because through KNNAM, the prediction is based on the average value of the nearest neighbor. Therefore, we wanted to investigate whether adding distance element will give better results. Finally, we want to find the best model to predict the chaotic data. Hence, several questions are raised:

1. Do NNA can be used to predict chaotic data?
2. Do use more neighbors can provide a better prediction?
3. Do add distance as weighting elements will predict the data better?
4. Which is the best model to predict the chaotic data?

3008 Logistic map data will be generated, where 3000 data were used to train the model while the rest is used to test the performance of the model. Correlation coefficient and average absolute error are used to view the performance of the model. The next section will describe the data and methodology. Then, results, discussion and conclusions section will answer all questions that arise above. Finally, there are some suggestions for future research at the end of this paper.

## 2 Data

We generate $N = 3008$ Logistic map data as (1) with $r = 4$ and $x(1) = 0.8$. Eight last data are kept to be compared with the predicted ones. Logistic map data that been used to train the models with $N = 3000$ are as Fig. 3.

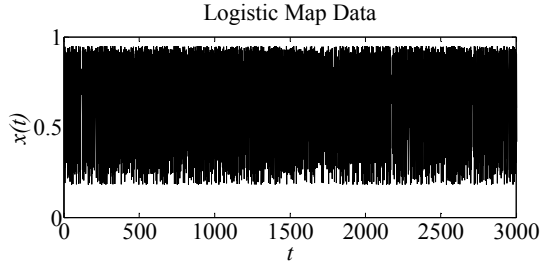



Figure 3: Logistic map data for $N = 3000$

## 3 Methodology

### *3.1 Zeroth-Order Approximation methods (ZOAM)*

As in [7], ZOAM is as follow. Say we have data of the system $\mathbf{X} = x_1, x_2, x_3, ..., x_N$ with $N$ total number of data and we want to predict the value of $x_{N+1}$. At first, we have to search the nearest neighbor to last known data, which is $x_N$. The nearest neighbor is a neighbor in neighborhood point with minimum Euclidean distance, $d$ where $d = |x_N - x_S|$, $s = 1,2,3,..., N-1$.

Say $x_p$ is the nearest neighbor to $x_N$. Then, $x_{p+1}$ is an approximation of $x_{N+1}$. For simple example, say we have $N = 10$, where the last known data is $x_{10}$. After calculating $d$, we found that $x_7$ is the nearest neighbor to $x_{10}$. Hence, $x_{7+1} = x_8$ is an approximation of $x_{11}$.

### 3.2 K-Nearest Neighbors Approximation method (KNNAM)

Moreover, for KNNAM, $k$ nearest neighbors were used and $x_{N+1}$ is the value of average approximation with

$$x_{N+1} = \frac{\sum_{i=1}^{k} x_{i+1}}{k} \tag{2}$$

Extending from above example, say we use three nearest neighbors, $x_7$, $x_3$ and $x_8$. Hence, the approximation of $x_{11}$ is $x_{11} = \frac{\sum_{i=1}^{3} x_{i+1}}{3} = \frac{x_8 + x_4 + x_9}{3}$.

### 3.3 Weighted Distance Approximation method (WDAM)

For WDAM, we added the distance value as a weight. If $d_i$ is the distance between $x_i$ and $x_N$, then

$$x_{N+1} = \frac{\sum_{i=1}^{k} (x_{i+1}) d_i}{\sum_{i=1}^{k} d_i} \tag{3}$$

Extending from above example again, say we use two nearest neighbors, $x_3$ and $x_8$. And the distance from $x_{10}$ to $x_3$ and $x_8$ are 0.2 and 0.4 respectively, hence,

$$x_{11} = \frac{\sum_{i=1}^{k} (x_{i+1}) d_i}{\sum_{i=1}^{k} d_i} = \frac{(x_3)(0.2) + (x_8)(0.4)}{0.2 + 0.4}.$$

### 3.4 Computing Step

In this study, parameter that we vary is the number of nearest neighbor, $k$. For ZOAM, $k = 1$ while for KNNAM $k = 2$ and $k = 3$ was used. Furthermore, $k = 2$ was used in WDAM. Below are the computing steps:

a. Generate $N = 3008$ data of Logistic map. 3000 data are used to train the models, 8 data are kept to be compared to the predicted ones.
b. Predict 8 data using ZOAM with $k = 1$, compare with real data using performance measure in Section 3.5.
c. KNNAM is done via Equation (2) with $k = 2$ and $k = 3$ and performance measure in Section 4.5 are used to compare with real data.

d. WDAM is done with $k = 2$ using Equation (3) and performance measures in Section 3.5 are used to compare with real data.

### *3.5 Performance Measure*

Correlation coefficient $cc$ was utilized to summarize information on how close the relationship between the real and predicted data. $cc$ value is ranging from -1.00 to +1.00. A $cc$ of 0.00 reflects that there is a zero correlation, or no relationship, between the real and predicted data. The closer a $cc$ is to 0.00, the weaker the relationship is. Oppositely, the closer $cc$ approaches plus or minus 1.00 the stronger the relationship is and reflects that the real and predicted data are close to each other. Meaning, the nearer the $cc$ value to plus or minus 1.00, the better the prediction is. The formula of $cc$ is:

$$cc = \frac{m\left(\sum_{w=1}^{m} x_w y_w\right) - \left(\sum_{w=1}^{m} x_w\right)\left(\sum_{w=1}^{m} y_w\right)}{\sqrt{\left[m\sum_{w=1}^{m} x_w{}^2 - \left(\sum_{w=1}^{m} x_w\right)^2\right]\left[m\sum_{w=1}^{m} y_w{}^2 - \left(\sum_{w=1}^{m} y_w\right)^2\right]}} \tag{4}$$

with $x_w$ and $y_w$ are real and predicted data where $w = 1,2,3,...,m$ is prediction day(s).

In addition, average absolute error, $e$ also used to observe the average different value between the predicted and the real ones. The formula of $e$ is:

$$e = \frac{\sum_{w=1}^{m} |x_w - y_w|}{m} \tag{5}$$

The smaller value of $e$ shows the prediction is better.

## 4 Results and Discussion

Tables 1 and 2 are the results of all three prediction models with various values of $k$. Figure 4-7 depict a comparison between the real data and prediction data. Figures 8 and 9 show the comparison graph of $cc$ and $e$ values for each model. Points in the graph (Figure 4-7) are not clear as the differences between real and predicted data are small. That is why we attach Table 1 as a reference. All predicted data agree with the real data. All the $cc$ are close to 1 and all $e$ are low. This suggests that, all prediction models are great and NNA is a good approach to predict the noise-free data such as Logistic map.

We are also interested to know how many past data (nearest neighbors) needed to perform prediction. We only choose the number of nearest neighbors of 1, 2 and 3. We want to check whether the more $k$ is used, the better prediction obtained. As we can see in Table 1 and 2, the prediction performances are better when $k$ is increasing from 1 to 2. However, the prediction performance decreases from $k = 2$ to $k = 3$. This shows that the growing number of nearest neighbors do not necessarily contribute to better prediction. Sometimes it is better, sometimes not. However, as a whole, all prediction is closer to the real value of the data. Therefore, we agree with Casdagli (in [17]) that a small number of nearest neighbors was enough to get a good prediction.

We would like to know if adding distance as weight will produce better prediction. From the $cc$ and $e$, it is clear enough to conclude that it is worthwhile to add distance elements as weights. We can see that WDAM is the best model to predict the chaotic data than other models. This means that the distance is quite important element to be added in order to obtain a better prediction.

Therefore, we will precede this method in our future research to real data (i.e. rainfall). Below are the answers for questions arise in Research Objectives part.

a. Do NNA can be used to predict chaotic data?
*From the values of $cc$ and $e$, we can see that all variations of the model gives a good prediction of the $cc$ values close to 1 and the values of $e$ are small. Therefore, yes, NNA can be used to predict chaotic data.*

b. Do use more neighbors can provide a better prediction?
*Increasing the number of nearest neighbors does not contribute to better prediction. Sometimes it is better, sometimes not. But, overall, for $k = 1$ to $k = 3$, the values of prediction is quite close to the real ones. Therefore, we agree with Casdagli (in [17]) that a small number $k$ is enough to get a better prediction.*

c. Do add distance as weighting elements will predict the data better?
*From the values of $cc$ and $e$, it is clear enough to conclude that the addition of the distance element is worth it. Therefore, the distance between the points of the neighborhood is quite important to be added in order to improve prediction performance.*

d. Which is the best model to predict the chaotic data?
*Through the values of $cc$ and $e$, it is clearly reflected that WDAM is the best model compared ZOAM and KNNAM.*

Table 1: Prediction results (data)

| REAL | ZOAM (1) | KNNAM (2) | KNNAM (3) | WDAM (2) |
|---|---|---|---|---|
| **0.6073** | 0.6079 | 0.6079 | 0.6079 | 0.60785 |
| **0.9062** | 0.9058 | 0.9058 | 0.9058 | 0.90580 |
| **0.3230** | 0.3243 | 0.3242 | 0.3244 | 0.32425 |
| **0.8309** | 0.8327 | 0.8327 | 0.8328 | 0.83265 |
| **0.5339** | 0.5294 | 0.5293 | 0.5294 | 0.52945 |
| **0.9456** | 0.9467 | 0.9467 | 0.9467 | 0.94670 |
| **0.1954** | 0.1916 | 0.1917 | 0.1917 | 0.19165 |
| **0.5973** | 0.5886 | 0.5888 | 0.5889 | 0.58875 |

Table 2: Prediction results (performance measures)

| *Model* | *cc* | *e* |
|---|---|---|
| **ZOAM (1)** | 0.999916392 | 0.002775 |
| **KNNAM (2)** | 0.999919000 | 0.002738 |
| **KNNAM (3)** | 0.999919167 | 0.002750 |
| **WDAM (2)** | 0.999919536 | 0.002725 |

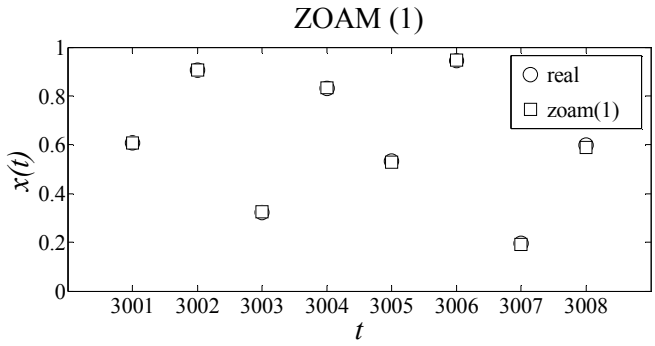



Figure 4: Comparison between the real and prediction data for ZOAM model.

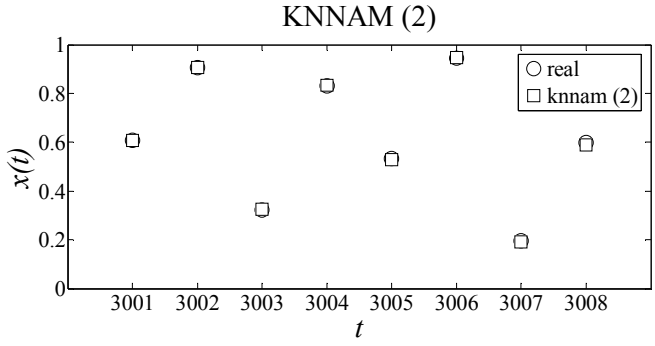



Figure 5: Comparison between the real and prediction data for KNNAM with 2 nearest neighbors model.

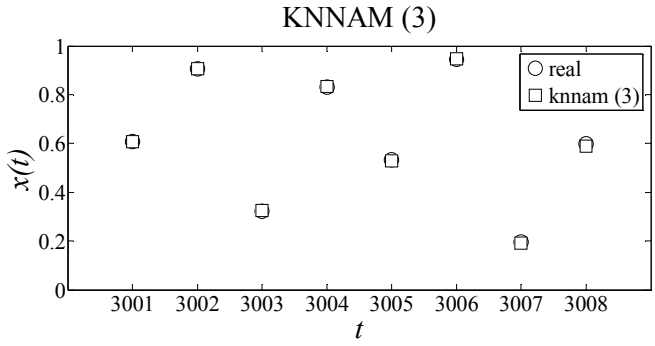



Figure 6: Comparison between the real and prediction data for KNNAM with 3 nearest neighbors model.

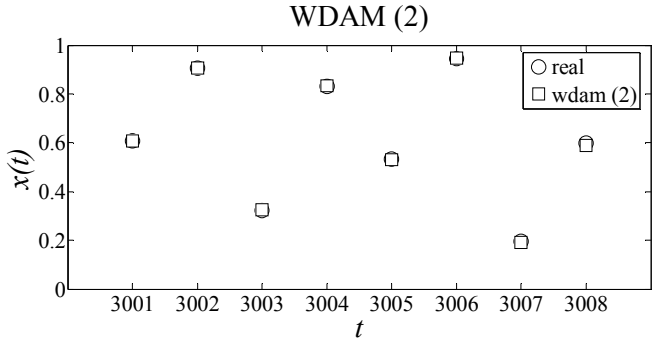



Figure 7: Comparison between the real and prediction data for WDAM with 2 nearest neighbors model.

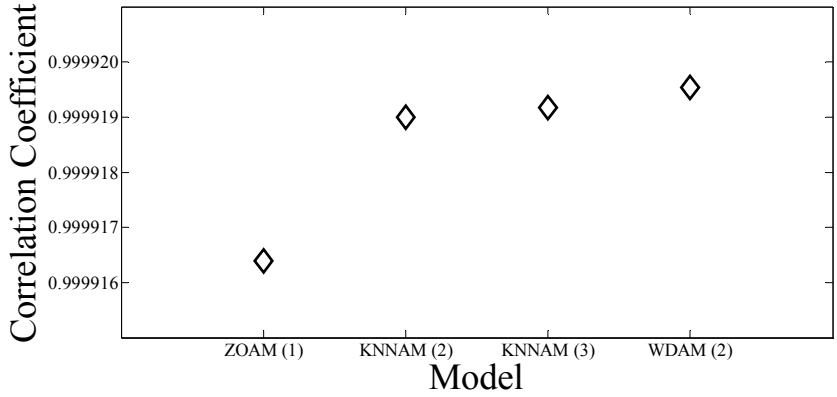



Figure 8: Comparison values of $cc$

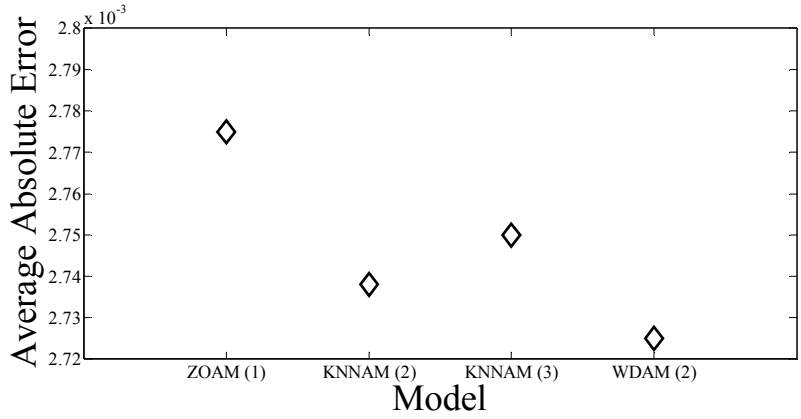



Figure 9: Comparison values of $e$

## 5 Conclusion and Further Studies

In this study, three prediction models namely ZOAM, KNNAM and WDAM based on nearest neighbor approach has successfully built and implemented on the chaotic data (Logistic map). All

prediction results are in a good agreement with the real value. This indicates that the data in the past is important in predicting future data. Through the values of correlation coefficient and average absolute error, it is clearly reflected that WDAM is the best model compared ZOAM and KNNAM. Hence, the distance between the points of the neighborhood is an important element to be added in order to improve prediction performance.

In the future, we would like to explore the method of how to find the optimal $k$ value. We know that a small number $k$ is enough to get a better prediction. But, we are not sure of the real value of appropriate $k$. We would also like to highlight that the study was conducted on synthetic system, a noise-free data. Since it has been successful, we recommend that before performing NNA on real data such as rainfall, the data must be cleaned first.

**Acknowledgments**

Special thanks go to Universiti Pendidikan Sultan Idris, Perak, Malaysia and Ministry of Higher Education, Malaysia for sponsoring this research studies.

**References**

[1] P. Radhakrishnan and S. Dinesh. An alternative approach to characterize time series data: Case study on Malaysian rainfall data. *Chaos, Solitons and Fractals*, 27:511–518, 2006.

[2] V. W. N. Betty, S. M. N. Mohd. & T. Fredolin. Deterministic behavior in Malaysian rainfall. *Proceedings of Applied Mathematics International Conference*, 2010.

[3] B. Sivakumar, S. Y. Liong, C. Y. Liaw and K. K. Phoon. Singapore rainfall behavior: Chaotic? *Journal of Hydrologic Engineering*, 38-48, 1999.

[4] B. Sivakumar, R. Berndtsson, J. Olsson, K. Jinno and A. Kawamura. Dynamics of monthly rainfall-runoff process at the Gota basin: A search for chaos. *Hydrology and Earth System Sciences*, 4:407-417, 2000.

[5] P. S. Addison*, Fractals and Chaos. An Illustrated Course*. IOP Publishing Ltd., 2001.

[6] J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, 2003.

[7] S. Velickov. *Nonlinear Dynamics and Chaos*. London. Taylor & Francis Group plc. 2004.

[8] J. D. Farmer & J. J. Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59:845-848, 1987.

[9] A. W. Jayawardena. Runoff forecasting using a local approximation method. *IAHS*, 167-171, 1997.

[10] B. Sivakumar, R. Berndtsson and M. Persson. Monthly runoff prediction using phase space reconstruction. *Hydrological Sciences Journal*, 46:377-387, 2001.

[11] M. N. Islam and B. Sivakumar. Characterization and prediction of runoff dynamics: A nonlinear dynamical view. *Advances in Water Resources*, 25:179–190, 2002.

[12] D. She and X. Yang. A new adaptive local linear prediction method and its application in hydrological time series. *Mathematical Problems in Engineering*, 2010:1-15, 2010.

[13] M. Siek and D. P. Solomatine. Real-time data assimilation for chaotic storm surge model using NARX neural network. *Journal of Coastal Research, Proceedings of the 11th International Coastal Symposium*, 1189 – 1194, 2011.

[14] B. Sivakumar, A. W. Jayawardena and T. M. K. G. Fernando. River flow forecasting: Use of phase-space reconstruction and artificial neural networks approaches. *Journal of Hydrology*, 265:225–245, 2002.

[15] A. W. Jayawardena and A. B. Gurung. Effect of noise in nonlinear hydrological time series analysis and prediction. *Hydrological Extremes: Understanding, Predicting, Mitigating*. 121-128, 1999.

[16] N. Z. A. Hamid and M. S. M. Noorani. Local approximation method: Application on daily rainfall prediction. *Proceedings of 1st AKEPT Young Researchers Conference & Exhibition 2011, Kuala Lumpur*, 531-541, 2011.

[17] A. Elshorbagy, S. P. Simonovic, & U. S. Panu. Estimation of missing stream flow data using principles in chaos theory. *Journal of Hydrology*, 255:123-133, 2002.