Development of waiting time predictor based Artificial Neural Network

V Limlawan\textsuperscript{1,2} and P Anussornnitisarn\textsuperscript{1,3}

\textsuperscript{1}Faculty of Engineering, Kasetsart University, Bangkok, Thailand
\textsuperscript{2}varis.lw@gmail.com, \textsuperscript{3}fengpta@ku.ac.th

Abstract. Queue is a part of our daily life. Everyday people have to wait in queue for services such as waiting for train, buying a cup of coffee. Unfortunately, it is difficult to eliminate waiting time in queue because it is very costly to do. Some business provides queue-length to each customer for reducing customer anxiety, but the given queue-length is often larger than the actual queue-length. Customers estimated the waiting time from the given queue-length and decide to leaves the queue, so the business may lose customer to their competitor. This paper aims to propose the design of the system that can provide the accurate estimated waiting time to each customer. The proposed system constantly updates queue-length and service rate. The updated queue-length and the updated service rate are used as the input of Artificial Neural Network (ANN) which is the waiting time predictor in the system. ANN with the proposed system is compared to the other predictors such as linear regression, the historical based predictor, Queue-Theory predictor. The result shows that ANN with the proposed system outperforms the others, and more than 95% of predicted waiting time by ANN is accurate within 5 minutes tolerances.

Keywords: queuing system, queuing theory, waiting time prediction, artificial neural network, delay estimation

1. Introduction

Queue is everywhere in daily activities and service such as bank, hospital, and canteen. People cannot avoid waiting in queue for services. Many studies show that long waiting in queue has a negative effect on service satisfaction although service quality is not quite bad [1, 2]. However, it is difficult to eliminate waiting time in queue because the costs and resources in service systems are limited.

To reduce dissatisfactions, many businesses provide entertainment stuffs for customers while waiting such as television, WiFi and coffee are provided for customers in a waiting area. Moreover, some business provides queue-length information to customers for reducing anxiety, but the given queue-length is often larger than the actual queue-length. Customers estimated waiting time from the queue-length information and decide to leave from the queue. Business loss revenue from leaving customers and may lose the customers to competitors. To solve the problem, providing accurate waiting time for each customer is another way to increase customer satisfaction. With accurate waiting time, customers can plan their activities while waiting for services.

The paper proposes a queue system that can provide the accurate waiting time to each customer. The remaining of this paper is organized as follows. Method of this study is described in section 2.
Experiment result and discussion are presented in section 3. Finally, this work is concluded in section 4.

2. Methods

2.1. Waiting time prediction

The goal of our proposed system is to propose the accurate waiting time estimation for each customer, so waiting time predictors are reviewed and discussed as follows.

2.1.1. Queue-theory based predictor. Queue-Theory based Predictor (QT) is based on a G/M/n queue system [3]. Predicting waiting time by QT is not complex and use only queue-length, service rate, and number of server in the prediction. Let $\hat{W}_{QT}$ be the waiting time predicted by QT, $q(t)$ be queue-length at time $t$, $u$ be service rate trained by the historical data, $s(t)$ be number of service at time $t$. $\hat{W}_{QT}$ is

$$\hat{W}_{QT} = \frac{q(t)+1}{s(t)u}$$  

Although QT is easy to implement, QT may not accurate in queue systems that has customer abandonment or reneging. The system does not update queue-length when customer leaves from the queue before receiving services. Queue-length used in QT is often larger than the actual queue-length.

2.1.2. Historical based predictor. Last customer to enter service (LES) is the historical based predictor which use the waiting time of the customer who recently enter service [4, 5]. LES is not based on queue-length, so incorrect queue-length from customer abandonment does not affect LES. However, queue environment such as arrival rate, service rate may affect the performance of LES. The queue environment of the last customer who enters service might not be the same as the environment of the new customer. For example, the last customer waits for 10 customers in the queue before receiving service, but the new customer has only 2 people in front of him in the queue. Common sense we know that the waiting time of new customer shorter than the last customer.

2.1.3. Linear regression. Linear Regression (REG) is applied in many applications [6]. In waiting time prediction, variants of linear regression such as regression spline, quartile regression were applied to predict waiting time in a call centre and an emergency room [7, 8]. Queue-length and number of server are used as the main independent variables in those predictors.

In this paper, the standard linear regression is used in the experiment. The independent variables are queue-length divided by number of server, and the response is waiting time. The linear regression for waiting time prediction is

$$\hat{W}_{REG} = \beta_0 + \beta_1 \frac{q(t)}{s(t)}$$  

where $\hat{W}_{REG}$ is waiting time predicted by the linear regression, $\beta_0, \beta_1$ are unknown regression coefficients, $q(t)$ is Queue length at time $t$ and $s(t)$ is the number of server at time $t$.

Naturally, linear regression performs well in the problem that response and independent variables have linear relationship. LREG may not accurate if queue-length and number of server do not have linear relationship with waiting time. Moreover, incorrect queue-length may affect the performance of LREG as QT.

2.1.4. Artificial neural network. Artificial Neural Network (ANN) is inspired by nervous system of human brain. ANN is a powerful tool that can solve multi-dimensional inputs and outputs problems [9,
10]. ANN has been applied in many applications such as stock prediction, wind forecasting [11]. In waiting time prediction, ANN has been applied to predict waiting time in call-center and bank [7, 12].

In those works, the architecture of ANN is a single hidden layer feedforward, and the main inputs are queue-length and number of server. The results of ANN are slightly better than QT and REG in the queue system with no abandonment, but the results of ANN is somewhat in the queue system with abandonment.

Based on those works, there are no works which consider incorrect queue-length problem from customer abandonment. Incorrect queue-length may affect the performance of waiting time predictor which has queue-length in the predictive models such as QT, REG, ANN.

2.2. The proposed queue system

Nowadays, ticket queue system is commonly installed to managing a queue system. Customers arrive and get a ticket which contains queue information such as queue-length, estimated waiting time. After customer got the ticket, customers wait for calling at waiting area.

The ticket queue system make customers feel more relax than the physical queue. Instead of standing in queue, customer can do anything during waiting. However, the disadvantage of the ticket queue system is inaccurate queue information [13]. The ticket queue system does not update queue-length when customer left the system before receiving service, so queue-length and estimated waiting time is often larger than the actual waiting time.

To solve the problem, the proposed system is designed to provide the accurate estimated waiting time to each customer. The design of the system is illustrated in figure 1. Unlike the ticket queue, the proposed system constantly updated queue-length. All cases of arrival and departure are counted by the system. Moreover, service time is also monitored and used to adjust service rate which is the main variable in several waiting time predictors. Every time the system updated queue-length and service rate, the system will send them to waiting time predictor. The predictor will use them to estimate waiting time for a new customer.

![Figure 1. Diagram of the proposed queue system.](image)

To monitor queue-length and collect service time, we roughly plan to use the video surveillance and computer vision technologies. Recently, video surveillance has been commonly found in many places, and its cost is inexpensive for every organization. With computer vision technologies, video surveillance can be used for many purposes such as car detection, people counting. The detail of using the technology is beyond the scope of this paper, so we will not discuss in detail.

Service time of each customer is sent to Exponential Weight Average (EWMA) control chart for adjusted service rate in waiting time predictor. EWMA control chart is a tool for detecting mean shift [14]. Let \( u_0 \) be an initial service rate from the historical data, \( u \) be the current service rate, \( \sigma \) be the
standard deviation of service rate from the historical data, $z_i$ be service rate of customer $i$, $\lambda$ and $L$ be EWMA parameters. The step of EWMA for controlling service rate is as follows

Step 1 Initialize parameters $u = u_0$, $z_0 = u_0$, $\lambda$ and $L$

Step 2 Calculate $z_i$ from EWMA statistics as follows.

$$z_i = \lambda z_{i-1} + (1 - \lambda) z_i$$

(3)

Step 3: Calculate lower bound (LCL) and upper bound (UCL) as follows.

$$UCL = u_0 + L\sigma \sqrt{\frac{\lambda}{2 - \lambda}} (1 - (1 - \lambda)^2)$$

(4)

$$LCL = u_0 - L\sigma \sqrt{\frac{\lambda}{2 - \lambda}} (1 - (1 - \lambda)^2)$$

(5)

Step 4: If $z_i > UCL$ or $z_i < LCL$, then the service rate is updated by $u = z_i$

Step 5: Repeat step 2

Once the service rate and queue-length is updated, they will be sent to waiting time predictor. Based on the advantage of ANN, the main predictor of our system is ANN. The architecture of ANN in our system is the single hidden layer feedforward which has three layers as shown in figure 2. The inputs of ANN are queue-length, service rate and number of server. Based on [15], the single hidden layer feedforward is expressed as follow.

$$p = \begin{bmatrix} u(t) \\ Q(t) \\ S(t) \end{bmatrix}, \text{ for } t = 1, 2, 3,...$$

(6)

$$a^0 = p$$

(7)

$$a^{i+1} = f^{i+1}(w^{i+1}a^i + b^{i+1}), \text{ for } i = 1, 2, 3,...$$

(8)

$$a^2 = W_{ANN}$$

(9)

where $u(t)$ is the updated service rate at time $t$, $Q(t)$ is the updated queue-length at time $t$ and $S(t)$ is the number of server at time $t$, $t$ is discrete time, $p$ is the inputs from the external source, $a^i$ is the input vector of layer $i$, $b^i$ is the biased vector of layer $i$, $w^i$ is the weights vector of layer $i$, $f^i$ is the activate function of layer $i$.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{figure2.png}
\caption{A single hidden layer feedforward ANN.}
\end{figure}
2.3. Numerical experiment
To evaluate the proposed queue system, ANN with the proposed system is compared to QT without the system, LES, REG with the proposed system. The test system is described as follow.

2.3.1. Test system. The proposed system is evaluated by the dynamic queue system which was conducted by SIMAN ARENA® Rockwell software. The detail of the tested system is as follows.
- The system opens at 8.00 a.m. and closes at 16.00 p.m.
- The service time \( (u) \) is normal distribution with a mean of 1 and a standard deviation of 0.75.
- The load of the system \( (p) \) is dynamically changed in a replication.
- The arrival rate is Poisson distribution with a mean of \( \lambda \) which is adjusted based on the load \( \rho \) and service rate \( \mu \).
- The reneging time of customers is 2-Erlang distribution with a mean of 15.
- The summary of test system is shown in table 1.

In the numerical experiment, there are 3 cases for testing ANN with the proposed system. The cases are divided by the shift of service rate. There is no mean service time shift in case 1, and there is a negative shift and positive shift on the case 2 and the case 3 respectively. The shift size is +0.5 and -0.25 in case 2 and case 3. For all cases, the shift in service time begins at 11.00 a.m.

| Time (hr) | Arrival Rate | Average Service Time (u) | Load (p) | Renege Rate |
|-----------|--------------|--------------------------|----------|-------------|
| 8.00-9.00 | Norm(1,0.75) | 0.85                     |          |             |
| 9.00-10.00| Norm(1,0.75)| 0.85                     |          |             |
| 10.00-11.00| Norm(1,0.75)| 0.95                     |          |             |
| 11.00-12.00| Norm(1±Shift size,0.75)| 0.99                   |          | 2-Erlang distribution with a mean of 15.|
| 12.00-13.00| POISSON(\( \lambda \)) depends on load (p) and avg. service time (\( \mu \))| | 0.99 |
| 13.00-14.00| Norm(1±Shift size,0.75)| 0.95                     |          |             |
| 14.00-15.00| Norm(1±Shift size,0.75)| 0.85                     |          |             |
| 15.00-16.00| Norm(1±Shift size,0.75)| 0.85                     |          |             |

2.3.2. Parameter setting. For training ANN and REG, we simulated 30 replications of the mixed case which combined all cases together, and we simulated 10 replications for each test case. For ANN, the architecture of ANN is a single hidden layer feedforward. The number of neuron in the hidden layer is 5. The activated function of the hidden layer is tan-sigmoid function, and the activated function of the output layer is pure linear function. ANN is trained by backpropagation algorithm (traingd) in MATLAB R2016a. The learning rate of the backpropagation algorithm is 0.05. For other predictors, QT, REG and LES are coded by Excel VBA. Moreover, \( \lambda \) and L of EWMA control chart are set at 0.1 and 2.7 respectively.

2.3.3. Performance measurement. To compare the performance of ANN with the proposed system with the other predictors, Mean Square Error (MSE) is used as the performance measurement in the experiment. The error and MSE are as follows.

\[ e_i = W_i - \hat{W}_i \]  

(10)
\[
MSE = \frac{1}{n} \sum_{i=1}^{n} e_i^2
\]  
(11)

where \( e_i \) is the prediction error, \( W_i \) is the actual waiting time of customer \( i \), \( \hat{W}_i \) is the predicted waiting time of customer \( i \) and \( n \) is the number of the test sample.

Moreover, percentage of accuracy within an acceptable tolerance PA is used to measure the performance of the best predictor. Let \( l \) be the acceptance tolerance, \( n \) be number of test samples, \( e_i \) be an error of \( i \) sample, \( c_i \) is the accuracy indicator of \( i \) sample. The percentage of accuracy is defined as follow.

\[
PA \pm l = \frac{1}{n} \sum_{i=1}^{n} c_i \times 100
\]
(12)

where,
\[
c_i = \begin{cases} 
1 & , e_i \in [-l,l] \\
0 & , e_i \notin [-l,l] 
\end{cases}
\]

3. Result and Discussion

The comparison of MSE for each predictor is summarized in table 2. QT is the best predictor in case 1. MSE of ANN with the system is less than QT in case 2 and case 3, by 5.7% and 25.6% respectively. MSE of LES is approximate 4.3 for all cases. REG is the worse predictor in all cases. The comparison of predicted waiting time of QT and ANN in case 3 is illustrated in figure 3. As shown in the comparison, ANN with the system is slightly better than QT in case 3.

| Case                  | QT   | LES   | REG (with system) | ANN (with system) |
|-----------------------|------|-------|-------------------|-------------------|
|                       | MSE  | MSE   | %Diff*            | MSE              | %Diff*            | MSE | %Diff* |
| Case 1: No shift      | 2.10 | 4.17  | -98.9             | 6.52             | -210.7            | 2.23 | -6.30  |
| Case 2: Positive shift| 2.11 | 4.50  | -113.4            | 6.31             | -199.2            | 1.99 | 5.70   |
| Case 3: Negative shift| 6.48 | 4.47  | 31.1              | 6.28             | 3.1               | 4.82 | 25.6   |

Figure 3. Comparison of actual waiting time and predicted waiting time of QT and ANN plot in Case 3.
Table 3 summarized the accuracy percentage of ANN with the system which is the best predictor in MSE comparison. The result shows that more than 95% of predicted waiting times are accurate within 5 minutes, and more than 99% of predicted waiting times are accurate within 10 minutes.

Based on the result, ANN with the system outperforms other predictors in the case that has the shift in service time, but QT is the best predictor in the system that does not have the shift in the service rate. On the other hand, REG with the system is the worse predictor in all cases because the training set is not specific for any cases, and REG has no ability to adapt the model when the service time was shifted.

| Case               | Accuracy (%) within ±1 min | Accuracy (%) within ±3 min | Accuracy (%) within ±5 min | Accuracy (%) within ±10 min |
|--------------------|----------------------------|----------------------------|-----------------------------|-----------------------------|
| Case 1: No shift   | 62.00%                     | 94.90%                     | 98.80%                      | 99.90%                      |
| Case 2: Positive shift | 61.10%                  | 95.00%                     | 99.50%                      | 100.00%                     |
| Case 3: Negative shift | 41.30%                  | 83.30%                     | 96.70%                      | 99.90%                      |

4. Conclusion
In this paper, we proposed the design of the queue system that can improve the accuracy of waiting time prediction for each customer. The proposed system solves the losing queue information in the traditional ticket queue system by constantly updating queue-length and service rate. Due to the availability of camera devices and computer vision technology, we roughly plan to use cameras as a sensor to monitor the queue-length and other queue information, so all cases of customer abandonment are counted and update in the system. For service rate, we use Exponential Weight Average (EWMA) control chart to monitor and adjust service rate of the system. Both service rate and queue-length updated from the system are sent to waiting time predictor which is Artificial Neural Network (ANN).

ANN with the proposed system is compared to Queue Theory predictor without the system, the historical based predictor, and linear regression predictor with the proposed system. The result shows that ANN with the system outperforms other predictors in the case that has customer abandonment and a shift in service rate. Moreover, more than 95% of predicted waiting time by ANN is accurate within 5 minutes.

The proposed system solved the losing queue information and can provide the accurate waiting time estimation for each customer. With the accurate waiting time, customers can manage their time and activities while they wait for a service. For future work, more cases on service time and more cases customer abandonment will be studied for investigating the effect of them.

5. Reference
[1] Tarloy S 1994 Waiting for Service: The Relationship between Delays and Evaluations of Service. *Journal of Marketing*. **52**, 56-69.
[2] Pruyn, A T H and Smidts A 1998 Effects of waiting on the satisfaction with the service: Beyond objective time measures. *International Journal of Research in Marketing*. **15**(4), 321-334.
[3] Whitt W 1999 Predicting queueing delays. *Management Science*. **45**, 880-888.
[4] Ibrahim R and Whitt W 2008 Real-time delay estimation in call centres. *Proceedings of the 2008 Winter Simulation Conf* pp. 2876-2883.
[5] Ibrahim R and Whitt W 2009 Real-Time Delay Estimation Based on Delay History *Manufacturing & Service Operations Management* **11**(3), 397-415.
[6] Montgomery D C and Runger G C 2010 *Applied statistics and probability for engineers*. 5th ed (Hoboken, NJ, Wiley).
[7] Thiongane M, Chan W and Ecuyer P L 2015 Waiting time predictors for multi-skill call centres. Proceedings of the 2015 Winter Simulation Conference pp. 3073-3084.
[8] Sun Y, Teow K L, Heng, B H, Oni C K and Tay S Y 2012 Real-Time Prediction of Waiting Time in the Emergency Department, Using Quantile Regression. Annals of Emergency Medicine. 60(3), 299 – 308.
[9] Abiodun O I, Jantan A, Omolara A E, Dada K V, Mohamed N A and Arshad H 2018 State-of-the-art in artificial neural network applications: A survey. Heliyon 4(1).
[10] Moghaddam A H, Moghaddam M H and Esfandyari M 2016 Stock market index prediction using artificial neural network. Journal of Economics, Finance and Administrative Science 21, 89-93.
[11] Ahmed A and Khalid M. 2017 Multi-step Ahead Wind Forecasting Using Nonlinear Autoregressive Neural Networks. Energy Procedia 134, 192-204.
[12] Satya Hermanto R P, Suharjito D and Nugroho A 2018 Waiting-Time Estimation in Bank Customer Queues using RPROP Neural Networks. Procedia Computer Science. 135, 35-42.
[13] Xu S H, Gao L and Ou J 2007 Service Performance Analysis and Improvement for a Ticket Queue with Balking Customers. Management Science. 53(6), 971-990.
[14] Montgomery D C 2012 Introduction to Statistical Quality Control 7th ed (John Wiley & Sons, Inc).
[15] Hegan M T, Demuth H B, Beale M H and Jess O D 2014 Neural Network Design 2nd eds. [online].