Journal of Convergence for Information Technology
Vol. 9. No. 12, pp. 54-61, 2019
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2019.9.12.054

Time Series Forecasting on Car Accidents in Korea Using Auto-Regressive Integrated Moving Average Model

Hyunkyung Shin
Associate Professor, Department of Financial Mathematics, Gachon University

자동 회귀 통합 이동 평균 모델 적용을 통한 한국의 자동차 사고에 대한 시계열 예측

신현경
가천대학교 금융수학과 부교수

Abstract Recently, IITS (intelligent integrated transportation system) has been important topic in Smart City related industry. As a main objective of IITS, prevention of traffic jam (due to car accidents) has been attempted with help of advanced sensor and communication technologies. Studies show that car accident has certain correlation with some factors including characteristics of location, weather, driver's behavior, and time of day. We concentrate our study on observing auto correlativity of car accidents in terms of time of day. In this paper, we performed the ARIMA tests including ADF (augmented Dickey-Fuller) to check the three factors determining auto-regressive, stationarity, and lag order. Summary on forecasting of hourly car crash counts is presented, we show that the traffic accident data obtained in Korea can be applied to ARIMA model and present a result that traffic accidents in Korea have property of being recurrent daily basis.

Key Words : Time series analysis, stochastic stationary time series, ARIMA model, Augmented Dickey-Fuller (ADF) test, Forecasting, Car accident data

요 약 최근 들어 IITS는 스마트 시티관련 산업계에서 중요한 주제로 떠오르고 있다. IITS의 주요 목적인 교통체증 (차량 사고에 기인한) 예방책들이 발전된 센서 및 통신 기술의 도움을 받아 다양하게 시도되었다. 관련 연구들에서는 자동차 사고와 사고 위치적 특성, 날씨, 운전자 행동, 시간 등 다양한 요인들과 상관 관계가 있음을 보여주고 있다. 우리 연구는 자동차 사고와 사고 발생 시간 사이의 상관관계에 주제를 집중했다. 본 논문에서는 ARIMA (Auto-Regressive Integrated Moving Average) 자동 회귀, 정상 및 지연 순서를 결정하는 세 가지 요소를 확인하기 위해 ADF (Augmented Dickey-Fuller)를 포함한 ARIMA 테스트를 수행했다. 본 연구 결과로서 시간 별 자동차 충돌 수 예측에 대한 요약을 제시하며, 한국 내 자동차 사고 데이터는 ARIMA 모델에 적용될 수 있음을 보여주었고, 국내 자동차 사고는 하루를 기준으로 일정한 주기가 존재하는 성격을 가지고 있다는 것을 제시했다.

주제어 : 시계열 분석, 확률 론적 고정 시계열, ARIMA 모델, ADF (Augmented Dickey-Fuller) 테스트, 예측, 자동차 사고 데이터

*Corresponding Author : Hyunkyung Shin(hyunkyung@gachon.ac.kr)
Received November 19, 2019
Revised December 6, 2019
Accepted December 20, 2019
Published December 28, 2019

1. Introduction

In signal processing, a time series is autoregressive when the series describes certain time depending processes [1]. Autoregressive model takes the previous steps as input and learns from the history to forecast next time step [2]. Generally, $AR(p)$ denotes autoregressive model of order p.

$$X[t] = c + \sum_{i=1}^{p} \varphi_i X[t-i] + \varepsilon_t \qquad (1)$$

, where c is constant, $\varphi_{1,\ldots,}\varphi_p$ are the parameters of the model, and is white noise. The number p defines lag variables $X[t-1],\ldots,X[t-p]$ [3]. Correlation between the lag variables and the output $X[t]$ is critical for predictability [4]. As discussed in machine learning, the higher correlation the more likely the past will predict [5,6]. For demonstration purpose, consider an unrealistic time series X defined as follows:

$$X[n] = \begin{cases} 1 \text{ if } & n = 7m+2 \\ 2 \text{ if } & n = 7m+4 \\ 0 \; otherwise & \end{cases} \qquad (2)$$

Fig. 1 is visualization of $X[n]$ for $1 \le n \le 24$.

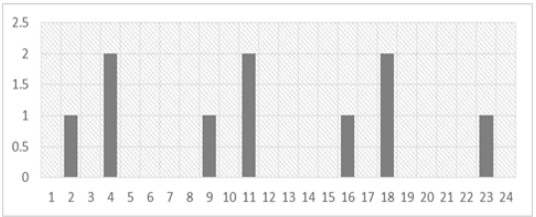

Fig. 1. Visualization of an example time series X[n]. It is autoregressive with the lag = 7

The X[n] defined above can be written as autoregressive model AR(7) as below:

$$X[n] = c + \sum_{k=1}^{7} a_k X[n-k] + \varepsilon_n \qquad (3)$$

For this example case, it is straightforward to find $c = 0, a_7 = 1, a_k = 0 (k \neq 7)$ and $\varepsilon_n = 0$. Once the values of coefficients a_k were identified, for an arbitrary m 〉 n, X[m] can be obtained from its history, i.e., forecastable. As seen in the example, predictability is equivalent to solvability of coefficient (a_k's) constraint equation. For the general problem case, ARIMA model is a basic and widely used. ARIMA models are denoted by

$$ARIMA(p,d,q) \qquad (4)$$

, where the three parameters account for seasonality, trend, and noise in data [7]. ARIMA is consisted of AR (auto regressive), I (integrated), and MA (moving average). First of all, the parameter p specifies the number of lags used in regression equation as below

$$\begin{aligned} x[t] - m &= a[1]*(x[t-1]-m) + \cdots \\ &+ a[p]*(x[t-p]-m) + e[t] \end{aligned} \qquad (5)$$

with

$$x[t] = X_d[t] \qquad (6)$$

, where $X_d[t]$ denotes d^{th} differencing of $X[t]$ which is the time series of interest, a's are model parameters of AR coefficient, m is a constant equivalent to mean value of $x[t]$, and $e[t]$ is error term compensating discretizing errors due to finite terms. For example, ARIMA(2,0,0) can be written as

$$\begin{aligned} x[t] - m &= a[1]*x[t-1] + a[2]*x[t-2] + e[t] \\ x[t] &= X_0[t] \end{aligned} \qquad (7)$$

, where the coefficients a's and e are as explained above. It is just another form of Markov process with two previous time steps. Secondly, in the integrated I(d) component, represents the degree of differencing. For example, computation of follows a rule as below

$$X_1[t] = X[t] - X[t-1]$$
$$X_2[t] = X[t] - X[t-2]$$
$$X_d[t] = X[t] - X[t-d] \quad (8)$$

. The main purpose of differencing is to stabilize non-stationary time series. Details are discussed later. Thirdly, moving average MA(q) component represents the error term as combination of the previous error terms. The parameter q determines the number of terms. The following formula shows use of parameter q

$$x[t] - m = a[1](x[t-1] - m) + \cdots + a[p](x[t-p] - m) + e[t] + b[1]e[t-1] + \cdots + b[q]e[t-q] \quad (9)$$

, where the terms are explained already above.

Selection of an ARIMA model for target application depends on size of lags for AR, differencing, and moving average and which can be written as a linear equation

$$Y_i = c + \xi_t Y_{dt-1} + \xi_p Y_{dt-p} + \cdots + \varsigma_1 e_{t-1} + \varsigma_q e_{t-q} + e_t \quad (10)$$

where p is lag, d is differencing step, q is size of moving average. Y_d is the Y differenced d times and c is a constant.

For its mathematical completeness, ARIMA has been applied to various scientific applications in the area of forecasting time series including financial, energy radiation, energy conservation, and other non-linear non-stationary time series behaviors [8-11]. However, up to our knowledge, no results were published for prediction on traffic accidents.

To achieve forecasting model on traffic accident using ARIMA, in this paper, we transformed police report data to time series. Details are discussed after the section of data description. With the time series data reconstructed from the raw police reports, certain criteria systemized by ARIMA tests on the time series should be satisfied prior to apply actual prediction task. Details on this process is also described at later section "Method". Figure 2 is a graph showing daily counts of car crashes reported to Police Department during three years of period in Korea. The red dotted line is the linear trend line, the blue dotted lines is 30 days moving average line. It shows typical behavior of autoregressive time series: season component such as less counts in earlier month of a year and more counts in end of a year, sharp noises in daily frequencies needing moving average, and fairly uniform oscillation pattern.

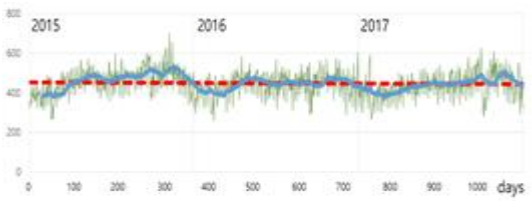

Fig. 2. Graph of daily count of car crashes in Korea during three years from 1.1.2015 to 12.31.2017. X-axis is number of days passed since Jan. 1. 2015. Y-axis denotes frequency of car accidents. Green graph represents daily frequency, red graph denotes linear trend line, and blue graph shows 30 step moving average

2. Data Description

The data used in this paper was obtained from Police Department after being pre-processed to avoid violation on personal information protection act and to protect deployment strategy of police department, personal information and police station information. It is basically police report data integrated in excel format. It should be mentioned that the raw police report data was preprocessed on the exact time of occurrence to protect police station information. Only the hour of accident is available to access. For an example, if an accident occurred at 9.30AM, it is recorded as 10AM.

Specification of data is presented in the Table 1 below. Total 491,545 records were collected during 1,095 days from Jan. 2015 to Dec. 2017.

Table 1. Specification of the car crash data in South Korea. 491,545 counts during the years 2015 to 2017

begin date	end date	total count	covered data	data source
2015.01.01	2017.12.31	491,545	South Korea	Police Department

3. Method

Prior to applying ARIMA model, the data described at the previous section needs to be converted into time series. The shortest time unit for the time series is interval of 1 hour since the raw data does not have minute value. For each time step, count up police reports to create a time series. The total data contains 1,095 days (=26,280 hours) which amounts to 26,280 time steps. The size of time series is too big to display, for the clarity of presentation, we sampled 1 month data (30 days = 720 hours) from Jan. 1 2015 to Jan. 30 2015. See Fig. 3 where the black curve is frequency and the red line is its linear trend line. the most important process of time series analysis is correct estimation of stationarity and convert into stationary if it is not. Visual inspiration suggests the graph is both stationary and auto-regressive. The red line in the middle of graph is linear trend with the value of linear coefficient 0.0019 which is almost flat, and periodicity of daily peak seems obvious. Therefore, this time series is a good candidate for an ARIMA model. We perform rigorous statistical tests to convince applicability of ARIMA model.

Outliers of time series including data entry error typically removed by moving average of order m

$$MA(m) = \frac{1}{m}\sum_{j=-k}^{k} \xi_{t+j} \qquad (11)$$

, where m = 2k + 1. In this study, we do not use moving average on the raw data (i.e. k = 0 for above formula). Fig. 3 shows shape of graph of raw time series without moving average. The red dotted line in the middle represent linear trend (0.0019x+15.318). This sample graph is obtained from January in 2015.

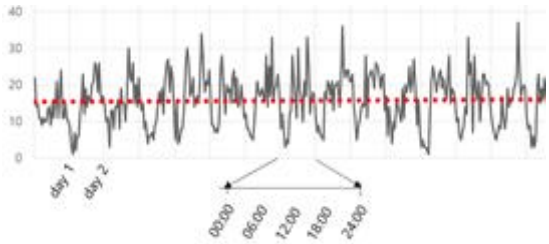

Fig. 3. Frequency data of car accidents. X-axis is date-time (HH:MM), Y-axis is frequency. The unit interval of x axis is 1 hour

The three fundamental parts of time series analysis for building a prediction model are seasonality, trend, and cycle [12]. Seasonality states variation in the data related to calendar cycles. From the Fig. 2, it shows less crash counts earlier month of year, which can say seasonal component. Trend refers overall pattern of time series. The blue dotted line in Fig. 2 is 30 days moving average to see trend component. Cycle denotes oscillated patterns which is not seasonal. Fig. 4 illustrates decomposition outputs with the sample time series data. Decomposition refers to the process of extracting the three components out of a time series. The residuals in the Fig. 4 is the remains of season, trend, and cycle.

The result of decomposition S_t, T_t, E_t reconstructs the raw data in the way of $X_t = S_t + T_t + E_t$.

Since ARIMA takes previous lags of time series to model its behavior, it requires that input time series is stationary. By definition, stationarity of a time series can be detected by time invariant of mean, variance, and auto-covariance.

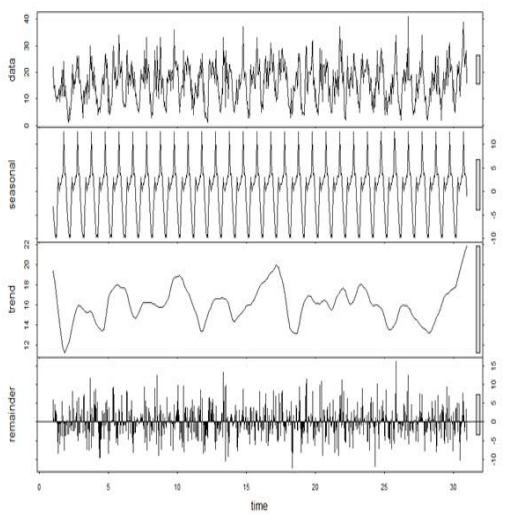

Fig. 4. Outputs of time series decomposition

Top row shows raw time series as input of decomposition. The rest three rows are output of decomposition. Seasonal, trend, and remainder components are displayed from the second row to last row. The augmented Dickey-Fuller (ADF) test is a formal statistical test for stationarity [13]. The null hypothesis assumes that the series is non-stationary. ADF procedure checks if variations in Y can be leveraged by lagged value and linear trend in time.The following script is a output of performing adf.test() command pre-installed in R. In the script, 'v' denotes array of data as illustrated in Fig. 3. 'ts(v)' indicates time series conversion.

Table 2 is the output of ADF test with our raw data. In this paper, "tseries" package in R was adopted to run the test. The following R-script shows procedures: the raw frequency data ('v') was converted to time series ('ts_carcrash') which in turn was decomposed by season, trend, cycle using 'stl' api command in R.

the resulting decomposed data('ts_carcrash_dec om') was adjusted by seasonal component ('ts_carcrash_deseason'). Finally through differencing it was centered on 0 value. Display of the series is impossible due to its huge size (26283 time steps).

```
ts_carcrash = ts(v, frequency = 24)  # number of observations in unit time (24 per day)
ts_carcrash_decom = stl(ts_carcrash[,1], s.window = "periodic")
ts_carcrash_deseason <- forecast::seasadj(ts_carcrash_decom)
ts_carcrash_deseason_diff1 <- diff(ts_carcrash_deseason, difference=1)
acf(ts_carcrash_deseason_diff1)
pacf(ts_carcrash_deseason_diff1)
```

Fig. 5. R-script of adftest in R

The processed data('ts_carcrash_deseason_diff1') is used to test. The table below is the real output of stationarity test. The p-value (<0.01) from Table 2 for ADF test indicates that the input time series is stationary. The small p-value indicates strong evidence against the null hypothesis. The alternative hypothesis is that the input time

series is stationary.

Table 2. ADF test outputs

Output of adf.test in R
Dickey-Fuller = -19.317, Lag order = 29, p-value = 0.01, alternative hypothesis: stationary

In order to select ARIMA model order (for the lag value) autocorrelation is useful [14]. Autocorrelation function (ACF) plots display correlation between series and its lag. Fig. 6 displays ACF and PACF (partial autocorrelation function) plots. As a summary statement, ACF determines MA(q) and PACF does AR(p) for an ARIMA model. The plots from R, shows blue dotted line which is 95% significance boundaries. Notice that lag1 implies 24 hours.

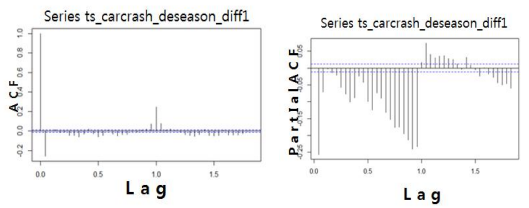

Fig. 6. Autocorrelation plots for stationarity testing and for choosing order parameters of ARIMA model

ACF figure above strongly suggests autocorrelation of input time series. The spikes at particular lags of the series give hint to the choice of p and q of ARIMA model. There are significant auto correlation at lag 0.1, and 1 while PACF plot shows significant spikes at the lag 1.0. The peak at 1 in ACF suggests presence of seasonal pattern as daily basis.

To fit to ARIMA model using 'forecast' package in R, 'auto.arima()' generates a set of optimal (p, d, q). As for the criteria of choosing the best, Akaike information criteria (AIC) is used. The following outputs (partial parts were cut out for better presentation) from R demonstrates model selection process using 'auto.arima()'. It shows the method minimizes AIC. AIC is an estimator of how much information would be lost by choosing the given model.

```
auto.arima(ts_carcrash_deseason_diff1, seasonal=TRUE, stationary = TRUE,
max.p = 24, d=1, stepwise = TRUE, approximation=TRUE, trace=TRUE)

 ARIMA(0,0,0)            with non-zero mean : 179037.7
ARIMA(1,0,0)(1,0,0)[24] with non-zero mean : 174115.9
ARIMA(0,0,1)(0,0,1)[24] with non-zero mean : 174598.8
 ARIMA(0,0,0)            with zero mean     : 179035.7
 ARIMA(1,0,0)            with non-zero mean : 177242.8
ARIMA(1,0,0)(2,0,0)[24] with non-zero mean : 172963.5
ARIMA(0,0,0)(2,0,0)[24] with non-zero mean : 176727.9
ARIMA(2,0,0)(2,0,0)[24] with non-zero mean : 171989.8
 ARIMA(2,0,0)(2,0,0)[24] with zero mean     : 171987.8
 ARIMA(2,0,0)(1,0,0)[24] with zero mean     : 173381.1
 ARIMA(1,0,0)(2,0,0)[24] with zero mean     : 172961.5
 ARIMA(3,0,0)(2,0,0)[24] with zero mean     : 171682.9
 ARIMA(13,0,0)(1,0,0)[24] with zero mean     : 172125.9
 ARIMA(15,0,0)(2,0,0)[24] with zero mean     : 170070.4
 ARIMA(16,0,1)(2,0,0)[24] with zero mean     : 168631
 ARIMA(16,0,1)(2,0,0)[24] with non-zero mean : 168633
 ARIMA(16,0,1)(1,0,0)[24] with zero mean     : 169424.3
 ARIMA(17,0,1)(2,0,0)[24] with zero mean     : 168481.3
 ARIMA(17,0,0)(2,0,0)[24] with zero mean     : 169824.1
 ARIMA(16,0,0)(2,0,0)[24] with zero mean     : 169946.7
 ARIMA(17,0,1)(2,0,0)[24] with non-zero mean : 168483.3
 ARIMA(17,0,1)(1,0,0)[24] with zero mean     : 169218.3
 ARIMA(18,0,1)(2,0,0)[24] with zero mean     : 168312.1
 ARIMA(18,0,0)(2,0,0)[24] with zero mean     : 169625.3
 ARIMA(19,0,2)(2,0,0)[24] with zero mean     : Inf
 ARIMA(18,0,1)(2,0,0)[24] with non-zero mean : 168314.1

 Best model: ARIMA(18,0,1)(2,0,0)[24] with zero mean
```

Fig. 7. R-script of ARIMA model in R

The selected model is (18, 0, 1) the model doesn't need differencing (d=0), and uses autoregressive terms of 18 lags, and moving average model of order 1. Summary part of outputs from 'auto.arima()' is presented separately as below.

```
Coefficients:
          ar1     ar2     ar3     ar4     ar5     ar6     ar7     ar8     ar9     ar10    ar11    ar12    ar13
      0.2403  0.0554  -0.0093  -0.0632  -0.0747  -0.0907  -0.0865  -0.0867  -0.0637  -0.0419  -0.0689  -0.1029  -0.0850
s.e.  0.0089  0.0071   0.0066   0.0065   0.0065   0.0065   0.0065   0.0065   0.0065   0.0065   0.0065   0.0065   0.0065
         ar14    ar15    ar16    ar17    ar18     ma1     sar1    sar2
      -0.0491  -0.0691  -0.0858  -0.0806  -0.0941  -0.8127  0.2074  0.1678
s.e.   0.0065   0.0065   0.0066   0.0067   0.0070   0.0066  0.0069  0.0065

sigma^2 estimated as 35.32:  log likelihood=-84124.86
AIC=168293.7   AICc=168293.8   BIC=168473.6
```

Fig. 8. Summary of coefficients in ARIMA model

The output above implies that the original time series can be fit with the following

$$X[n] = ar1^{*}X[n-1] + \cdots + ar18^{*}X[n-18] + ma1^{*}E[n] + \epsilon \quad (12)$$

, where ϵ is discrepancy between the original and the generated.

4. Results

We have shown that the time series data, from hourly frequency of car crashes during from 2015 to 2017 in Korea, is stationary and appropriate to adopt ARIMA model. Simulation result indicates (18,0,1) is the best parameter with corresponding 18 coefficient values and 1 coefficient for moving average term. The list of exact values for coefficient ar1,...,ar18 is presented in Table 3.

Table 3. List of coefficient of AR(p) for p = 18. Reading order is from left to right and then top to bottom

0.2403	0.0554	−0.0093	−0.0632	−0.0747	−0.0907
−0.0865	−0.0867	−0.0637	−0.0419	−0.0689	−0.1029
−0.0850	−0.0491	−0.0691	−0.0858	−0.0806	−0.0941

With the ARIMA model parameters selected by the procedures described above, forecast simulation is performed using the following R-script.

```
fit_carcrash_deseason_diff <-auto.arima(ts_carcrash_deseason_diff1)
forecast_carcrash_deseaon_diff <-forecast(fit_carcrash_deseason, h = 300)
```

Fig. 9. R-script of forecast simulation in R

The output of forecasting simulation is presented at Fig. 10. Unfortunately, the size of data point is too big to show prediction results which is attached at the end of existing data points. For the visualization purpose, prediction simulation with sampled data points (30 days amount) is demonstrated at Fig. 11.

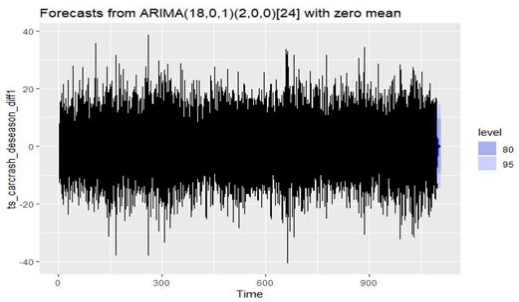

Fig. 10. Plot of forecasting result obtained from the R-script above

The extended lines for predicted time series is colored blue. It keeps reasonably well on the value range, oscillated shapes, and period of frequency cycles. This implies that our attempts to forecast the frequency of car crash counts by adopting ARIMA model is acceptable.

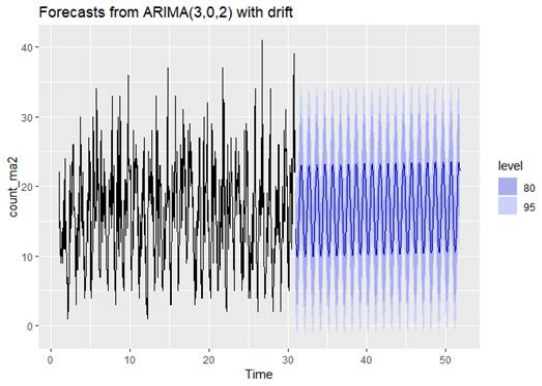

Fig. 11. The figure with large raw data is not visually helpful due to too many data points. This is sampled data (1 month) based forecasting output. The forecast plot shows its result clearly

5. Conclusion

Prevention of car accidents is one of the most important issues in industrialized countries in present time. In [15], comparison study of ARIMA and ARIMAX on traffic accidents in Nigeria. In predictability aspect, authors claim ARIMAX was superior. It will certainly be more critical issue in upcoming automatic driving era. Various factor analysis of car accident have contributed to improve situations. However, the car accidents occurs all the time as if there were no prevention strategies were ever applied. In this paper, we concentrated our research on temporal correlativity of car accidents to see if it is predictable. Our study is too early stage to make a statement on predictability of car accident. Our result shows that temporal

frequencies of car accidents are stationary enough to apply conventional ARIMA model. In layman's term, it says that we don't know where but we know when a car accident happens. The property found can be helpful to design a policy on prevention of car accidents.

In [16], big data analysis integrated with ARIMA was improved in terms of forecast traffic state. For the future work on predictive model on traffic state in automatic driving generation, we will have to design deep learning strategy to discover hidden feature of cause of accident. Temporal dependency of the newly found factor by using ARIMA test will be interesting.

REFERENCES

[1] D. P. Percival & A. T. Walden. (1998). *Spectral Analysis for Physical Applications*. Cambridge University Press.

[2] S. Theodoridis. (2015). *A Bayesian and Optimization Perspective. Academic Press, 9*,-51.

[3] F. Kadri, F. Harrou, S. Chaabane, Y. Sun & C. Tahon. (2016). Seasonal ARMA-based SPC charts for anomaly detection: Application to emergency department systems. *Neurocomputing*, *173*, 2102-2114.

[4] B. Soren. & M. Kulahci. (2011). *Time Series Analysis and Forecasting by Example*. Hoboken, N.J.: Wiley.

[5] C. Cheng et al. (2015). Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. *Iie Transactions*, *47(10)*, 1053-1071.

[6] M. Li & J.Y. Li. (2010). *On the predictability of long-range dependent series*. Mathematical Problems in Engineering

[7] M. Shukla & S. Jharkharia. (2013). Applicability of ARIMA models in wholesale vegetable market: an investigation. *International Journal of Information Systems and Supply Chain Management (IJISSCM), 6(3)*, 105-119.

[8] J. G. De Gooijer & R. J. Hyndman. (2006). 25 years of time series forecasting. *International journal of forecasting*, *22(3)*, 443-473.

[9] C. Voyant, M. Muselli, C. Paoli & M. L. Nivet. (2012). Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. *Energy, 39(1)*, 341-355.

[10] M. Lydia, S. S. Kumar, A. I. Selvakumar & G. E. P. Kumar. (2016). Linear and non-linear autoregressive models for short-term wind speed forecasting. *Energy conversion and management*, *112*, 115-124.

[11] H. O. Stekler. (2007). The future of macroeconomic forecasting: Understanding the forecasting process. *International Journal of Forecasting*, *23(2)*, 237-248.

[12] M. Theodosiou. (2011). Forecasting monthly and quarterly time series using STL decomposition. *International Journal of Forecasting*, *27(4)*, 1178-1195.

[13] R. J. Hyndman & Y. Khandakar. (2008). Automatic time series forecasting: The forecast package for R, *Journal of Statistical Software*, *26(3)*.

[14] F. Hadi & G. Joao. (2013). Event labeling combining ensemble detectors and background knowledge, *Progress in Artificial Intelligence*, 1-15.

[15] C. C. Ihueze & U. O. Onwurah. (2018). Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. *Accident Analysis & Prevention*, *112*, 21-29.

[16] T. Ma, Z. Zhou & C. Antoniou. (2018). Dynamic factor model for network traffic state forecast. *Transportation Research Part B: Methodological*, *118*, 281-317.

신 현 경(Hyunkyung Shin) **[정회원]**

· 2002년 : 뉴욕주립대학교 (스토니 부룩) 응용수학 박사
· 2007 ~ 현재 : 가천대학교 긍융수학과 부교수
· 관심분야 : 인공지능, 자연어처리, 영상처리, 가상현실, 수학정보교육
· E-mail : hyunkyung@gachon.ac.kr