https://doi.org/10.33778/kcsa.2024.24.2.151

보이스피싱 발생 추이 예측을 위한 시계열 모형 연구: 계절성과 외생변수 활용★

강 다 연*, 이 승 연*, 황 은 주**

요 약

최근 고금리와 고물가로 인해 민생의 불안정성이 가중되고 있는 현 사회에, 보이스피싱으로 인한 피해액 또한 증가하고 있다. 이러한 범죄는 기술 발전으로 인해 그 형태와 수법이 지속적으로 진화하고 있으며, 피해자들에게 심각한 금전적 및 정신적 피해를 야기하고 있다. 본 연구는 보이스피싱 발생건수를 더 정확하게 예측하기 위하여, 시계열 모형을 연구 비교하는 것을 목표로 한다. 보이스피싱 발생건수 데이터를 기반으로 ARIMA, SARIMA 모형과 외생변수로서 피해액, 검거건수, 검거인원의 조합을 고려한 SARIMAX 모형을 비교 분석한다. 표본 외 예측분석을 수행하여 예측값의 예측 성능을 검증한다. 예측구간을 추정하고 이의 경험적 포함확률을 도출함으로써 예측 모형의 우수성을 확인한다. 2024년 12월까지의 보이스피싱 월별 발생 건수를 예측하여, 향후 보이스피싱 대응 및 예방 전략 수립에 기여하고자 한다.

Time series models for predicting the trend of voice phishing: seasonality and exogenous variables approaches

Da-Yeon Kang*, Seung-Yeon Lee*, Eunju Hwang**

ABSTRACT

In recent years with high interest rates and inflations, which worsen people's lives, voice phishing crimes also increase along with damage. Voice phishing that becomes more evolved by technology developments causes serious financial and mental damage to victims. This work aims to study time series models for its accurate prediction. ARIMA, SARIMA and SARIMAX models are compared. As exogenous variables, the amount of damages and the numbers of arrests and criminals are adopted. Forecasting performances are evaluated. Prediction intervals are constructed along with empirical coverages, which justify the superiority of the model. Finally, the numbers of voice phishing up to December 2024 are predicted, through which we expect the establishment of future prevention strategies for voice phishing.

Key words : Voice Phishing, Time Series Analysis, Exogenous Variables, Forecast.

접수일(2024년 01월 24일), 수정일(1차: 2024년 04월 10일), 게재확정일(2024년 05월 27일)

★이 연구는 가천대학교 지원을 받아 수행되었음(202305040001).

* 가천대학교 응용통계학과 (주저자 이름순)

* These authors contributed equally to this work.

** 가천대학교 응용통계학과 부교수 (교신저자)

1. 서 론

보이스피싱(voice phishing)은 전화로 피해자의 불안감을 자극하여 개인정보를 편취하거나 결제를 유도함으로써 금전적 이익을 취하는 범죄 수법이다. 보이스피싱 범죄는 우리나라에서 처음 보고된 2006년 5월 이후로, 지난 17년 동안 사기의 형태와 수법이 지속적으로 발전해 왔다. 초기에는 상대적으로 정보에 취약한 노인층이 주요 범죄 대상이었으나, 현재는 연령이나 사회적 신분에 관계없이 피해자 범위가 확대되었다.[1] 최근에는 금융당국과 정부의 예방책 강화로 보이스피싱 건수는 줄어들고 있지만, 오히려 수법이 진화하면서 발생건수 대비 피해액은 증가하는 추세를 보인다. 그 수법이 끊임없이 진화하고 있는 만큼 이에 대한 대응 방안 또한 발전되어야 한다.[2]

본 연구는 보이스피싱 발생건수를 보다 정확하게 예측할 수 있는 시계열 모형을 선정하는 것을 목표로 한다. 이를 위해 보이스피싱 발생건수 데이터를 기반으로 ARIMA, SARIMA 모형과 피해액, 검거건수, 검거인원 등의 외생변수를 포함한 SARIMAX 모형을 적용하여 분석한다. 각 모형의 최적 차수를 결정하고, 이를 데이터에 적합시킨 뒤 비교 분석하여 가장 우수한 모형을 선정한다. 선정된 모형을 활용하여 표본 외 예측분석을 수행하여 예측 성능을 검증하고, 예측구간의 경험적 포함확률을 구하여 예측모형의 타당성과 정확성을 평가한다. 마지막으로 최종 모형을 통하여 예측된 2024년 12월까지의 보이스피싱 발생 건수 데이터를 제시한다.

2. 관련 연구

김도윤(2020)[3]에 따르면, 보이스피싱은 주로 행위지와 결과 발생지가 일치하지 않는 조직적 범행으로 국제 범죄성과 조직 범죄성을 가지며, 범행 수법과 조직 구성이 사회 상황을 반영하여 빠르게 변한다고 설명한다. 이 연구는 한국과 중국의 관련 법제와 정책적 동향에 대해서도 논의한다. 곽영암(2022)[4]은 코로나 상황에 따라 비대면 거래, 비대면 일상생활이 확산되며 진화하고 있는 보이스피싱의 현황과 사례를 분석하고 정부, 금융기관, 소비자단체, 이용자 측면에서 개선방안을 모색하였다. 추정호(2022)[5]는 보이스피싱 발생 추이의 예측을 위하여 X-12 계절성 조정 방법론으로 계절성을 조정하고, ARIMA 모형을 이용하여 2022년 보이스피싱 발생을 예측하였다.

보이스피싱에 관련된 선행연구를 검토한 결과, 보이스피싱 범죄의 계절성을 반영한 연구는 선행되었으나 보이스피싱 발생에 영향을 미치는 변수를 분석에 포함한 연구는 존재하지 않는다는 점을 발견했다. 이에 따라, 본 연구에서는 범죄 발생의 계절성을 고려하고, 피해액, 검거건수, 검거인원을 외생변수로 사용하여 분석을 진행한다. 피해액이 증가하면 범죄에 대한 사회적 경계심이 높아져 발생 빈도에 영향을 줄 수 있고, 검거건수와 인원이 증가하면 범죄 억제 효과로 이어져 발생률을 낮출 수 있다. 이러한 외생변수들의 추가는 기존 연구들에 비해 보이스피싱 발생 예측을 높이는 데 기여할 것으로 기대된다.

3. 방법론 및 분석 자료

3.1 시계열 모형 방법론

3.1.1 ARIMA(p, d, q)

ARIMA(AutoRegressive Integrated Moving Average)모형은 자기회귀(AR), 차분(Difference), 이동평균(MA)을 통합하며, 누적된 데이터의 경향성을 분석하는 데 매우 효과적이다. 이러한 특성으로 사회 문제와 관련된 시계열 분석에 적합하기에 초기 분석 모형으로 설정한다.[6] ARIMA(p,d,q) 모형 Y_t의 수식은 다음과 같다.

$$(1-\sum_{i=1}^{p}\Phi_i B_i)(1-B)^d Y_t = (1+\sum_{j=1}^{q}\Theta_j B^j)\epsilon_t \qquad (1)$$

여기서 Φ_i는 자기회귀계수, d는 차분의 차수, Θ_j는 이동평균계수, B는 후진 연산자, ϵ_t는 시점 t에서의 오차항을 의미한다. 비정상(nonstationary) 시계열 데이터를 d번 차분하여 정상(stationary) 시계열로 안정화한 후, 과거 시계열이 현재의 시계열에 영향을 미치는 자기회귀 차수(p), 과거 오차항에 대한 영향을 받아 설명되는 이동평균 차수(q)를 정한다. 보이스피싱은 대부분 조직적인 역할 분담하에 집단이 수행하는 조직범죄

의 형태를 취한다.[4] 보이스피싱의 발생건수는 조직의 과거의 범행 횟수와 경험을 기반으로 과거의 경향성을 따를 것으로 예상된다. 즉, ARIMA 모형은 과거 발생건수에 비추어 미래의 발생건수를 예측하는 데 유용하다.

3.1.2 SARIMA(p, d, q)(P, D, Q, s)

SARIMA(Seasonal AutoRegressive Integrated Moving Average) 모형은 ARIMA 모형의 확장 버전으로, 계절적 패턴이나 주기성을 가진 데이터에 특히 적합하다.[7] SARIMA(p,d,q)(P,D,Q,s) 모형의 수식은 다음과 같다.

$$(1-\sum_{i=1}^{p}\Phi_i B^i)(1-\sum_{i=1}^{P}\Phi_i^* B^{is})(1-B)^d(1-B^s)^D Y_t$$
$$=(1+\sum_{j=1}^{q}\Theta_j B^j)(1+\sum_{j=1}^{Q}\Theta_j^* B^{js})\epsilon_t \quad (2)$$

여기서 Φ_i^*,Θ_j^*는 각각 계절적 자기회귀계수, 계절적 이동평균계수이며, D는 계절적 차분의 차수, s는 계절의 주기를 의미한다. 계절적 비정상 시계열인 경우 계절적 차분(seasonal difference)을 통하여 추세가 존재하는 계절적 비정상 시계열을 정상 시계열로 안정화한다. 이전 연구[5]에 따르면, 화폐의 이동에 따른 계절적 변동이 존재하며, SARIMA 모형은 이러한 계절성을 고려하여 보다 정확한 예측을 가능하게 한다.

3.1.3 SARIMAX(p, d, q)(P, D, Q, s)

SARIMAX(Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) 모형은 SARIMA 모형에 외생변수(exogenous variables)를 추가한 버전이다.[8] 다음은 본 연구에서 고려하고 있는 SARIMAX(p,d,q)(P,D,Q,s) 모형의 수식이다.

$$(1-\sum_{i=1}^{p}\Phi_i B^i)(1-\sum_{i=1}^{P}\Phi_i^* B^{is})(1-B)^d(1-B^s)^D Y_t$$
$$=(1+\sum_{j=1}^{q}\Theta_j B^j)(1+\sum_{j=1}^{Q}\Theta_j^* B^{js})\epsilon_t \quad (3)$$
$$+\beta_1 X_{1t}+\beta_2 X_{2t}+\beta_3 X_{3t}$$

여기서 β_1,β_2,β_3는 외생변수 계수이며, X_{1t},X_{2t},X_{3t}는 시간 t에서의 외생변수 값이다. SARIMAX는 계절성을 반영하는 SARIMA 모형의 기능에 더해, 본 연구에서는 피해액, 검거건수, 검거인원과 같은 외생변수를 추가함으로써 예측력을 향상시킬 수 있다.

3.2 분석 자료

보이스피싱 예측 분석을 위한 시계열 데이터로서, '발생건수', '피해액', '검거건수', '검거인원'을 고려해 볼 수 있다. 또한 사이버 범죄 예방교육 및 홍보 관련 데이터, 전담 인력 확대 및 국제 공조사례 등도 보이스피싱 발생 건수에 영향을 미칠 것이다. 본 연구에서는 먼저 피해액, 검거건수, 검거인원을 고려한 분석을 수행하고자 한다. 다른 관련 데이터는 향후 연구에서 고려해 볼 것이다.

<표1> 분석 데이터 일부

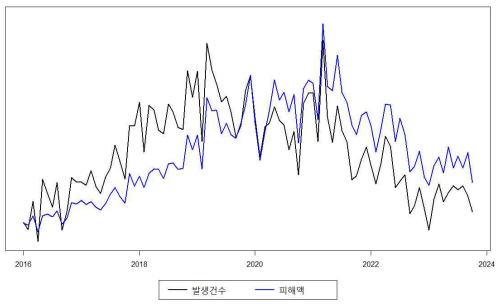
연도	월	발생건수	피해액	검거건수	검거인원
2016	1	1147	96	407	612
2016	2	1029	88	535	772
2016	3	1480	126	749	977

<표2> 기초통계량(n = 94)

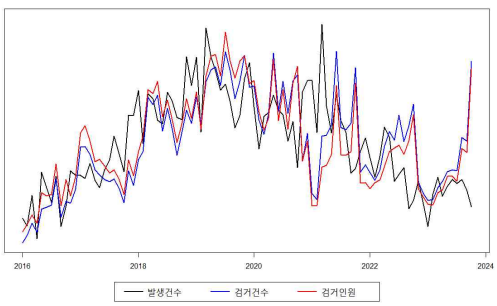
	발생건수	피해액	검거건수	검거인원
최솟값	842	58	407	612
1분위수	1723	254	1515	1731
중앙값	2271	390	2212	2331
평균	2265	404	2179	2528
3분위수	2785	555	2275	3378
최댓값	4017	952	3797	4900
표준편차	712	194	850	1030
분산	506695	37611	721983	1060365

데이터는 2016년부터 2023년 10월까지의 월별 데이터로, 정보공개 포털을 통해 경찰청에서 제공받은 데이터이다. 분석의 주요 대상인 발생건수는 보이스피싱 범죄의 발생 횟수를 나타내며, 나머지 피해액, 검거건수, 검거인원은 외생변수로 사용된다. 피해액은 보이스피싱으로 인한 피해 금액을 억 원 단위로 나타내며, 검거건수와 검거인원은 보이스피싱 범죄에 대한 검거된 사건 수와 관련 인원수를 의미한다. 제공된 데이터는 결측치가 없는 완전한 데이터셋으로, <표1>은 분석

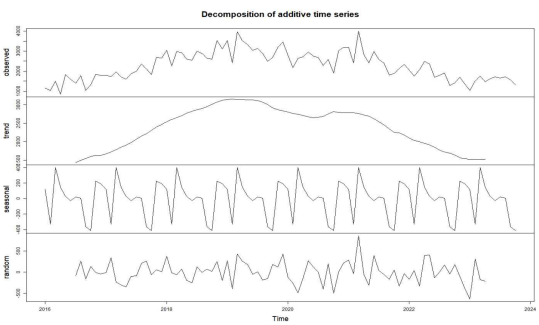
데이터의 일부이다. <표2>는 각 변수에 대한 기초통계량이며 반올림하였다.



(그림 1) 발생건수, 피해액 시계열 그래프(표준화)



(그림 2) 발생건수, 검거건수, 검거인원
시계열 그래프(표준화)



(그림 3) 발생건수 추세 및 계절성 분해

(그림 1)의 그래프를 보면 2019년과 2020년의 보이스피싱 발생건수와 피해금액은 최고점을 기록한 후 감소하는 추세를 보인다. 코로나 시기 이후로 발생건수 대비 피해액이 증가하는 것으로 보아, 건당 피해 규모가 확대되고 있다고 할 수 있다. (그림 2)에서는 2021년의 검거건수와 검거인원은 줄어든 반면, 발생건수는 증가하였다. 이러한 현상은 코로나19로 인해 증가한 비대면 거래와 관련성이 있다.[4] 최근의 검거건수 및 검거인원의 급증은 정부의 보이스피싱 대응 TF팀을 통한 강력한 단속 및 신고 창구 일원화로 인한 결과로 보인다.[9] 데이터 분해는 추세와 계절성을 이해하는 데 중요한 과정이다. (그림 3)은 보이스피싱 발생건수 데이터를 추세와 계절성으로 분해한 결과이다. 이 그래프에서는 발생건수가 증가한 후 감소하는 명확한 추세를 보인다. 또한, 계절성 분석에서는 데이터가 일정한 패턴을 따르고 있음이 나타나, 이는 추세뿐만 아니라 계절성도 데이터에 중요한 요소임을 나타낸다.

4. 분석결과

4.1 정상성과 자기 상관성

ARIMA 기반의 모형을 적합하기 전, 해당 시계열 데이터의 정상성 여부를 검증하기 위해 ADF(Augmented Dickey-Fuller) 검정을 시행한다. 정상성은 시계열 분석에서 필수적인 조건으로, 이를 충족시키지 않는 데이터는 차분, 로그 변환 등을 통해 정상 시계열로 변환될 필요가 있다.

4.1.1 정상성 검정

<표 3> ADF 검정

ADF	발생건수	피해액	검거건수	검거인원
통계량	-1.697	-1.773	-2.775	-2.668
P-value	0.433	0.394	0.062	0.08

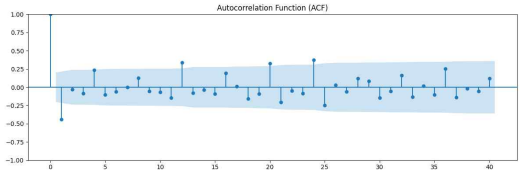
ADF 검정의 귀무가설은 시계열에 단위근(unit root)이 존재한다는 것이며, 이는 해당 시계열이 비정상(non-stationarity)이라는 것을 의미한다. ADF 검정 결과에 따르면, 보이스피싱 발생건수와 모든 외생변수에서 비정상성이 관찰되었다.

4.1.2 일반 차분한 시계열의 정상성 검정

<표 4> 일반 차분한 시계열 데이터 대한 ADF 검정

ADF	발생건수	피해액	검거건수	검거인원
통계량	-5.383	-4.872	-4.91	-7.235
P-value	0.000	0.000	0.000	0.000

일반 차분을 통하여 모든 변수의 추세에 대해 정상성을 확보하였다. ACF(AutoCorrelation Function) 그래프를 통해 다양한 시차(lag)에서의 시계열 데이터 간 상관관계를 시각적으로 분석하였다. 분석 결과, 대부분의 시차에 대한 상관계수가 신뢰구간 내에 위치하는 것이 관찰되었다. 이는 데이터가 1번 차분(d = 1)을 통해 정상성을 획득했음을 나타낸다. ACF 그래프에서 시차 12, 20, 24에서만 신뢰구간을 벗어나는 상관계수가 나타나, 이들 시차에서의 상관관계가 통계적으로 유의미함을 보여준다. 그러나 유의미한 상관관계가 4의 배수마다 반복적으로 나타나는 패턴을 보임에 따라, 계절성 주기(seasonality, s)를 4로 결정하였다.



(그림 4) 1번 차분한 데이터의 ACF 그래프

4.1.3 계절성 차분 데이터의 정상성 검정

이전 단계에서 나타난 계절성이 있는 데이터의 정상성 변환을 위하여, lag = 4를 기준으로 데이터에 계절 차분(seasonal difference)을 적용하였다.

<표 5> 계절성에 대한 ADF 검정

ADF	발생건수	피해액	검거건수	검거인원
통계량	-5.383	-4.872	-4.91	-7.235
P-value	0.000	0.000	0.000	0.000

<표 5>에서는 계절 차분 후의 시계열 데이터에 대한 정상성 검정 결과를 보여준다. 여기서 귀무가설이 강력히 기각되므로, 데이터가 정상성을 갖는다는 것을 확인할 수 있다. 따라서 모형 설정 시, 일반 차분의 차수(d)는 1, 계절 차분의 차수(D)는 1로 설정한다.

4.2 모수 추정 결과

이전의 과정을 통해 정해진 차수와 Python의 pmdarima 라이브러리의 'auto.arima' 함수를 사용하여 모형의 최적 차수를 결정하였다. 이 과정에서 계절성을 고려하지 않은 일반 차분(d)의 값을 1로 고정한 결과, ARIMA(0,1,1) 모형으로 차수를 결정하였다. 계절성을 고려한 모형에 대해서는 d=1, D=1, 계절성 주기(s)를 4로 설정한 상태로 진행하였다. 이를 통해 SARIMA(2,1,1)(0,1,1,4) 모형이 계절성을 고려한 상황에서 최적의 모형으로 확인되었다. 최종 모형 선택을 위하여 <표6>에서와 같이 AIC, BIC, RMSE, MAE, MAPE를 성능 지표로써 활용하였다.

AIC(Akaike Information Criterion)은 모형의 복잡도와 설명력의 균형으로 값이 작을수록 더 좋은 모형이라고 판단할 수 있다. BIC(Bayesian Information Criterion)은 AIC에 비해 더 많은 데이터의 수에 대해 페널티를 부여하며 마찬가지로 값이 작을수록 좋다.

RMSE(Root Mean Square Error)는 오차 제곱의 평균에 제곱근을 취한 것으로, 큰 오차에 더 많은 가중치를 부여하며, MAE(Mean Absolute Error)는 예측값과 실제값 사이의 절대 오차의 평균으로, RMSE에 비하여 이상치에 덜 민감하다. MAPE(Mean Absolute Percentage Error)는 각 오차를 실제 값에 대한 백분율로 계산해 오차의 상대적 크기를 이해할 수 있다.

Ljung-Box 잔차 검정은 시계열 데이터의 잔차에 자기상관이 있는지를 검정하는 통계적 방법으로, 이 검정의 귀무가설(H0)은 자기상관이 없다는 가설이다.

다양한 적합 지표들을 비교한 결과, ARIMA(0,1,1) 모형은 AIC, BIC, RMSE, MAE에서 상대적으로 높은 값을 보였다. 계절성을 고려한 SARIMA 모형은 ARIMA 모형에 비해 적합도가 개선되었다. 외생변수를 포함하는 SARIMAX(2,1,1)(0,1,1,4) 모형은 모형 적합도에서 눈에 띄는 향상을 보였다.

Ljung-Box 잔차 검정 결과에 따르면, SARIMAX 모형은 0.05의 유의수준에서 높은 p-value를 나타내어 다른 모형들보다 우수한 적합성을 보였다. 피해액을 주요 외생변수로 선정한 이유는 이 변수를 제외하면 모형 적합도 지표가 상대적으로 낮아지는 경향이 있기 때문이다. 이는 피해액이 보이스피싱 범죄와 관련하여

<표 6> 모형 적합 지표

모형	외생변수	AIC	BIC	RMSE	MAE	MAPE	Ljung-Box P-value
ARIMA	-	1389.30	1394.360	429.1	337.4	16.2%	0.83
SARIMA	-	1327.56	1340.000	416.15	332.7	16.83%	0.77
SARIMAX	피해액, 검거건수	1216.13	**1233.549**	233.2	183.3	9.8%	**0.98**
	피해액, 검거인원	1216.67	1234.088	233.28	183.63	9.78%	0.91
	모든 외생변수	**1214.28**	1234.188	**226.24**	**179.67**	**9.51%**	0.97

중요한 외생변수로 작용하고 있음을 나타낸다. 모든 외생변수를 포함한 모형과 피해액 및 검거건수만을 포함한 모형이 뛰어난 적합도를 보였다.

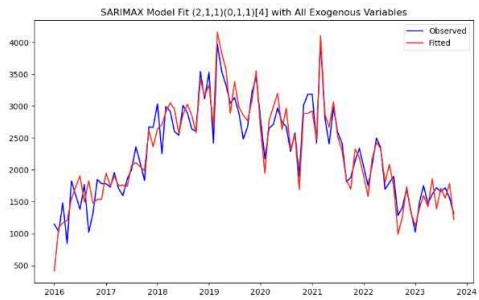
<표 7> 모든 외생변수를 포함한 모형 적합

	coef	std err	P-value
피해액	4.0446	0.305	0.000
검거건수	-0.3875	0.194	0.045
검거인원	0.3005	0.154	0.050
ar.L.1	-1.1354	0.337	0.001
ar.L.2	-0.4545	0.172	0.008
ma.L.1	0.5967	0.371	0.108
ma,S,L4	-0.7906	0.118	0.000

<표 8> 검거인원을 제외한 모형 적합

	coef	std err	P-value
피해액	0.8952	0.289	0.000
검거건수	-0.0340	0.048	0.478
ar.L.1	-1.1139	0.382	0.004
ar.L.2	-0.4423	0.189	0.019
ma.L.1	0.5778	0.421	0.170
ma,S,L4	0.7634	0.117	0.000

<표7, 표8>에서 두 모형을 비교하기 위해 각 계수의 유의성을 검토한 결과, 피해액과 검거건수는 통계적으로 유의미하지만, 검거인원은 유의성 면에서 경계선상에 위치하는 것으로 나타났다. 이에 따라, 피해액과 검거건수만을 포함한 모형을 살펴보았다. 검거건수는 분석에 있어서 중요한 외생변수로 작용하지만, <표 8>에 따르면 회귀계수의 유의성 검정 결과는 이와 일치하지 않는 것으로 보인다. 이러한 결과를 바탕으로, 최종적으로 선택된 적합 모형은 모든 외생변수를 고려한 SARIMAX (2,1,1)(0,1,1,4) 모형이다.



(그림 5) 최종 선택된 모형 적합

4.3 예측 분석

4.3.1 일단계예측(one-step ahead forecast)

최종적으로 선택된 모든 외생변수를 포함한 모형을 사용하여 세 기간에 대한 예측을 수행하고, 이를 실제 데이터와 비교하여 모형의 예측 성능을 평가하였다.

<표 9> 예측기간별 성능 지표

	m	RMSE	MAE	MAPE
21-01~ 23-10	34	204.6700	167.7990	8.7816
22-01~ 23-10	22	193.9625	156.7414	9.5356
23-01~ 23-10	10	196.6336	166.9178	10.8040

2021년 1월부터 2023년 10월까지 테스트 데이터의 수가 34개인 성능지표를 확인하여 보면, RMSE와 MAE 값은 가장 높지만, MAPE 값이 낮은 것으로 보인다. 이는 MAPE 수식의 분모에 있는 실제 관측값

때문에 높은 값을 가질수록, 동일한 절대 오차가 발생하더라도 MAPE 값이 작아진다는 특성 때문이다.

4.3.2 다단계예측(multi-step ahead forecast)

다단계예측(multi-step ahead forecast) 분석은 시계열 데이터를 기반으로 미래의 여러 시점에 대한 예측을 수행하는 분석 기법이다. 테스트 데이터의 수가 22개일 때의 one-, two-, three-step 예측, 즉 현재 시점으로부터 1, 2, 3 시점 이후의 데이터를 예측하였다.

<표 10> 예측 시점별 오차지표(m = 22)

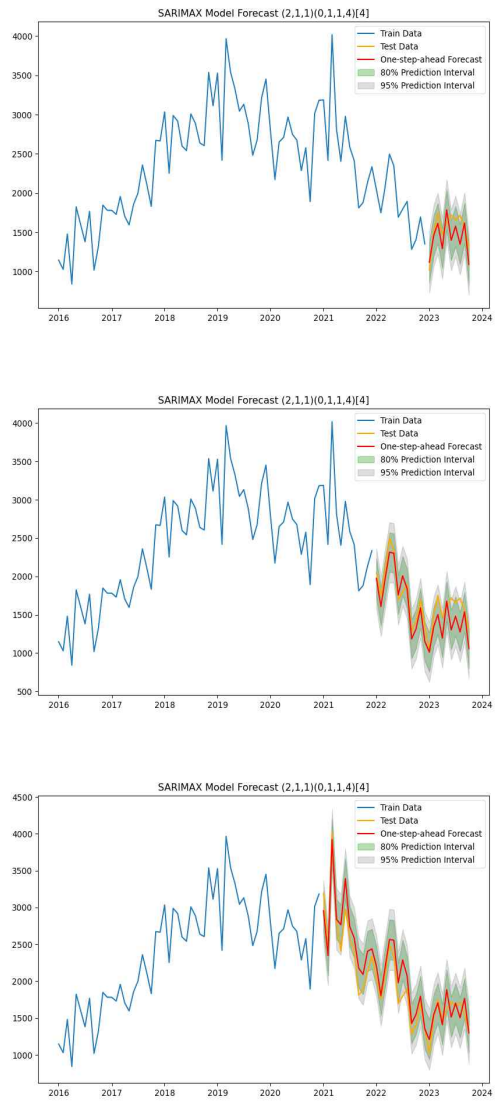
	RMSE	MAE	MAPE
one-step	193.9625	156.7414	9.5356
two-step	261.5866	211.0500	12.9149
three-step	212.1120	174.6446	10.4351

2022년 1월부터 2023년 10월까지의 기간 동안 수행된 예측 시점별 결과를 비교해 보았을 때, one-step 예측이 RMSE, MAE, MAPE 모든 면에서 가장 낮은 오차를 보였다. 이는 one-step 예측이 각 시점의 데이터를 가장 정확하게 예측하는 경향이 있음을 의미한다. two-step 예측오차가 three-step 예측오차보다 커진 이유는 (그림 1)의 발생건수 그래프에서 보듯이 최근 22개 원데이터의 패턴이 지그재그 형태로 증가감소가 반복되는 과정에서 two-step의 더 큰 오차를 발생시킨 것으로 파악된다.

4.3.3 예측구간 및 경험적 포함확률(Prediction interval and empirical coverage)

(그림 6)은 m=10, 22, 34일 때의 80%, 95% 예측구간을 나타낸 그래프이다. 예측구간의 성능을 확인하기 위하여 경험적 포함확률(empirical coverage)을 구하였다. 경험적 포함확률이란 관찰된 데이터가 특정 예측구간 내에 포함되는 비율을 나타내는 지표이다. <표 11>은 세 가지 예측기간 동안 80%, 95% 예측구간에 대한 경험적 포함확률을 정리한 표이다.

전반적으로 예측구간은 높은 신뢰도를 보여준다. 각 기간에 대한 경험적 포함확률을 비교해 보면, 2021년 1월부터 2023년 10월까지의 기간에서는 80% 예측구간과 95% 예측구간 모두 적절한 포함확률을 보인다. 80%, 90% 예측구간 모두 테스트 데이터의 수가 커질 때, 경험적 포함확률은 이론상의 포함확률(nominal coverage probability)에 근사한 값임을 확인할 수 있다. 즉, 2021년 1월부터 2023년 10월까지의 기간에서의 경험적 포함확률은 각각 0.8529, 0.9706으로서 다른 기간에서보다 80%, 95%에 더 가까운 값이며, 이는 예측모형 및 예측구간 분석의 신뢰성을 보여준다.



(그림 6) 예측값 및 80% 90% 예측구간
m=10 (상), m=22 (중), m=34 (하)

<표 11> 예측기간별 경험적 포함확률

	예측 구간	경험적 포함확률	예측 구간 내 데이터 수	테스트 데이터 수
21-01~ 23-10	80%	0.8529	29	34
	95%	0.9706	33	34
22-01~ 23-10	80%	0.8636	19	22
	95%	0.9090	20	22
23-01~ 23-10	80%	0.8	8	10
	95%	1.0	10	10

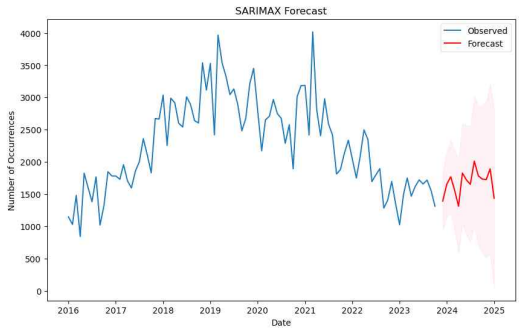
<표 12>는 각 예측 시점별 경험적 포함확률을 분석한 결과를 나타낸 표이다. 예측기간은 2022년 1월부터 2023년 10월까지이며, 이 기간에 대한 테스트 데이터의 수는 총 22개이다. one-step 예측의 경우, 80% 예측구간에서 높은 포함확률을 보이지만, 95% 신뢰구간에서는 다른 기간의 예측에 비해 상대적으로 낮은 포함확률을 보인다. two-step 예측은 95% 예측구간에서 포함확률이 one-step과 동일하지만, 80% 예측구간에서는 포함확률이 더 낮다. three-step 예측에서는 두 예측구간 모두에서 높은 포함확률을 보인다.

<표 12> 예측 시점별 경험적 포함확률

	예측 구간	경험적 포함확률	예측 구간 내 데이터 수	테스트 데이터 수
one-step	80%	0.8636	19	22
	95%	0.9090	20	22
two-step	80%	0.7273	16	22
	95%	0.9090	20	22
three-step	80%	0.9090	20	22
	95%	0.9545	21	22

4.3.4 최종 적합 모형 예측 결과

(그림 7)은 SARIMAX(2,1,1)(1,1,0,4) 모형을 사용하여 2024년 12월까지의 예측 결과를 시각화한 것이다. 이는 Multi-step 예측 방법을 적용한 것으로, 예측기간이 증가함에 따라 결과가 점점 특정 값으로 수렴하는 경향을 보일 것이다. <표 13>은 이에 대한 결과로, 예측값과 예측구간의 하한값, 상한값을 나타낸 것이다.



(그림 7) 2024년 12월까지의 예측

<표 13> 2024년 12월까지의 예측 발생건수

	하한값	예측 발생건수	상한값
2023-11	938	1391	1845
2023-12	1153	1653	2154
2024-01	1204	1769	2334
2024-02	952	1565	2179
2024-03	592	1312	2033
2024-04	1044	1826	2608
2024-05	874	1720	2565
2024-06	752	1653	2554
2024-07	1005	2012	3019
2024-08	708	1783	2858
2024-09	590	1734	2878
2024-10	520	1727	2934
2024-11	577	1891	3206
2024-12	47	1436	2825

5. 결 론

본 연구는 보이스피싱 범죄의 발생건수에 대하여 ARIMA, SARIMA, SARIMAX 모형을 비교 분석 후 예측하였다. 분석에 사용된 데이터는 2016년부터 2023년 10월까지의 데이터이다. 보이스피싱 발생건수 데이터의 예측 성능이 가장 우수한 모형으로서, 계절성과 외생변수를 고려한 SARIMAX(2,1,1)(1,1,0,4) 모형이 선정되었다. 이 모형의 외생변수는 피해액, 검거건수, 검거인원 모두를 포함한다. 표본 외 예측분석을 수행하여 훈련 데이터, 테스트 데이터로 나누어 예측값의 예측 성능을 검증하였다. 예측기간을 다양하게 설정하였고, 다단계예측 기법을 사용하여 예측 시점별 오차를 비교하여 성능을 확인하였다. 예측구간을 추정하고

이것의 경험적 포함확률을 도출함으로써 예측 모형의 우수성을 확인하였다. 2024년 12월까지의 보이스피싱 월별 발생 건수를 예측값 및 상한값, 하한값을 제시하였다.

본 연구의 차별성 및 우수성은 시계열 분석에서 보이스피싱에 영향을 주는 데이터, 즉, 피해액, 검거건수, 검거인원을 외생변수로 채택하여 분석하여 예측 성능을 향상시켰다는 점이다. 이를 입증하기 위하여 기존 연구에서 제시된 동일한 기간의 데이터로 위의 세 외생변수를 포함하고 계절성 주기가 4인 SARIMAX 모형으로 분석한 결과, 예측오차가 놀라울 만큼 개선되었다. [5]의 논문에서의 기간과 동일하게 2018년 1월부터 2021년 12월까지를 훈련 데이터로, 2022년 1월부터 12월까지 테스트 데이터로 사용하여 표본 외 예측분석을 진행하였을 때, SARIMAX 모형의 예측 결과의 RMSE 값은 113.0811, MAE 값은 78.5784, MAPE 값은 4.4453으로 계산되었다. 이는 동일한 기간의 데이터를 ARIMA 모형에 적용한 기존 연구[5]의 예측 성능 결과 대비, 예측오차의 높은 향상력을 보여준 것이다.

앞서 언급한 바와 같이 분석에 사용된 외생변수 데이터 외에, 관련 데이터를 추가하여 향후 연구를 고려해 볼 수 있다. 보이스피싱은 사회 경험이 부족한 사회 초년생과 인지 능력이 상대적으로 떨어지는 노인층의 금융 피해가 큰 경향이 있다. 이들을 대상으로 하는 사이버 범죄 예방 교육이나 홍보 활동은 보이스피싱 발생건수 감소에 영향을 줄 것이다.[10] 또한 보이스피싱 검거인원 및 검거건수를 증가시키는 주요 요소에는 보이스피싱 전담 수사 기관의 설립, 전담 인력의 확대, 제도적 개선 및 국제적 공조 강화를 들 수 있다. 이와 관련된 데이터로 보이스피싱 수사팀의 적정 인력 모형[11]을 활용한 시계열 분석은 후속 연구로서의 도전적인 연구가 될 것이며, 이는 더욱 향상된 예측 성능을 제공할 수 있으리라 기대한다.

보이스피싱은 단지 금전적 손실을 넘어서 피해자에게 심각한 정신적 고통을 주고, 이로 인해 사회적 안정성이 저하되어 국가적 차원에서도 큰 손실이 발생할 수 있다.[12] 본 논문에서 진행된 시계열 예측 분석 결과는 정부와 금융 기관이 보이스피싱 대응 전략 수립에 도움을 줄 것으로 예상한다. 본 연구와 더불어, 보이스피싱 예측력 향상을 위한 연구는 더욱 활발하게 수행되리라 기대하며, 전문가들의 열정 있는 연구는 보이스피싱 예방 정책에 관한 예산 및 인원 투입에 합리적인 근거를 제시할 수 있으므로 기대효과가 크다고 할 수 있다.

참고문헌

[1] 조호대, “보이스피싱 발생 및 대응방안”, 한국콘텐츠학회 논문지, 제12권, 제7호, pp. 176-182, 2012.

[2] https://www.etoday.co.kr/news/view/2273041 (검색일: 2023.01.18)

[3] 김도윤, “한·중 전기통신금융사기범죄 및 관련 제도의 현황과 시사점”, 中國法硏究(Chinese Law Review), 제41권, pp. 159-182, 2020.

[4] 곽영암, “코로나 사태에 따른 비대면 생활의 증가와 보이스피싱 사례연구”, 한국경영컨설팅연구, 제22권, 제5호, pp. 171-184, 2022.

[5] 추정호, 주용휘, 엄정호, “ARIMA 모형을 이용한 보이스피싱 발생 추이 예측”, 융합보안논문지, 제22권, 제3호, pp. 79-86, 2022.

[6] 김재문, 장성호, 김성수, “시계열 모형과 기계학습 모형을 이용한 풍력 발전량 예측 연구”, 응용통계연구, 제34권, 제5호, pp. 723-734, 2021.

[7] 김재호, 김장영, “SARIMA모형을 이용한 코로나19 확진자수 예측”, 한국정보통신학회논문지, 제26권, 제1호, pp. 58-63, 2022.

[8] 이근철, 이희정, 구훈영, “SARIMAX 모형을 이용한 부산항 컨테이너 물동량 예측”, 한국경영과학회, 제40권, 제2호, pp. 1-13, 2023.

[9] https://www.korea.kr/news/policyNewsView.do?newsId=148920883 (검색일: 2023.12.08.).

[10] https://www.fss.or.kr/fss/bbs/B0000188/view.do?nttId=134451&menuNo=200218 (금융감독원 보도자료, 2024.03.08.).

[11] 정웅, “보이스피싱 범죄추세와 수사 대응체제의 발전방향”, 한국공안행정학회보, 제29권, 제4호. pp.461-484, 2020.

[12] 김경진, 서준배, “보이스피싱 현황과 정책제언”. 시큐리티연구. 제66호, pp. 111-128, 2021.

〔저자소개〕

강 다 연 (Da-yeon Kang)
2024년 8월 가천대학교 응용통계학과 학사
email : kkddyy0830@naver.com

이 승 연 (Seung-yeon Lee)
2020년 3월 ~ 현재 가천대학교 응용통계학과 학사 과정
email : leesyun122@naver.com

황 은 주 (Eunju Hwang)
2002년 5월 미국 인디애나대학교 이학박사
2014년 3월 ~ 현재 가천대학교 응용통계학과 조교수, 부교수
email : ehwang@gachon.ac.kr