IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

Running Mean Smoothing Technique of Extreme Exploratory Data Analysis with Application Using R Software

Barjess H. Mohammed[1], Qasim N. Husain [2]

[1,2] Mathematics Departments, Education for Pure Science College, Tikrit University, 34001 Baghdad, Iraq.

ARTICLE INFO

Article history:
Received 13 October 2025
Revised 13 October 2025
Accepted 13 January 2026
Available online 13 January 2026

Keywords:
Exploratory data analysis
Running means
Smoothing
Extreme data

ABSTRACT

Exploratory Data Analysis (EDA) is a pivotal stage in statistical modelling and data analysis, comprising methods that support the identification of unknown patterns in data observations. Among the techniques employed, running mean smoothing, that sometimes known as the moving averages smoother, is openly applied to minimize the passing variations and highlight long-term trends. This paper clarifies the application of the running means smoothing method using the R software. The present way requires replacing each data observation with the average of its neighbourhood within a known span, thereby detecting errors and outliers and enhancing the ability to interpret. The analysis shows how R opens the door to elastic and efficient achievement of smoothing for both contrived and real data. The outcomes proved that running mean as a technique of smoothing is effective in lighting invisible structures, detecting outliers, and providing insights into temporal dynamics, making it an important instrument in EDA.

1. Introduction

Exploratory Data Analysis (EDA) involves revealing patterns, detecting outliers and extremes, examining hypotheses, and summarizing data preparatory to data analysis. Smoothing techniques like running means, running medians, play a critical part in the procedure of EDA. In particular, the running means technique provides a modified view of the raw data without losing the original patterns. This technique offers an advantage in avoiding anomalous observations, which often indicate noise or missing values in the data.

Running means are an important smoothing way in exploratory data analysis, particularly for identifying outliers. In this context, Husain et al. (2017) [1] discussed the concepts of Running means and running medians, linking them to the concept of extreme data and supporting these smoothing techniques in the field of data analysis in the agricultural sector, highlighting their ability to treat and smooth such data. Dawson (2011) [2] also discussed the importance of identifying outliers in boxplots and accurately interpreting their statistical significance. Schwertman et al. (2004) [3] then proposed an improvement to Tukey's method for detecting outliers to include more general cases. Walfish (2006) [4] introduced an overall review of outlier detection methods, supporting the reliability of using the moving average as a dynamic reference. Fitrianto et al. (2022) [5] then compared multiple methods for detecting outliers, including boxplots, highlighting their effectiveness and superiority in some contexts. Babura et al. (2018) [6] demonstrated the use of box plots in analyzing very extreme data, such as flood values, and illustrated their

Corresponding Author: qasim11@tu.edu.iq

https://doi.org/10.62933/r0e91705

analytical properties under severe conditions. Byrne et al. (2000) [7] presented the use of neural networks in medical decision support, reinforcing the importance of combining visual analysis with intelligent tools. Husain et al. (2017) [8] performed an exploratory study of extreme data using box plots detect significant differences.

This paper's goal is to clarify the running means procedure as an efficient data analysis tool for daily health data, focusing on detecting extreme values and outliers in such kind of dataset, and to smooth the data carves of various geographical areas in Salah al-Din Governorate, relying on a comprehensive temporal and analytical representation.

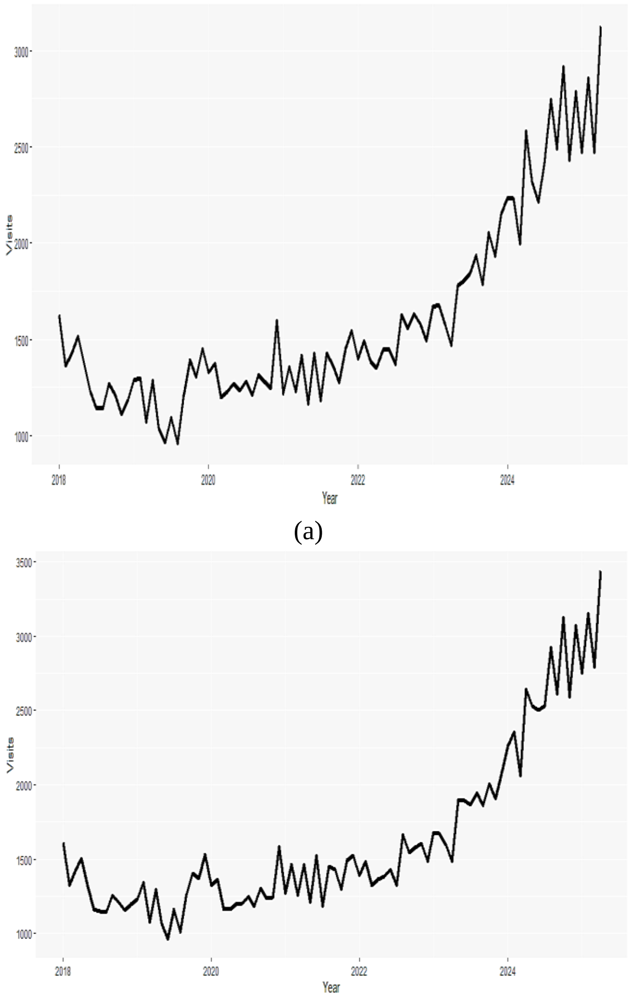
2. Data Analysis

The data used in this analysis are monthly data on the number of patients visiting herbal clinics in Salah al-Din Governorate for interval (January 2018-April 2025) with number of observations 88, which were obtained from the Directorate of Health in Salah al-Din Governorate. This data was divided or analyzed into two types: data on male and female visitors of different ages, as shown in Table 1.

Table 1. number of patients visiting herbal clinics in Salah al-Din Governorate

Female patients visitors												
Year	**January**	**February**	**March**	**Abril**	**May**	**June**	**July**	**August**	**September**	**October**	**November**	**December**
2018	**1628**	**1364**	**1422**	**1515**	**1384**	**1229**	**1140**	**1139**	**1268**	**1209**	**1112**	**1181**
2019	**1286**	**1301**	**1071**	**1290**	**1044**	**966**	**1093**	**957**	**1208**	**1396**	**1308**	**1453**
2020	**1327**	**1377**	**1202**	**1232**	**1270**	**1237**	**1283**	**1211**	**1318**	**1279**	**1246**	**1599**
2021	**1215**	**1359**	**1227**	**1419**	**1164**	**1431**	**1183**	**1426**	**1366**	**1276**	**1454**	**1546**
2022	**1399**	**1495**	**1390**	**1353**	**1448**	**1454**	**1367**	**1627**	**1555**	**1634**	**1583**	**1491**
2023	**1669**	**1682**	**1573**	**1469**	**1782**	**1803**	**1847**	**1940**	**1786**	**2054**	**1930**	**2147**
2024	**2236**	**2232**	**1999**	**2581**	**2322**	**2211**	**2416**	**2748**	**2487**	**2916**	**2430**	**2788**
2025	**2473**	**2858**	**2469**	**3130**								
Male patients visitors												
Year	**January**	**February**	**March**	**Abril**	**May**	**June**	**July**	**August**	**Septembe r**	**October**	**November**	**December**
2018	**1614**	**1325**	**1419**	**1501**	**1322**	**1160**	**1146**	**1143**	**1255**	**1210**	**1154**	**1198**
2019	**1230**	**1340**	**1072**	**1293**	**1065**	**957**	**1164**	**1004**	**1261**	**1403**	**1371**	**1529**
2020	**1321**	**1360**	**1158**	**1158**	**1195**	**1201**	**1247**	**1183**	**1305**	**1232**	**1240**	**1586**
2021	**1269**	**1461**	**1254**	**1462**	**1209**	**1521**	**1178**	**1453**	**1427**	**1294**	**1493**	**1525**
2022	**1390**	**1485**	**1322**	**1361**	**1384**	**1432**	**1319**	**1664**	**1544**	**1577**	**1607**	**1484**
2023	**1676**	**1672**	**1596**	**1480**	**1896**	**1891**	**1863**	**1948**	**1857**	**2005**	**1904**	**2081**
2024	**2262**	**2357**	**2059**	**2645**	**2529**	**2504**	**2528**	**2923**	**2611**	**3129**	**2593**	**3074**
2025	**2750**	**3155**	**2790**	**3445**								

Step 1: Enter the data by creating a data frame containing the months 2018–2025, then constructing the monthly time series with a frequency of 12 starting from 2018-01, and then plotting the raw series before analysis as in Figure 1.



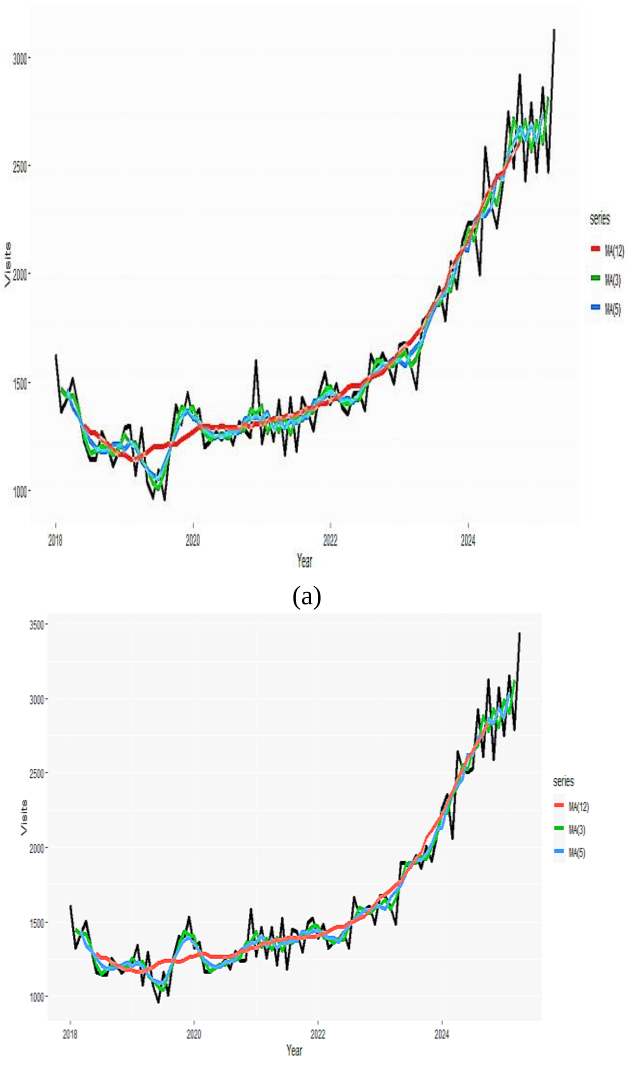
(a)

(b)

Figure 1. (a) Data series plot for female patient visitors, (b) Data series plot for male patient visitors

Figure (1a) shows that female visits were relatively stable between 2018 and 2020, with an average of approximately 1,200 visits per month, before beginning a gradual increase in 2021, followed by a clear jump starting in 2023, reaching more than 3,000 visits in early 2025. Figure (1b) shows a similar pattern for males, with a slight difference in monthly visit levels, but it also highlights the same strong upward trend in recent years.

Step 2: Optionally calculate the moving averages by selecting MA(3), MA(5), and MA(12), and then plotting the three moving averages over the series as shown in Figure 2.



(a)

(b)

Figure 2. (a) Analysis of monthly (Female) using moving averages over three different periods, (b) Analysis of monthly (Male) using moving averages over three different periods

In Figure (2), the results for females showed that the long-term average (MA12) reflects a steady upward trend without a clear break, while the shorter averages (MA(3) and MA(5)) reveal short-term seasonal fluctuations, but they all move in the same upward direction. This indicates that demand for these services has been steadily increasing since late 2022. The data for males showed almost the same pattern, with relative stability in the first years and then a sharp increase starting in 2023, confirming the presence of a structural change in the behavior of demand for these clinics.

This is detailed for females as follows: During the period 2018–2020, visits were low and fluctuated around 1,000–1,500 visits per month, with relative stability. During the period 2021–2022, there was a slight increase in visits, with a clear upward trend beginning in late 2022. From 2023 onward, visits increased sharply, exceeding 3,000 visits per month in 2025, with the steep climb continuing .The annual moving average (MA(12)) illustrates the overall long-term trend, confirming continuous growth without significant interruption. The shorter moving averages (MA(3)) and (MA(5)) reveal short-term fluctuations but move in the same direction as the overall upward trend. The overall trend has been strongly upward since 2022, potentially indicating increased demand for clinic services. Short-term moving averages are useful for detecting sudden or seasonal changes, while long-term moving averages confirm overall stability. The data demonstrate a typical pattern that can support time-based forecasting models such as ARIMA or hybrid models for future planning.

Step 3: Select the order of the MA(q) model. Orders from 1 to 15 were tested, and the AIC and BIC were extracted, a table was built, and then the best order was determined based on the AIC and BIC criteria [9-10-11], which is as shown in Table 2. While Moving Average (MA) Model Coefficients and Interpretations calculating in table 3. Then the training set error measures were calculated, which are shown in Table 4.

Table 2. The values of Information criteria (AIC & BIC)

	MA(q)	AIC	BIC
Female patient visitors	MA(1)	1298	1305
	MA(2)	1253	1263
	MA(3)	1237	1249
	MA(4)	1211	1226
	MA(5)	1204	1221
	MA(6)	1188	1208
	MA(7)	1192	1215
	MA(8)	1186	1211
	MA(9)	1180	1207
	MA(10)	1179	1208
	MA(11)	1178	1210
	MA(12)	1170	1205
	MA(13)	1173	1211
	MA(14)	1166	1206
	MA(15)	1160	1202
Male patient visitors			
	MA(1)	1320	1327
	MA(2)	1266	1275
	MA(3)	1250	1263
	MA(4)	1221	1236
	MA(5)	1215	1232
	MA(6)	1200	1220
	MA(7)	1201	1223
	MA(8)	1189	1214
	MA(9)	1189	1216
	MA(10)	1190	1219
	MA(11)	1185	1217
	MA(12)	1173	1213
	MA(13)	1181	1218
	MA(14)	1177	1217
	MA(15)	1179	1216

Table 3. Moving Average (MA) Model Coefficients and Interpretations

	Coefficient	Values	Standard error	Interpretation
Female patient visitors	MA(1)	0.3523	0.1889	Strong correlation with the previous one month
	MA(2)	0.8242	0.1501	Strong correlation with the previous two months
	MA(3)	0.6602	0.1803	Strong correlation with the previous three months
	MA(4)	0.8813	0.2680	Very strong correlation with the previous fourth month
	MA(5)	0.6657	0.1840	Medium correlation with the fifth month
	MA(6)	0.8100	0.1942	Strong correlation
	MA(7)	0.6703	0.2020	Medium correlation with the seventh month
	MA(8)	0.8789	0.2029	Very strong correlation
	MA(9)	0.7492	0.2620	Strong correlation with the ninth month
	MA(10)	0.9927	0.2239	The highest correlation, approaching 1, indicates a clear cyclical pattern
	MA(11)	0.6573	0.1907	Moderate correlation
	MA(12)	0.9153	0.2855	Strong correlation with the same month of the previous year (annual seasonality)
	MA(13)	0.1423	0.1760	Weak and insignificant correlation
	MA(14)	0.7363	0.1743	Strong correlation
	MA(15)	0.4336	0.1461	Moderate correlation
	Intercept	1686.3242	141.8705	
Male patient visitors	MA(1)	0.3034	0.1418	Moderate correlation with the previous month
	MA(2)	0.8012	0.1076	Strong correlation with the previous two months
	MA(3)	0.6979	0.1495	Strong correlation with the previous third month
	MA(4)	0.9556	0.1890	Very strong correlation with the fourth month
	MA(5)	0.6892	0.1588	Moderate correlation with the fifth month
	MA(6)	0.8356	0.1666	Strong correlation with the sixth month
	MA(7)	0.7284	0.1671	Moderate to strong correlation with the seventh month
	MA(8)	0.8737	0.1909	Strong correlation with the eighth month
	MA(9)	0.7539	0.1909	Strong correlation with the ninth month
	MA(10)	0.8729	0.1636	Strong correlation with the tenth month
	MA(11)	0.6036	0.1702	Medium correlation with the eleventh month
	MA(12)	1.0348	0.1830	Strong correlation with the twelfth month (annual seasonality)
	MA(13)	0.4030	0.1532	Relatively weak correlation with the thirteenth month
	MA(14)	0.7443	0.1823	Strong correlation with the fourteenth month
	MA(15)	0.4515	0.1287	Medium correlation with the fifteenth month
	Intercept	1741.7059	161.9034	

Table 4. Training set error measures

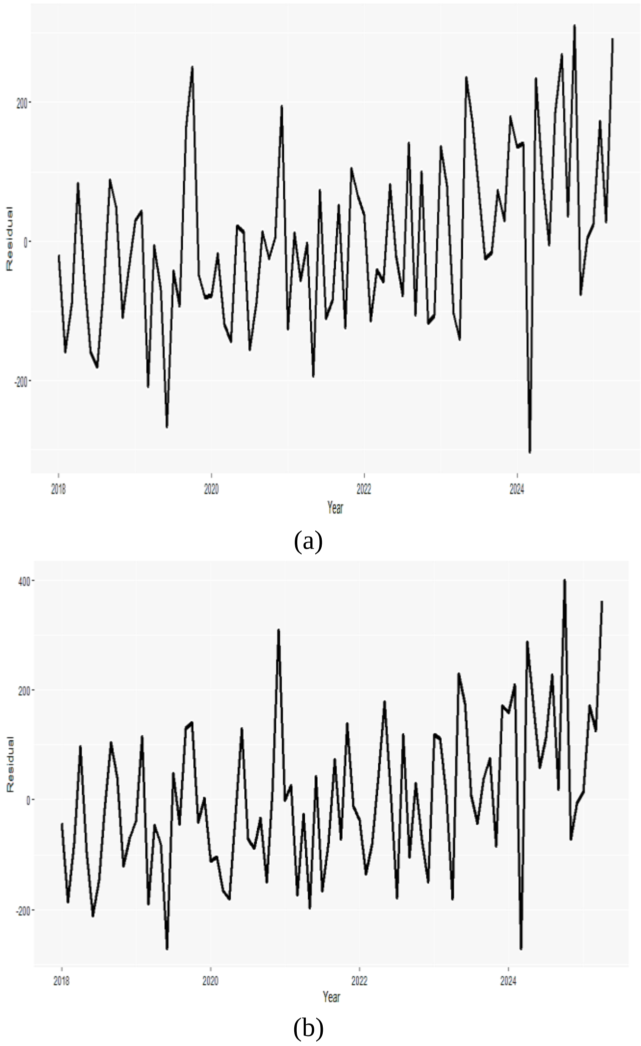
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Female	2.061817	125.7589	100.6358	-1.091244	6.538782	0.615531	0.05986081
Male	1.002975	154.4024	125.2294	-1.408556	7.815156	0.7425172	0.001067863

The AIC and BIC were calculated as shown in Table (2). The table shows that the AIC value gradually decreases with increasing model rank, with the MA(15) model recording the lowest value for both males and females, indicating it best fits the data.

Accordingly, the model coefficients were estimated as shown in Table (3), which displays the estimated values and standard deviations for each coefficient. For females, the high values of some coefficients, such as MA(10) and MA(12), indicate a strong correlation with the values in the corresponding months of the previous year, reflecting a clear annual seasonal pattern. The same applies to the male data, where the MA(12) coefficient appears to be approximately greater than 1, reinforcing the hypothesis of annual seasonality in visitation. The constant coefficient (intercept) in both models indicates a basic average monthly visitation of approximately 1,700 visits, which represents the baseline after removing the effect of temporal correlations.

Table (4), which displays the error indicators in the training set, shows that the average absolute error percentage (MAPE) [12] is relatively low, reaching 6.53% for females and 7.81% for males, indicating good accuracy of the model in prediction. Furthermore, the MAPE values are less than 1, which means that the model outperforms simple random models.

Step 4: Diagnose the residuals for the AIC model. The residuals were extracted from the (res_aic = residuals). Then, plot the residuals series as time and a ggplot of the residuals shown in Figure 3. Then, a histogram with density on the same scale using the (aes(y=..density..) + geom_density) instruction shown in Figure 4. Then, a QQ-plot of the residuals was drawn shown in Figure 5. Finally, a Ljung–Box test was performed up to lag 24 and the result was printed.



(a)

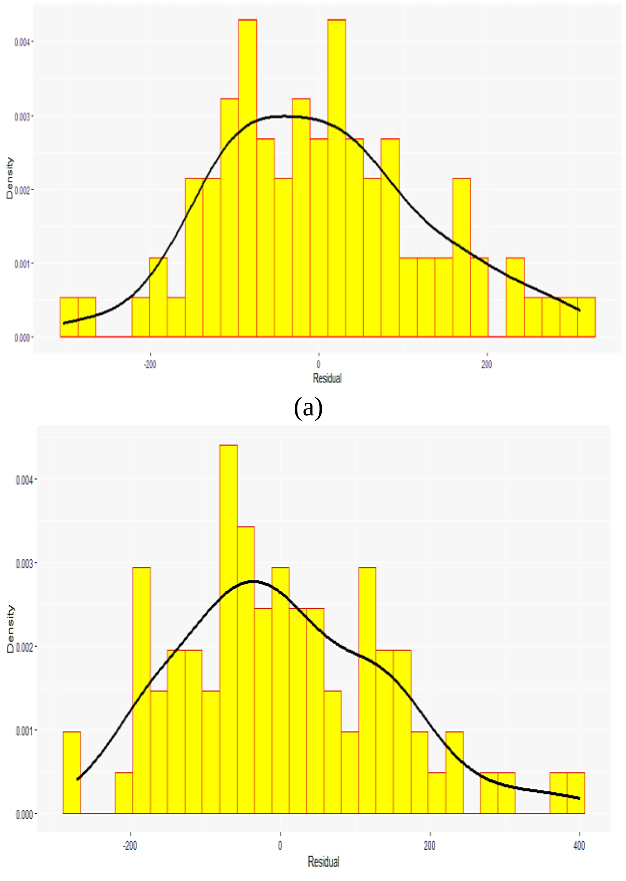
(b)

Figure 3. (a) ggplot of the residuals (Female), (b) ggplot of the residuals (Male)

Figure 3 (a) displays the residuals for the ARIMA model selected according to the AIC criterion.

- The residuals fluctuate between approximately -250 and +250, and the fluctuations are relatively constant between 2018 and 2022. Starting in mid-2023, the fluctuations increased to more than ±250.
- The distribution around zero: Most values are distributed around zero without a clear deviation. This indicates that the model does not have a clear bias in its predictions.
- The autocorrelation does not show a clear cyclical pattern in the graph, indicating that the residuals are quasi-random. This is consistent with the results of the Box-Ljung test, which showed a high p-value (0.2683), meaning there is no significant autocorrelation.
- The anomaly periods: Some outliers appear towards the end of 2024, indicating unexpected events in the data that the model did not capture well. These values could reflect sudden changes in visit volume.

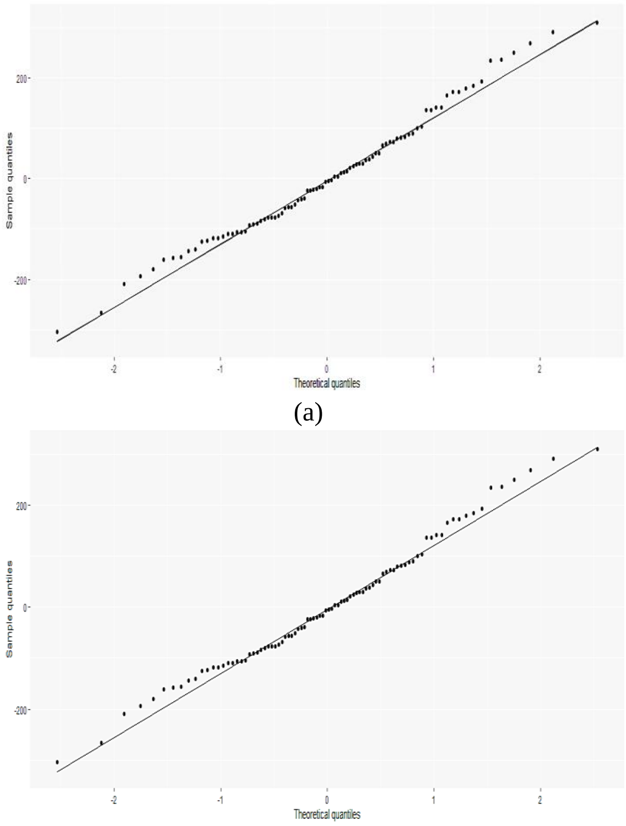
The model fits and represents the data well. However, the extreme values in recent periods call for consideration of updating the model or introducing additional explanatory variables. The lack of a clear pattern in the residuals means that the predictions generated by the model are largely reliable.



(a)

(b)

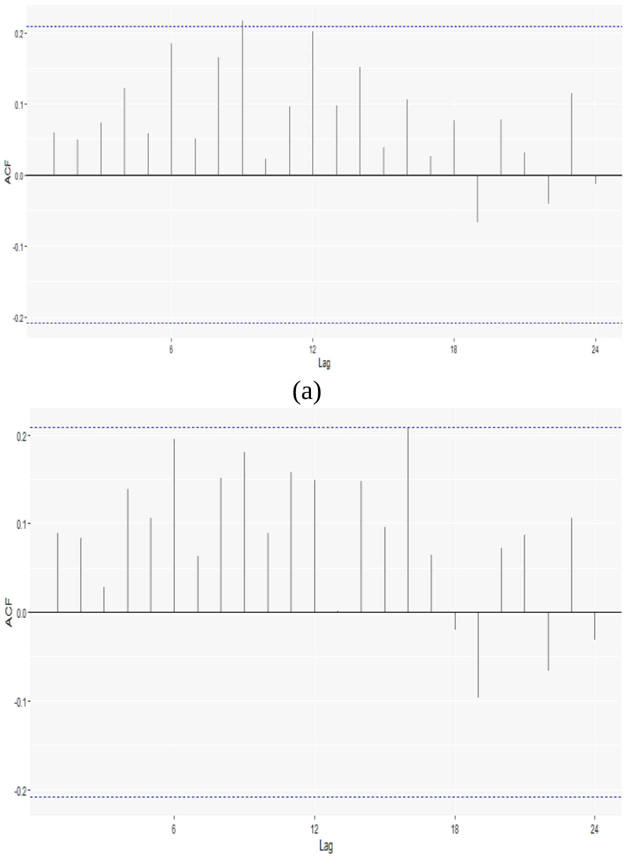
Figure 4. (a) histogram with density (Female), (b) histogram with density (Male)



(a)

(b)

Figure 5. (a) QQ-plot of the residuals (Female), (b) QQ-plot of the residuals (Male)



(a)

(b)

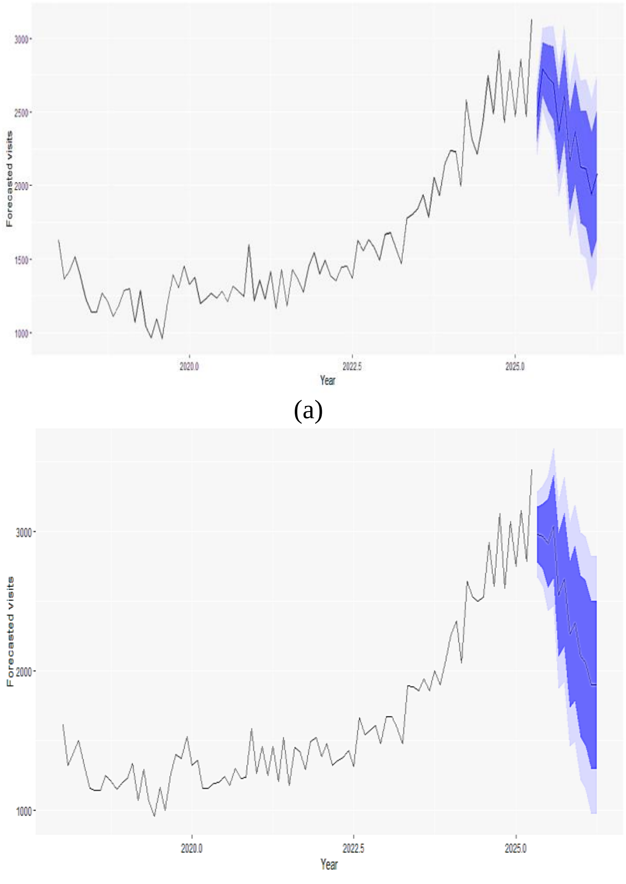
Figure 6. (a) ACF of the residuals (Female), (b) ACF of the residuals (Male)

Table 5. Box-Ljung test

data		Residuals for the AIC	
	X-squared	df	p-value
Female	27.806	24	0.2683
Male	32.798	24	0.1084

Examining the model residuals, as shown in Figure (3), reveals that the values fluctuate around zero without any clear deviation, indicating that the model is unbiased. The residuals for females and males mostly range between ±250, with some outliers at the end of 2024 that may reflect unexpected events not captured by the model. Figure (4) shows the probability distribution of the residuals combined with the density curve, which is close to normal. The QQ plot in Figure (5) confirms that the residuals are distributed approximately along the normal distribution line. The autocorrelation plots of the residuals in Figure (6) show the absence of clear correlation patterns, which is confirmed by the results of the Box–Ljung test in Table (5), where the probability values are high (0.2683 for females and 0.1084 for males), indicating the absence of residual autocorrelation.

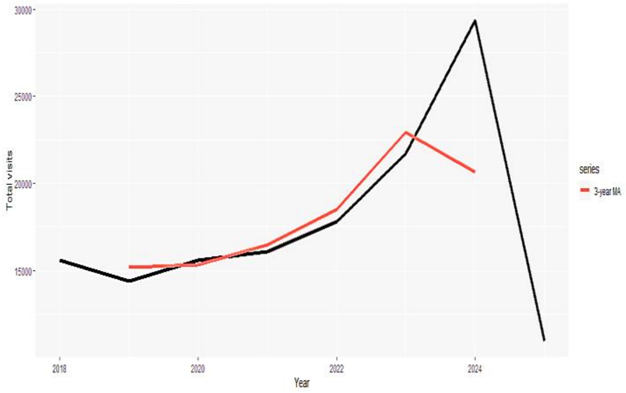
Step 5: Forecast for 12 months and plot it as shown in Figures 7 and 8 respectively.



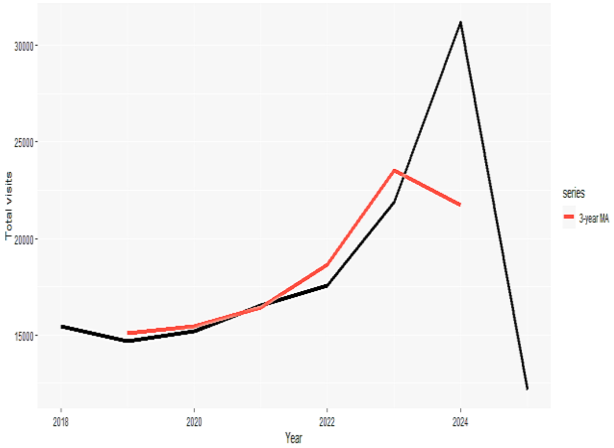
(a)

(b)

Figure 7. (a) 12-month forecast (Female), (b) 12-month forecast (Male)



(a)



(b)

Figure 8. (a) Annual totals with 3-year moving average (Female) visitors, (b) Annual totals with 3-year moving average (Male)

The forecast for the next twelve months, as shown in Figure (7), showed a continued upward trend in the number of monthly visits, for both females and males, while maintaining a clear annual seasonality. These results enhance the reliability of the model for use in future planning and resource management. Figure (8) also shows the annual totals and the overall trend using a three-year annual moving average, confirming a sustainable growth trend with no signs of decline.

Overall, the analysis shows that the data contain a clear upward pattern and significant annual seasonality, and that high-order MA models such as MA(15) are capable of accurately representing this behavior. The error indices and statistical diagnostics of the residuals also indicate that the model is suitable for short-term forecasting and can be relied upon to develop operational and organizational plans for clinics to meet the growing demand.

References

[1] Husain, Q. N., Adam, M. B., Shitan, M., & Fitrianto, A. (2017). Exploratory Extreme Data Analysis for Farmer Mac Data. *Malaysian Journal of Mathematical Sciences*, *11*, 1-16.

[2] Husain, Q. N., Adam, M. B., Shitan, M., & Fitrianto, A. (2016). Extension of Tukey’s smoothing techniques. *Indian J Sci Technol*, *9*(28), 1-5.

[3] Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 33-44.

[4] Velleman, P. F. (1982). Applied nonlinear smoothing. *Sociological Methodology*, *13*, 141-177.

[5] Hyndman, R. J. (2025). Moving averages. In *International encyclopedia of statistical science* (pp. 1559-1562). Berlin, Heidelberg: Springer Berlin Heidelberg.

[6] Somorjai, R. L., & Jarmasz, M. (1999). Exploratory data analysis of fMR images: Philosophy, strategies, tools, implementation. *NeuroImage*, *9*, S45-S45.

[7] Härdle, W. (2012). *Smoothing techniques: with implementation in S*. Springer Science & Business Media.

[8] Deveaux, R. D. (1999). Applied smoothing techniques for data analysis.

[9] Noori, N. A., & Mohammad, A. A. (2021). Dynamical approach in studying GJR-GARCH (Q, P) models with application. *Tikrit Journal of Pure Science*, *26*(2), 145-156.

[10] Abdullah, H. H., Khalaf, N. S., & Noori, N. A. (2024). Comparison of non-linear time series models (Beta-t-EGARCH and NARMAX models) with Radial Basis Function Neural Network usingReal Data. *Iraqi Journal For Computer Science and Mathematics*, *5*(3), 38.

[11] Abdullah, H. H., Noori, N. A., Mohammad, A. A., Khaleel, M. A., & Khalaf, N. S. (2025). Improving Financial Volatility Modeling Using neutrosophic Logic and Applying the GJR-GARCH Model. *Passer Journal of Basic and Applied Sciences*, *7*(2), 678-687.

[12] Ibrahim, M. Q., Mohammed, A. A., Khalaf, A. A., & Noori, N. A. (2025). Comparative Analysis of Seasonal Mixed Integrated Autoregressive Moving Average and Feed-Forward Neural Networks Models in Predicting United State Natural Hazard Casualties. *Baghdad Science Journal*, *22*(7), 2360-2374.