Pedestrian Movement Prediction and Pattern Analysis

Vladimir Toshev [1], Kaloyan Karamitov [2], Dessislava Petrova-Antonova [1,2]

[1] Faculty of Mathematics and Informatics, Sofia University, Sofia, Bulgaria – vntoshev@fmi.uni-sofia.bg, d.petrova@fmi.uni-sofia.bg
[2] GATE Institute, Sofia University, Sofia, Bulgaria – kaloyan.karamitov@gate-ai.eu, dessislava.petrova@gate-ai.eu

Keywords: ARIMA Model, Error-Trend-Seasonality Model, Pedestrian Activity, Pedestrian Movement Analysis and Patterns

Abstract

Understanding pedestrian movement patterns is essential for designing accessible and pedestrian-friendly urban environments. By analysing them, city planners can optimize resource allocation, enhance walkability, and improve infrastructure planning. This study examines pedestrian patterns in District Lozenets of Sofia city, Bulgaria, and develops a predictive model to estimate foot traffic based on environmental factors such as weather conditions and pedestrian density at different times of day.

The study utilizes data for pedestrian movement collected in the GATE Institute's City Living Lab in District Lozenets, employing radar-based sensors and incorporating weather data to analyse pedestrian movement trends. Furthermore, it explores various statistical models, such as ARIMA and ETS models, to forecast pedestrian traffic. Results indicate strong seasonal trends, with weekday peaks during commuting hours and a decline in foot traffic due to adverse weather conditions such as rain and snow. The SARIMA model demonstrates high accuracy in predicting short-term pedestrian movement patterns, outperforming alternative models in capturing both seasonal variations and long-term trends.

The findings provide valuable insights for urban planners, event organizers, and policymakers, enabling data-driven decisions for infrastructure improvements, public safety strategies, and pedestrian-friendly city planning. The study also highlights the potential of digital twin technology in urban mobility research, demonstrating the benefits of integrating real-time pedestrian data for predictive analytics.

1. Introduction

The importance of walking as a means of transportation has been steadily increasing due to its environmental friendliness and health benefits. In recent years, the necessity for walkable cities has been further underlined through issues of air pollution, traffic congestion, and urban sprawl. Therefore, prioritising more secure and pedestrian-friendly infrastructure, as well as understanding pedestrian movement and its patterns, would introduce a much-needed change in the quality of everyday life.

Urban morphology has been proven to play a crucial role in shaping pedestrian behaviour and movement patterns. It provides insight into how different urban forms influence walking habits and accessibility: street layout, building density, and public spaces are all important factors that determine the walkability of an area (Mostafa et al., 2022). As a result, increased walkability would promote more socially cohesive and vibrant communities. Moreover, integrating pedestrian data and predictive models offers invaluable tools for urban planners and policymakers.

The significance of pedestrian movement prediction and pattern recognition can be seen in the areas of Urban Planning and Infrastructure Management, Improvement in Walkability and Accessibility, Strategic Planning for Future Needs and Assessment of Meteorological Phenomena Effect and Event Impact. By analysing the patterns in pedestrian movement, city planners can make informed decisions in their choice of space allocation and infrastructure design (e.g. position and width of sidewalks, prioritisation of public transportation etc.). Moreover, insights from the study can help enhance the pedestrian experience, making the city more accessible and user-friendly by promoting a more walkable environment or by serving as a starting point in identifying key areas needing improvement. The ability to detect and predict pedestrian trends could assist cities in the optimisation of infrastructure developments and other urban improvements. Assessing how events (e.g. music festivals, sports games etc.) influence pedestrian numbers and movement will be informative regarding event planning and crowd management strategies. Predicting pedestrian traffic in accordance with weather conditions could also be of help when planning activities such as snow removal and short-term construction schedules. Due to the dynamic nature of urban spaces, continuous monitoring and adaptation are necessary to ensure that pedestrian needs are effectively met (Wang et al., 2017).

This study aims to analyze the pedestrian movement in the District of Lozenets of Sofia city, Bulgaria. It proposes a predictive model that estimates foot traffic based on various environmental factors, including weather conditions and pedestrian density at different times of the day. Specifically, it seeks to analyze pedestrian density, evaluate the environmental impact, and compare prediction models to identify the most effective approach for identifying patterns and forecasting pedestrian movement. The study is one of the showcases of the urban digital twin of Sofia city, aiming to simulate, analyze and visualize the urban environment and processes by applying the digital twin basic idea – "design, test and build first digitally" and thus showing the added value of data for decision making (Kumalasari at al., 2023).

The rest of the paper is structured as follows. Section 2 outlines the related work. Section 3 describes the data collection, while Section 4 presents the methodology followed to conduct the study. Section 5 shows the obtained results, which are further discussed in Section 6. Finally, Section 7 concludes the paper.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-G-2025-1447-2025 | © Author(s) 2025. CC BY 4.0 License.

2. Related Work

The key factors that have been found to influence pedestrian movement within urban spaces are population growth, weather conditions, time of the day, location and special events (Wang et al., 2017; Elzeni et al., 2022; Anciaes et al. 2017; Lindelöw et al., 2014; Aultman-Hall et al., 2009; Vanky et al. 2017, Shaaban and Muley, 2016; Kim, 2015). Furthermore, pedestrian activity and numbers are also influenced by the availability of bus stop locations and mid-block crossings (Xu et al., 2019). Weather conditions have also been found to impact crossing behaviour (Fourkiotis et al., 2022). In addition, a comprehensive review of the impact of urban morphology on pedestrian decision-making has been compiled (Elzeni et al., 2022).

Regression models have been employed to evaluate the impacts of weather conditions on pedestrian traffic and pedestrian activity, as well as to develop a specific model for pedestrian volumes during peak hours (Aultman-Hall et al., 2009; Vanky et al., 2017). Moreover, the Student's t-test was used to evaluate the difference in mean pedestrian volumes between hours in consideration of different environmental conditions (Aultman-Hall et al., 2009). Using correlation analysis, a statistically significant correlation was found between predictor variables like temperature, wind speed, relative humidity, and precipitation, with humidity and wind speed having the highest Pearson's Correlation Coefficient (Aultman-Hall et al., 2009). Studies focusing on pedestrian route choices employed a Multinomial Probit (MNP) model to account for route correlation and to relax the independence of irrelevant alternatives (IIA) assumption (allowing for correlation among alternative routes between the same origin-destination pair) and a logistic regression model to predict pedestrian route choices and to address the issue of separation in binary response models (Kim, 2015; Gim and Ko, 2017; Gou and Loo, 2013). The use of Maximum Likelihood Estimation (MLE) and Firth logistic regression has been explored as an alternative to conventional logistic regression methods for dealing with the separation issue of empirical data (Gim and Ko, 2017). The ARIMA model offers a better performance to predict the pedestrian traffic in comparison to support vector regression (SVR) and multiple linear regression (MLR) methods (Wang et al., 2017). Additionally, the possibility of modelling pedestrian flow rates from image data has also been explored. A CNN-LSTM approach was employed to predict pedestrian traffic flow rate and density. The number of new pedestrians and the total number of pedestrians were estimated via an STL CNN. Furthermore, a Dual Image Input CNN - MTL approach was used to extract common features from two consecutive images and task-specific layers to anticipate new and total pedestrian counts. Finally, pedestrian traffic flow rate and density for the next hour were forecasted using time-series data and MTL LSTM Modelling (Joshi and Silver, 2022).

3. Data Collection

The study utilizes data collected in GATE Institutes' City Living Lab from July 2022 to August 2024. The lab is equipped with 50 autonomous radar sensors for pedestrian counting (Figure 1). An innovative digital signal processing is applied, allowing bidirectional counting of people passing from left to right and in the opposite direction of the sensor. The counted pedestrians are stored in the internal memory of the device and transmitted at regular intervals via a wireless network using the LoRaWAN protocol at 868MHz.

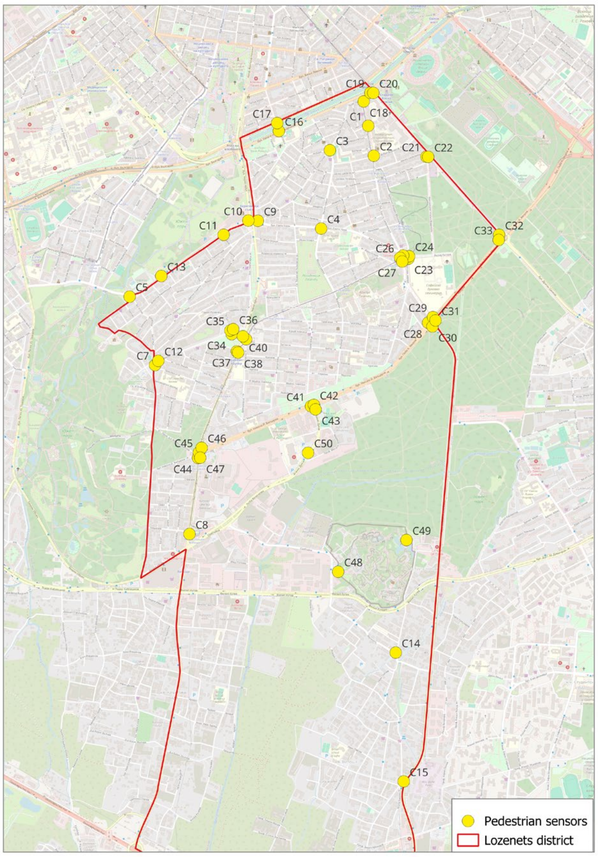

Figure 1. City Living Lab pedestrian sensor network.

Due to infrastructure renovation of the City Living Lab, data is not available between May 21 and May 28, 2024 (Figure 2). Therefore, the data was split into two datasets: before and after June 29, 2023. A decision was made to perform the analysis on the two datasets separately, to explore the effects of the missing data on the study results. The original sensor readings follow the structure:

- DATE-TIME: date and time at which the current record was made.
- PEOPLE-Radar-Left: number of pedestrians passed the radar and turned left when the record was made.
- PEOPLE-Radar-Right: number of pedestrians passed the radar and turned right when the record was made.
- Station: the unique ID of each sensor labelled from C1 to C50.

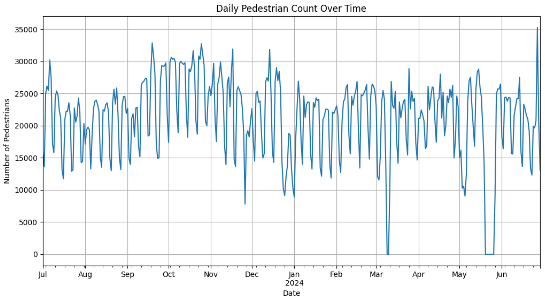

Figure 2. Missing data due to City Living Lab renovations.

The study focuses on the district of Lozenets, therefore the weather data for the study period was obtained from the Open-Meteo API (Zippenfenig, 2024), with parameters: latitude 42.70, longitude 23.40, elevation 560 m, and time zone MSK.

The dataset includes information for the date and time, temperature (°C), rain levels (mm), snowfall levels (cm), and wind speed (km/h).

4. Methodology

This section presents the methodology followed to perform the study, describing the evaluation metrics, correlation analysis and predictive models.

4.1 Evaluation Metrics

The study employs three evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) to assess model accuracy. The MAE is the average of the absolute differences between predicted and actual values. It measures the magnitude of the prediction errors in a model without considering their direction (positive or negative). A lower MAE indicates better model performance (Willmott and Matsuura, 2005).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{1}$$

where x_i = actual value
y_i = predicted value
n = number of observations.

The RMSE is the square root of the average of the squared differences between predicted and actual values. It gives more weight to large errors compared to MAE, making it sensitive to outliers. A lower RMSE indicates a better-fitting model (Chai and Draxler, 2014).

$$RMSE(x,y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

where x_i = actual value
y_i = predicted value
n = number of observations.

R^2, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It indicates how well the model explains the variance in the actual data. An R^2 value of 1 indicates a perfect fit, while a value of 0 indicates no predictive power.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{y}_i)^2} \tag{3}$$

where x_i = actual value
y_i = predicted value
$\bar{y}$ = mean of the actual values
n = number of observations.

4.2 Correlation Analysis

To explore the correlation between pedestrian count and temperature/wind speed, Pearson's R, also known as the Pearson correlation coefficient (PCC) (Pearson, 1895), was employed. It is a measure of the linear correlation between two variables. Its values range between -1 and 1, where 1 indicates a perfect positive correlation: as one variable increases, the other increases in a perfectly linear fashion, 0 indicates no correlation: there is no linear relationship between the variables, and -1 indicates a perfect negative correlation: as one variable increases, the other decreases in a perfectly linear fashion.

The correlation between pedestrian count and rain/snow was explored using a T-test for Categorical/Binary Variables (Mishra et al., 2019). The T-test is a statistical method used to determine if there is a significant difference between the means of two groups. Although it is typically used for continuous data, it is also applicable in cases where the independent variable is categorical/binary, and the dependent variable is continuous. In the context of the study's rain/snow scenario, the T-test can be used to compare the mean pedestrian counts on days with rain or no rain (or snow vs. no snow) to determine if there is a statistically significant difference in the number of pedestrians based on weather conditions. The T-statistics show the size of the difference relative to the variation in the sample data, where the p-value indicates whether the difference is statistically significant, with $p < 0.05$ indicating a significant difference in pedestrian counts between rainy and non-rainy days.

Finally, to explore the correlation between pedestrian count and temperature ANOVA for Multi-Category Variables was used (Fisher, 1992). ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more groups to determine if there are statistically significant differences among them. While a T-test compares the means of two groups, ANOVA is used when the independent variable has more than two categories (i.e., a multi-category variable). In the current study, the independent variable is categorized as a categorical variable with three or more categories (e.g., "low", "medium", and "high" temperature) and the dependent variable is a continuous variable (e.g., pedestrian count). ANOVA tests whether there is a statistically significant difference between the means of the dependent variable across the different categories of the independent variable.

4.3 ETS Model Prediction

The Error, Trend, Seasonality (ETS) model is a type of exponential smoothing method used for time series forecasting (Hyndman and Athanasopoulos, 2018; Jofipasi et al., 2018). It decomposes a time series into three components:

- Error (E): The random variations or noise in the data.
- Trend (T): The long-term movement in the data, indicating whether the series is increasing, decreasing, or stable over time.
- Seasonality (S): The repeating patterns at fixed intervals, such as daily, weekly, or monthly cycles.

The ETS model can adapt to different types of time series data by adjusting the way it handles each of these components (additively or multiplicatively). Depending on how these components interact, different variations of the ETS model are available. When predicting pedestrian counts based on time series data, the ETS model is highly useful because it handles both trend and seasonality, which are common in pedestrian flow patterns. As seen already, the data has seasonality in both daily patterns and weekly patterns (more pedestrians during rush hours, fewer at night, weekdays vs. weekends), as well as monthly/seasonal effects. A general increasing or decreasing trend in pedestrian traffic over time, depending on various external factors, is also present.

4.4 SARIMA Model Prediction

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is an extension of the ARIMA model, specifically designed to handle seasonal and non-seasonal components in time series data (Hyndman and Athanasopoulos 2018; De Livera et al., 2011). It combines autoregressive (AR) terms, moving average (MA) terms, and differencing to make a time series stationary, along with seasonal counterparts to capture recurring patterns at fixed intervals (such as daily, weekly, or monthly patterns). The SARIMA model is often represented by the notation SARIMA (p, d, q) (P, D, Q), where:

- p, d and q represent the non-seasonal AR, differencing, and MA terms, respectively.
- P, D and Q represent the seasonal AR, differencing, and MA terms, respectively.
- S represents the seasonal period, representing the length of the season (e.g., 12 for monthly data with a yearly cycle, 7 for daily data with weekly cycles).

The combination of these components allows the model to capture both short-term dependencies (non-seasonal) and longer-term seasonal patterns that repeat at regular intervals. Exogenous variables are external factors that can influence the outcome of the time series. When added to the SARIMA model, the model becomes SARIMAX, where "X" stands for the exogenous variables. These exogenous variables can enhance the model by providing additional information that might explain variation in the data beyond seasonal and autoregressive components.

To utilize the ARIMA model, an Augmented Dickey-Fuller (ADF) test is necessary (Mushtaq, 2011). The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine whether a given time series is stationary or contains a unit root (i.e., it is non-stationary). A stationary time series has a constant mean, variance, and autocorrelation structure over time, which is essential for many time series forecasting models, such as ARIMA and SARIMA, to function properly. When determining appropriate parameters for the ARIMA models, the autocorrelation function (ACF) plot, and the closely related partial autocorrelation (PAC) plot are used (Yakubu and Saputra, 2022). The Partial Autocorrelation Function (PACF) is a statistical tool used in time-series analysis to measure the correlation between a time series and its lagged values. In other words, it shows the direct relationship between an observation and its lagged values by filtering out the effects of shorter lags. The PACF helps to isolate the "pure" relationship between a time series and a particular lag by removing the influence of all shorter lags. This makes the PACF particularly useful for identifying the appropriate order of an Autoregressive (AR) model.

To model pedestrian counts, three variations of the ETS model (Additive, Multiplicative, No Seasonality) were tested alongside two ARIMA-based models:

- SARIMA (Time-Only), which uses only the time-series pedestrian count data.
- SARIMAX (Time + Weather) which employs the weather features (temperature, rain, snowfall, wind speed) as additional exogenous variables.

5. Results

This section presents the results from the analysis, including trends analysis, correlation analysis and short-term modelling.

5.1 Pedestrian Trend Analysis

Analysis of daily pedestrian count, shown in Figure 3, reveals clear seasonality, with higher activity in warmer months and considerable day-to-day fluctuation related to events like holidays and the school year. The most noticeable drops in pedestrian traffic can be attributed to Christmas, New Year’s Eve and Easter. Additionally, the decrease in pedestrian count is particularly apparent during the late summer and fall months (August through September). Weekends consistently show a lower number of pedestrian movements, reflecting a drop in commuter traffic.

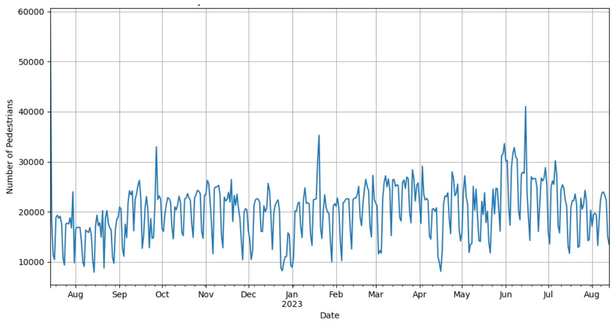

Figure 3. Daily pedestrian count.

The heatmap of pedestrian count by day and time highlights a peak in the mid-week activity from Tuesday to Thursday, with the highest pedestrian traffic during typical working hours, decreasing toward the weekend (Figure 4).

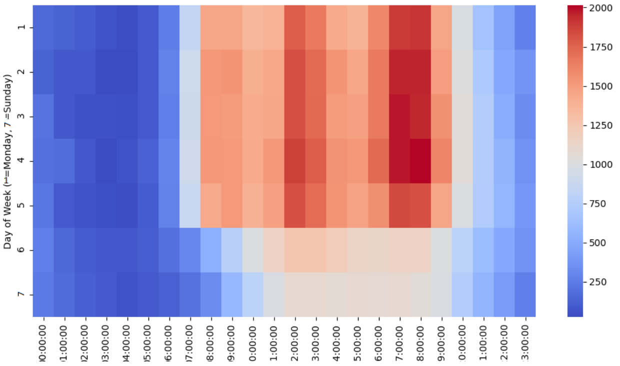

Figure 4. Average pedestrian count by weekday and day hour.

Daily traffic shows distinct morning (8:00 - 10:00 AM), midday (11:00 AM - 2:00 PM), and evening peaks (5:00 - 7:00 PM), corresponding to commuting and lunch breaks (Figure 5). Weekends have reduced pedestrian traffic, where moderate midday peaks appear on Saturday, while the number of pedestrians remains the lowest on Sunday. The median pedestrian number further confirms these weekday consistencies and weekend declines, though mid-week outliers suggest occasional high-traffic events (Figure 6). Hourly data shows predictable daily cycles, characterized by minimal activity overnight and a steady decline following the evening commute (Figure 5).

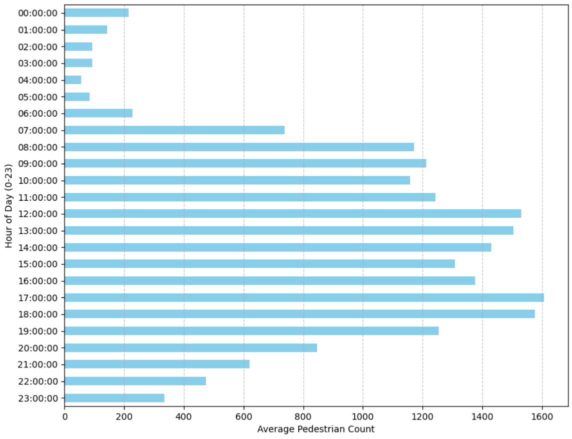

Figure 5. Pedestrian count by hour of day.

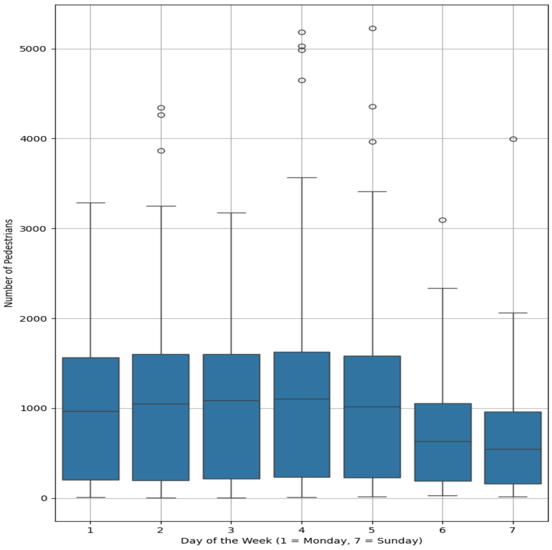

Figure 6. Pedestrian count by day of the week.

5.2 Correlation Analysis

The PCC indicated a moderate positive correlation between temperature and pedestrian count, suggesting that higher temperatures are associated with increased pedestrian traffic. Additionally, the results from the ANOVA method presented an f-statistic of 273.790, confirming that there is a significant variation in pedestrian counts across different temperature categories. The T-tests for snow and rain showed significant differences in counts during rain and snow, with both conditions reducing pedestrian activity.

5.3 Short-Term Modelling

The ACF plot of the residuals, presented in Figure 7, shows no significant autocorrelation, showing that the SARIMA model appropriately captures the time series structure. The autocorrelation at lag 0 is 1, but at all other lags, the autocorrelation is close to 0, and the points generally fall within the confidence intervals. The absence of substantial deviations beyond the confidence bounds suggests that the residuals of the SARIMA model exhibit no significant autocorrelation. This indicates that the model has effectively captured the underlying structure of the time series. This plot indicates that the model residuals exhibit characteristics of white noise, suggesting a well-fitted model.

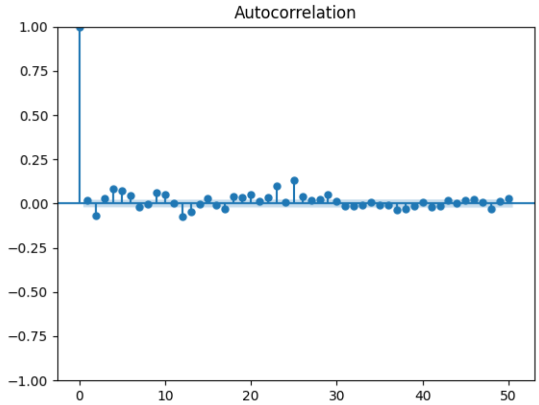

Figure 7. ACF plot of the residuals of the SARIMA model

The ACF shows significant correlations at multiple lags (positive peaks at lags 1, 6, and 12, for example), suggesting seasonality or repeating cycles in the time series data (Figure 8). The gradual decay of the peaks suggests a strong and sustained correlation across multiple lags.

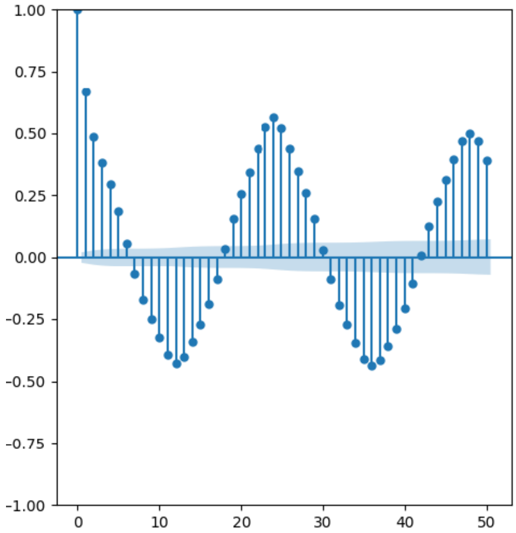

Figure 8. ACF plot of the SARIMA model.

The most notable characteristic is the significant spike at lag 1, indicating a pronounced autoregressive component, AR(1) (Figure 9). Subsequent spikes are comparatively smaller and exhibit a rapid decline after lag 1, implying that once the first lag is accounted for, additional lags contribute minimally to the predictive power of the autoregressive process.

Out of the three ETS model variations, the Additive model performs best for short-term (24-hour) and long-term (168-hour) forecasts. In comparison, the multiplicative model slightly outperforms for medium-term (48-hour) forecasts (Table 1). The No-Seasonality model consistently underperforms across all timeframes, with significantly higher errors, which can be expected, as it lacks the mechanism to account for these predictable, cyclical fluctuations in pedestrian traffic. The SARIMA model, on the other hand, outperformed its ETS counterparts in capturing both regular fluctuations and non-seasonal variations, achieving an accuracy of 0.98 for a 12-hour forecast and over 0.9 for a 24-hour forecast (Figure 10). Thus, it is a preferable model for predicting pedestrian counts (Table 2).

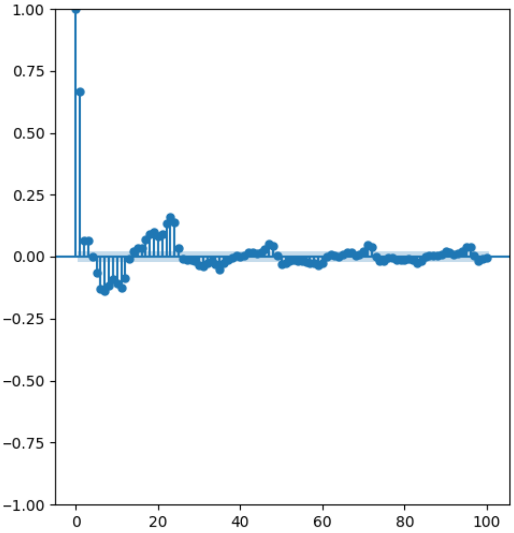

Figure 9. PACF plot of the SARIMA model

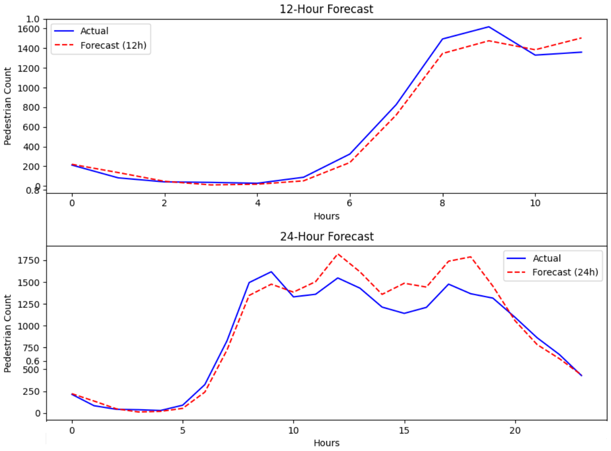

Figure 10. SARIMA Model short-term performance.

Timeframe	Best performing model	MAE	RMSE
(24-Hour)	ETS(Additive)	139.11	177.68
(48-Hour)	ETS(Multiplicative)	251.03	363.18
(168-Hour)	ETS(Additive)	228.97	336.76

Table 1. Best performing ETS models comparison.

Timeframe	Model	MAE	RMSE	R^2
(12-Hour)	2022/2023 Data	68.24	86.31	0.98
(24-Hour)	2022/2023 Data	124.99	167.25	0.91
(48-Hour)	2023/2024 Data	130.42	160.46	0.90
(168-Hour)	2023/2024 Data	178.86	239.70	0.77

Table 2. Best performing SARIMA models comparison.

6. Discussion

One of the key limitations of the applied approach lies in its inability to account for external factors that have a significant impact on pedestrian traffic, such as popular events and proximity to points of interest (e.g., malls, parks, schools, and transit hubs). These variables can drive large spikes or drops in pedestrian counts that the SARIMA model, while adept at handling seasonality and trends, is not equipped to capture without further customization or additional external information. For example, pedestrian numbers often surge during public events such as concerts, sporting events, festivals, or parades. These events are irregular and often unpredictable based solely on the historical time series data. Without incorporating such events as exogenous variables, the SARIMA model cannot explain the sudden spikes or sharp drops seen on outlier days. Consequently, the model's forecasts during these periods are likely to be inaccurate, as it treats these outlier events as noise rather than relevant factors. Furthermore, the distance between the observation points and key points of interest like shopping malls, parks, schools, and transportation hubs can strongly influence pedestrian counts. Areas near these high-traffic locations tend to experience regular pedestrian flows influenced by shopping hours, school timetables, and public transportation schedules. However, the current approach does not consider the geographical and spatial characteristics of the data, which are crucial for understanding how pedestrian traffic fluctuates based on proximity to such locations.

A more sophisticated approach to address these limitations would involve the clustering of data based on variables such as proximity to points of interest, and the customization of the time series model for different clusters. The benefits would include improved forecasting accuracy, targeted analysis and more dynamic and context-aware models. By tailoring the model for specific clusters, the model's ability to capture localized seasonal patterns and responses to external influences can be significantly enhanced, thereby improving the accuracy of forecasts. Clustering the sensors also allows for more targeted analysis and decision-making. For example, public planning authorities can use forecasts specific to high-traffic areas near malls or schools to optimize public safety, infrastructure, and service delivery. Moreover, by incorporating external factors such as events, holidays, and proximity to key points of interest, the model becomes more dynamic and context-aware, leading to more realistic and actionable forecasts for pedestrian management.

7. Conclusion

This study explores various statistical models, such as ARIMA and ETS models, to forecast pedestrian traffic. Pedestrian counts by day of the week appear to be consistent with lower counts during the weekend, exhibiting a Morning and Evening peak, which could contribute to people’s daily commute, with a gradual decline after 8 pm. Peak traffic days appear to be Tuesday, Wednesday and Thursday, with lower numbers on Monday and a gradually declining traffic towards the weekends.

Considering weather conditions, rain and snow significantly decrease pedestrian counts, with snow having a more pronounced negative effect. Temperature has a moderate positive impact on pedestrian traffic, meaning more people are outside when it’s warmer. Wind speed has a weak but statistically significant relationship with pedestrian counts, but the effect is not as pronounced. The ANOVA model suggests that pedestrian counts vary significantly across different temperature ranges, likely indicating a sweet spot of temperature where pedestrian traffic peaks (e.g. mild, comfortable weather may encourage higher foot traffic).

The SARIMA model was utilized to forecast pedestrian counts based on time series data that exhibited clear seasonal patterns, daily and weekly cycles, and occasional outliers. The model was selected due to its ability to account for both non-seasonal and seasonal components within the data, providing a comprehensive framework for handling the complex dynamics of pedestrian traffic, helping effectively capture the seasonal fluctuations in pedestrian traffic, particularly the daily and weekly trends, which were evident from the heatmaps and time series plots of the data. Residual analysis indicated that the model performed reasonably well, with no significant autocorrelation in the residuals and no evident patterns, suggesting that the core dynamics of the data were well captured by the SARIMA model. The Additive ETS model was tested for comparison but did not perform as well in capturing the seasonality and trends inherent in the pedestrian data. The SARIMA model, on the other hand, provided better accuracy in predicting both regular fluctuations and non-seasonal variations. However, certain areas of the model, particularly during outlier days, require further refinement, as the residuals displayed larger deviations during these periods.

Acknowledgements

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155 and the deployEMDS project, funded by Digital Europe Programme DIGITAL-2022-CLOUD-AI-03 under grant agreement no. 101123520.

References

Anciaes, P. R., Nascimento, J., Silva, S., 2017. The distribution of walkability in an African city: Praia, Cabo Verde. *Cities*, Volume 67, 2017, Pages 9-20, ISSN 0264-2751. https://doi.org/10.1016/j.cities.2017.04.008.

Aultman-Hall, L., Lane, D., Lambert, R. R., 2009. Assessing Impact of Weather and Season on Pedestrian Traffic Volumes. *Transportation Research Record*, 2140(1), 35-43. https://doi.org/10.3141/2140-04.

Chai, T., Draxler, R. R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

De Livera, A. M., Hyndman, R. J., Snyder, R. D., 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496), 1513–1527. https://doi.org/10.1198/jasa.2011.tm09771

Elzeni, M. M., ElMokadem, A. A., Badawy, N. M., 2022. Impact of urban morphology on pedestrians: A review of urban approaches. *Cities*, Volume 129, 2022, 103840, ISSN 0264-2751. https://doi.org/10.1016/j.cities.2022.103840.

Fisher, R.A., 1992. Statistical Methods for Research Workers. In: Kotz, S., Johnson, N.L. (eds) *Breakthroughs in Statistics*, Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_6

Fourkiotis, M., Kazaklari, C., Kopsacheilis, A., Politis, I., 2022. Applying deep learning techniques for the prediction of pedestrian behaviour on crossings with countdown signal timers. *Transportation Research Procedia*, Volume 60, 2022, Pages 536-543, ISSN 2352-1465. https://doi.org/10.1016/j.trpro.2021.12.069.

Gim, T.-H. T., Ko, J., 2017. Maximum Likelihood and Firth Logistic Regression of the Pedestrian Route Choice. *International Regional Science Review*, 40(6), 616-637. https://doi.org/10.1177/0160017615626214

Hyndman, R.J., & Athanasopoulos, G., 2018. *Forecasting: Principles and Practice*, 2nd ed. OTexts, Melbourne, Australia. https://OTexts.com/fpp2

Jofipasi, C. A., Miftahuddin, M., Sofyan, H., 2018. Selection for the best ETS (error, trend, seasonal) model to forecast weather in the Aceh Besar District. *IOP Conference Series: Materials Science and Engineering*, 352, 012055. The 7th AIC-ICMR on Sciences and Engineering 2017 18–20 October 2017, Banda Aceh, Indonesia. https://doi.org/10.1088/1757-899X/352/1/012055

Joshi, R., Silver, D., 2022. Pedestrian Traffic Prediction using Deep Learning. The International FLAIRS Conference Proceedings, 35. https://doi.org/10.32473/flairs.v35i.130731.

Kim, H., 2015. Walking distance, route choice, and activities while walking: A record of following pedestrians from transit stations in the San Francisco Bay area. *Urban Design International, Volume* 20, 144–157 (2015). https://doi.org/10.1057/udi.2015.2.

Kumalasari, D., Koeva, M., Vahdatikhaki, F., Petrova Antonova, D., Kuffer, M., 2023. Planning Walkable Cities: Generative Design Approach towards Digital Twin Implementation. *Remote Sensing*, 15(4), 1088. https://doi.org/10.3390/rs15041088

Lindelöw, D., Svensson, Å., Sternudd, C., Johansson, M., 2014. What limits the pedestrian? Exploring perceptions of walking in the built environment and in the context of every-day life. *Journal of Transport and Health*, 1(4), 223-231.

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., Pandey, G., 2019. Application of student's *t*-test, analysis of variance, and covariance. *Ann Card Anaesth*, 2019 Oct-Dec, 22(4):407-411. doi: 10.4103/aca.ACA_94_19. PMID: 31621677; PMCID: PMC6813708.

Mushtaq, R., 2011. Augmented Dickey Fuller Test. http://dx.doi.org/10.2139/ssrn.1911068.

Pearson, K., 1895. Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London*, 58, 240–242. https://doi.org/10.1098/rspl.1895.0041

Shaaban, K., Muley, D., 2016. Investigation of Weather Impacts on Pedestrian Volumes. *Transportation Research Procedia*, Volume 14, 2016, Pages 115-122, ISSN 2352-1465. https://doi.org/10.1016/j.trpro.2016.05.047.

Vanky, A., Verma, S., Courtney, T., Santi, P., Ratti, C., 2017. Effect of weather on pedestrian trip count and duration: City-scale evaluations using mobile phone application data. *Preventive Medicine Reports*. 8. 10.1016/j.pmedr.2017.07.002.

Xu, W., Ruiz, N., Pierce, K., Huang, R., Meyer, J., Duthie, J., 2019. Detecting Pedestrian Crossing Events in Large Video Data from Traffic Monitoring Cameras, 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 3824-3831. https://doi.org/10.1109/BigData47090.2019.9005655.

Wang, X., Liono, J., Mcintosh, W., Salim, F., 2017. Predicting the city foot traffic with pedestrian sensor data. The 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.4108/eai.7-11-2017.2273699.

Willmott, C. J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79-82. https://doi.org/10.3354/cr030079

Yakubu, U., Saputra, M., 2022. Time Series Model Analysis Using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for E-wallet Transactions during a Pandemic. *International Journal of Global Operations Research,* 3(3), 80-85. https://doi.org/10.47194/ijgor.v3i3.168.

Guo, Z., Loo, B. P. Y., 2013. Pedestrian environment and route choice: evidence from New York City and Hong Kong. *Journal of Transport Geography*, 28, 2013, 124-136, ISSN 0966-6923. https://doi.org/10.1016/j.jtrangeo.2012.11.013.

Zippenfenig, P., 2024. Open-Meteo.com Weather API, Version 1.4.0. https://doi.org/10.5281/zenodo.14582479.