

# MACHINE-LEARNING FORECASTING OF EVENT-RELATED CROWD INFLUX AND URBAN DENSITY IN SMART CITIES EVIDENCE FROM PUBLIC EVENT DATA AND MOBILE NETWORK SIGNALS IN EMILIA-ROMAGNA

Adonay A. Veiga  
Claudia J. B. Abbas

---

*Reliable anticipation of crowd build-up is a practical requirement for contemporary smart-city governance, particularly where transport systems, event programming, and public-safety planning intersect. This study presents a machine-learning framework for forecasting two operationally important outcomes in Emilia-Romagna, Italy: the class of event-related increases in local people presence, and short-horizon future density within census-defined sub-areas (ACEs). The analysis integrates automatically harvested public-event information, daily weather covariates, and anonymised mobile-network observations sampled every 15 minutes. Event data were collected for August–September 2019 and August–September 2020, yielding 5,123 retrieved records before preprocessing; 3,133 valid event-day instances were used for the classification task. For density forecasting, the base learning matrix comprised 58,696 hourly windows derived from 506 ACEs observed across 116 days.*

*A comparative evaluation of tree ensembles, kernel methods, and neural networks was conducted in a supervised-learning setting. For event-impact classification, Gradient Boosted Decision Trees achieved the strongest overall performance, with a peak accuracy of 0.86 on the best-performing feature configuration. For density forecasting, short-horizon custom accuracy was highest for boosted trees and multilayer perceptrons, whereas Support Vector Regression produced the lowest error profile and the most gradual deterioration as the forecasting horizon extended to 24 hours. The analysis also shows that a simple binary event-presence flag does not improve density forecasting, and that neighbouring-ACE variables contribute far less than initially expected. Framed for urban development and smart-city applications, the study demonstrates how structured event intelligence and mobility analytics can support transport coordination, public-space management, and policy-support systems for safer and more efficient cities.*

*Index Terms* — smart cities; urban mobility; public events; crowd analytics; machine learning; policy support systems; urban planning

---

## INTRODUCTION

Data-rich urban environments increasingly allow city authorities to move from reactive management of crowding and congestion to anticipatory planning. Among the most valuable, yet operationally challenging, tasks in this context are estimating the local inflow associated with public events and forecasting near-term changes in people density across the urban fabric [1]. These capabilities matter not only for event organisers, but also for municipal transport planners, emergency managers, commercial operators, and regional policy makers responsible for maintaining the continuity and safety of city services [2].

The study reported here addresses these challenges through a comparative machine-learning framework grounded in two linked prediction problems. The first estimates the class of event-related increases in local presence [3]. The second forecasts future crowd density inside ACEs (*Aree di CEnsimento*), the census-defined sub-areas used to partition Emilia-Romagna. Both tasks are designed for deployment in a policy-support setting in which model outputs can inform shuttle allocation, access management, transport messaging, and preventive crowd control [4].

The manuscript has been written for a smart-city and urban-development readership. Accordingly, the emphasis is not limited to algorithmic performance alone. Equal attention is given to data governance, the structured representation of urban events, model interpretability, and the planning value of predictive outputs. The contribution is therefore methodological and applied at the same time: it shows how heterogeneous public-event information and mobile-network observations can be translated into an operational decision aid for mobility services [5].

## URBAN CONTEXT, DATA SOURCES, AND STUDY DESIGN

### *Observation window and territorial unit*

The empirical setting is the Italian region of Emilia-Romagna. Two observation windows were analysed: August–September 2019 (61 days) and August–September 2020 (61 days), for a total of 122 days. The spatial unit of analysis is the ACE, a census-defined sub-area used to represent geographically bounded portions of the region. The regional partition includes 506 ACEs, which form the basis of the density-forecasting task.

### *Data streams*

Three data streams were combined.

1. **Public-event data.** Event information was automatically harvested from municipal, regional, and private tourism websites. Approximately 3,100 event records were retrieved for August–September 2019 and about 2,000 for August–September 2020, for a total of 5,123 raw records.
2. **Weather data.** Daily weather covariates were collected from a public meteorological source and included minimum, maximum, and mean temperature, wind speed, humidity, pressure, rainfall, and precipitation type.
3. **Mobile-network data.** Anonymised observations from a major Italian mobile network operator were sampled every 15 minutes for each ACE. The data include aggregated descriptors such as age, gender, nationality, contract type, and commuter-versus-resident status.

These three streams jointly support the article’s two predictive tasks: event-related inflow classification and ACE-level density forecasting.

#### *Why this study fits urban development and smart-city research*

From an urban-development perspective, the analytical value of the study lies in its direct connection to service coordination and spatial management. Forecasting where and when people concentrations intensify supports decisions on transport reinforcement, crowd routing, temporary access controls, retail staffing, and public-space surveillance. The work therefore sits squarely within the scope of smart-city scholarship concerned with data-informed governance, mobility planning, and the practical orchestration of urban resources.

## **EVENT INFORMATION MODEL AND AUTOMATED RETRIEVAL**

### *Structured event representation*

Because no single authoritative registry existed for regional public events, a dedicated event database was built. The data model stores each event through a structured representation that includes identifier, title, description, qualifier, category, theme, location, start and end dates, organiser, source URL, target users, and event dimension. The *dimension* field is especially important because it transforms qualitative event descriptions into a usable proxy for expected attendance.

Table 1: Attendance-dimension scale used for event labelling.

<b>Dimension</b>	<b>Expected attendance range</b>
1	0–49 people
2	50–199 people
3	200–999 people
4	1,000–9,999 people
5	10,000–99,999 people
6	100,000 people or more

The dimension label was assigned through a two-stage procedure. First, when location type was available, venue characteristics were used as a proxy for expected attendance: libraries, restaurants, and small halls typically indicated small events, whereas stadiums, arenas, and major fairgrounds were associated with large events. Second, for events without a reliable location type, category and town information were used to infer the class. This allowed the event database to be standardised for downstream machine learning even in the absence of direct attendance counts.

### *Automated retrieval pipeline*

A dedicated software pipeline written in Java and ASP populated the event database. The workflow scanned more than 30 selected websites that disseminate event information in Emilia-Romagna; extracted event attributes from HTML, JSON, and metadata; geocoded locations through automated matching of place names and addresses; removed duplicates; harmonised formats; identified recurring events; and injected additional events not fully captured online, such as local festivities or contextual events. Across 2019–2020, the software

retrieved nearly 18,000 major events in the region and more than 37,000 records when minor recurrent events were also counted. The months used in this study were then extracted from that broader archive.

---

**Algorithm 1** Operational data-to-prediction workflow

---

- 1: Collect daily weather variables for the study window
  - 2: Crawl selected public websites and parse structured event metadata
  - 3: Geocode venues and assign events to ACEs
  - 4: Deduplicate records and standardise event descriptors
  - 5: Estimate event dimension using venue, category, and town information
  - 6: Ingest anonymised mobile-network observations at 15-minute resolution
  - 7: Build the event-impact classification dataset
  - 8: Build the ACE density-forecasting dataset using rolling 24-hour windows
  - 9: Train and tune classification models on event-level records
  - 10: Train and tune regression models on density windows
  - 11: Export the best-performing predictors to a policy-support environment
- 

## PREDICTIVE TASKS AND MODELLING STRATEGY

### *Task 1: Event-related inflow classification*

The first task predicts the class of percentage increase in local presence associated with a future event. After preprocessing, 3,133 valid event-day instances remained from the initial 5,123 retrieved records. Records were excluded when geolocation was missing, when the event fell outside ACEs covered by the mobile data, or when date fields were inconsistent; multi-day events were expanded into daily records.

The event-level design matrix contained 14 predictor fields before one-hot encoding, plus the target class. The predictors included average peak, average peak per day, average peak per month, category, theme, province, duration, day of the week, target attendance, dimension, minimum temperature, maximum temperature, average temperature, and precipitation type. The target was defined as the percentage increase in visitors calculated from the maximum observed peak on the event day relative to the average peak for that weekday. Five classes were used:

$$[0, 25]\%, \quad (25, 50]\%, \quad (50, 75]\%, \quad (75, 100]\%, \quad \text{and} \quad > 100\%.$$

The dataset was split into 2,193 training examples (70%) and 940 test examples (30%). The class distribution was strongly imbalanced, especially for the two middle-high classes.

Table 2: Configuration of the event-impact classification dataset.

Quantity	Training set	Test set
[0, 25]%	1,152	494
(25, 50]%	137	59
(50, 75]%	26	11
(75, 100]%	17	7
> 100%	861	369
<b>Total</b>	<b>2,193</b>	<b>940</b>

Random Forest, Extra Trees, Support Vector Machines, Multilayer Perceptron, Gradient Boosted Decision Trees, and AdaBoost were compared. Hyperparameters were tuned through grid search for tree ensembles and random/grid search for kernel methods. For the neural model, an AutoML-assisted architecture search produced a shallow network with one hidden layer of 256 neurons, ReLU activation, and a five-neuron softmax output layer.

To probe feature relevance, four dataset variants were created. Variant 1 removed only `Duration`; Variant 2 removed `Duration`, `MIN_Temp`, and `MAX_Temp`; Variant 3 also removed `AVG_Temp`; Variant 4 further removed `Average peak per day`. This design allowed feature contribution to be assessed through controlled ablation.

### *Task 2: ACE density forecasting*

The second task predicts crowd density in an ACE at horizons of 1, 2, 4, 8, 12, and 24 hours. The initial regression matrix comprised 58,696 rows and 432 columns, where each row represented a rolling 24-hour history for a given ACE-day and each column stored hourly values of one of 18 mobile-data attributes over the preceding day (18 attributes  $\times$  24 hours). A further 1,728 candidate columns were added initially to summarise the average, maximum, minimum, and total values of neighbouring ACE attributes over the same 24-hour window.

The final design was sharpened after feature analysis. Neighbouring-ACE variables ranked very low in importance (for example, beyond the 80th position in a representative Random Forest run), and repeated tests showed negligible gains in custom accuracy. These variables were therefore removed from the final reported models. A second dataset version added a binary indicator of whether a public event was scheduled in the target ACE; this modification did not improve performance and was not retained.

Decision Trees, Extra Trees, Random Forest, Support Vector Regression, Multilayer Perceptron, Gradient Boosted Decision Trees, and AdaBoost were evaluated. The training split was 75%, with 25% held out for testing. Model quality was assessed using mean absolute error (MAE), root mean squared error (RMSE), and a custom accuracy metric that counts a forecast as correct when it deviates by less than 5% from the observed value.

## **RESULTS**

### *Event-impact classification results*

The classification experiment shows a clear hierarchy among the candidate methods. Tree ensembles consistently dominated the single neural baseline, and the best overall result was obtained by Gradient Boosted Decision Trees on Variant 1, with accuracy equal to 0.86. Extra Trees reached 0.85 on the same variant, while Random Forest remained stable at 0.83. Support Vector Machines improved substantially when the weakest feature was removed, rising from 0.634 on the original dataset to 0.74 on Variant 1. The MLP architecture underperformed throughout.

Table 3: Overall classification accuracy by model and dataset variant.

Model	Original	V1	V2	V3	V4
Random Forest	0.83	0.83	0.81	0.78	0.79
Extra Trees	0.83	0.85	0.76	0.80	0.76
SVM	0.634	0.74	0.73	0.74	0.73
MLP	0.70	0.52	0.58	0.52	0.56
GBDT	<b>0.85</b>	<b>0.86</b>	<b>0.82</b>	<b>0.75</b>	<b>0.77</b>
AdaBoost	0.84	0.83	0.81	0.77	0.76

Beyond aggregate accuracy, the classification task reveals two substantive patterns. First, the very small size of the (75, 100]% class severely constrained recall for all models. Second, predictive behaviour was much stronger for the extreme > 100% class than for the two thinly populated middle-high classes. In operational terms, the framework is therefore more reliable for distinguishing routine events from strong surges than for finely discriminating among scarce intermediate cases.

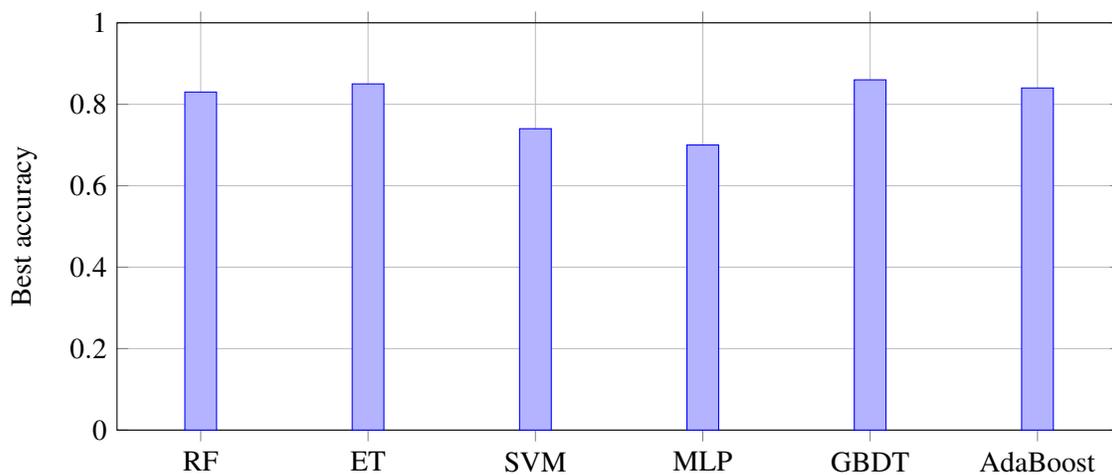


Figure 1: Best observed classification accuracy for each model family across the feature variants.

### *Density-forecasting results*

As expected, forecast quality deteriorated with longer lead times. The decline, however, was model-specific. Random Forest and MLP delivered high short-horizon custom accuracy, but their error levels rose sharply with time. Support Vector Regression exhibited the smallest error profile and the smoothest performance decay across the full horizon range. Boosted ensembles retained strong short-term accuracy and remained competitive through 24 hours.

Table 4: Custom accuracy for density forecasting on the original dataset.

Model	1 h	2 h	4 h	8 h	12 h	24 h
Random Forest	0.920	0.830	0.690	0.650	0.650	0.620
SVR	0.825	0.820	0.804	0.796	0.792	0.779
MLP	0.948	0.900	0.834	0.748	0.700	0.659
GBDT	0.944	0.893	0.795	0.731	0.710	0.680
AdaBoost	0.944	0.883	0.795	0.735	0.720	0.706

The lowest absolute errors were achieved by Support Vector Regression, whose MAE increased only from 0.046 at 1 hour to 0.059 at 24 hours, while RMSE ranged from 0.202 to 0.214 over the same interval. By contrast, the MLP returned the strongest 1-hour custom accuracy (0.948), but its 24-hour accuracy fell to 0.659. AdaBoost and GBDT both exceeded 0.94 custom accuracy at 1 hour; AdaBoost maintained 0.706 at 24 hours, slightly above GBDT (0.680).

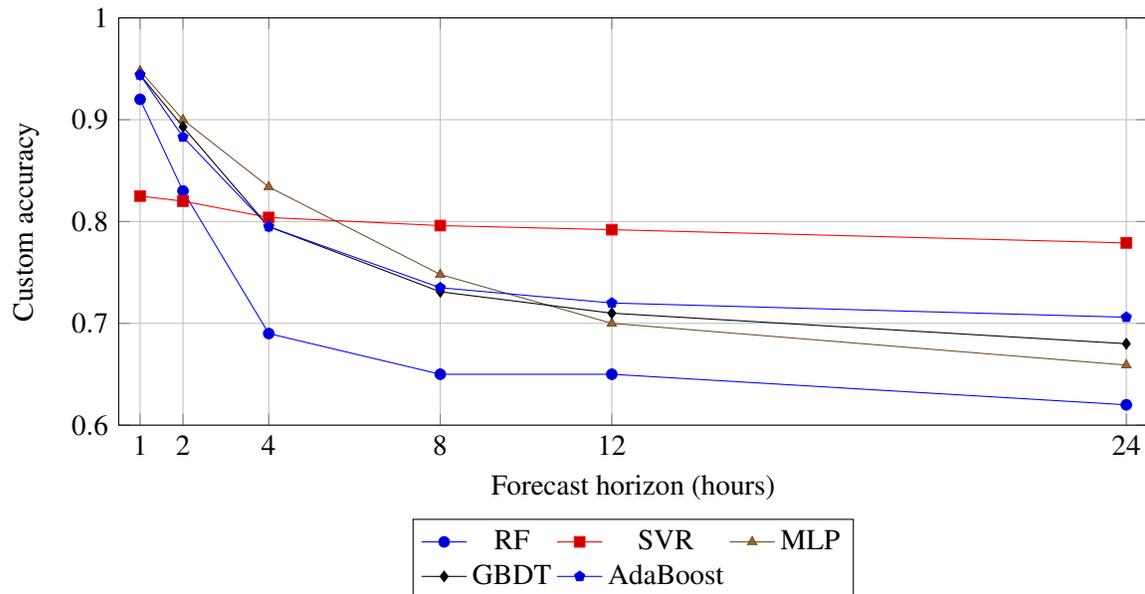


Figure 2: Custom accuracy declines with forecast horizon for all density-forecasting models, but the rate of decline differs substantially across model families.

## DISCUSSION

### *Feature relevance and dataset design*

Two design findings are especially important. First, for the classification task, the strongest results were obtained when only the Duration variable was removed. This indicates that the event’s temporal length did not materially improve prediction once the model already knew the event category, weather profile, and baseline crowding context [6]. By contrast, removing temperature and average-peak variables degraded tree-based performance, confirming that these fields carry much of the signal.

Second, for the density task, neighbouring-ACE variables contributed far less than originally expected.

Although the regression problem was initially framed as depending on both within-ACE and adjacent-area dynamics, feature-ranking analyses showed that external-area summaries were not among the dominant predictors [7]. This finding is operationally useful because it allows a leaner deployment pipeline with reduced learning time and lower computational burden.

#### *Planning implications for urban management*

The practical significance of the results extends beyond model comparison. In a smart-city policy-support system, the classification output can be used to flag events likely to induce sharp local surges, while the density forecasts can guide time-sensitive service decisions. Three immediate applications are apparent:

1. **Transport reinforcement.** Forecasted crowd increases can support decisions on extra buses, shuttles, and station staffing [8].
2. **Traffic and access management.** Short-horizon density forecasts can trigger temporary route changes, dynamic messaging, or timed access controls [9].
3. **Commercial and public-service adjustment.** Expected increases in local presence can inform opening hours, staffing levels, and security allocation in event-adjacent zones [10].

In this sense, the article contributes to urban development not merely by predicting crowd patterns, but by identifying how predictive signals can be converted into concrete management actions.

#### *Computational considerations*

The most accurate low-error regression model, SVR, also imposed the highest computational cost. On the reference hardware reported in the experiments (Intel i7-1255U, 1.70 GHz, 16 GB RAM), training time increased from 70 minutes for 1-hour forecasting to 1,446 minutes for 24-hour forecasting. This trade-off is substantively important: urban authorities may prefer the smoother error profile of SVR for strategic forecasting, but shorter training and update cycles may favour boosted trees in operational settings requiring frequent retraining.

## **LIMITATIONS AND FUTURE RESEARCH**

The study also exposes limitations that should shape future work.

**Event-database completeness.** Even with a broad web-retrieval strategy, some large events may remain outside official dissemination channels, including private, political, or informal gatherings. A centralised regional reporting hub for major events would improve coverage and reduce dependence on fragmented web sources.

**Attendance estimation.** The attendance-dimension field is indispensable, but it remains a proxy. More reliable organiser-provided estimates, especially for events expected to attract more than 500 participants, would improve both class balance and model realism.

**Class imbalance.** The scarcity of training examples in the (50,75]% and (75,100]% classes weakens discrimination among intermediate surge levels. Additional data collection across a longer time horizon would help stabilise these predictions.

**Temporal scope.** The analysis covers two late-summer windows, one of which falls in a pandemic-affected year. Broader seasonal coverage would allow stronger generalisation across tourism cycles, school calendars, and annual event calendars.

Future studies should therefore pursue longer-term data collection, direct organiser attendance estimates, and richer integration of external urban signals such as transit schedules or venue access controls.

## CONCLUSION

This manuscript presents a complete and policy-oriented account of how machine learning can be used to forecast event-related crowd influx and near-term urban density in a smart-city setting. Using structured public-event records, weather covariates, and anonymised mobile-network signals from Emilia-Romagna, the study evaluates multiple model families against two operational tasks central to urban governance.

The results show that Gradient Boosted Decision Trees provide the strongest overall solution for classifying event-related inflow increases, while Support Vector Regression yields the most stable low-error profile for density forecasting over longer horizons. Just as importantly, the analysis clarifies which data elements matter most: temperatures and baseline crowd indicators are central to event-impact classification, whereas a binary event flag and neighbouring-ACE summaries contribute far less to density forecasting than expected.

For urban development and smart-city practice, the contribution is clear. Predictive mobility analytics become most valuable when they are translated into transport scheduling, crowd-risk mitigation, and more efficient allocation of public resources. The framework presented here offers a realistic path from heterogeneous urban data to actionable planning intelligence.

## REFERENCES

- [1] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [2] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42.
- [3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- [4] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
- [5] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- [6] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.
- [7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

- [8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
- [9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.
- [10] Bellodi, E., Zese, R., Petrovich, C., Frascella, A., & Bertasi, F. (2024). Predicting the impact of public events and mobility in Smart Cities. *IET Smart Cities*, 6, 253–275.

Adonay A. Veiga, Department of Technology, State University of Feira de Santana, Feira de Santana 44036-900, Brazil

Claudia J. B. Abbas, Department of Technology, State University of Feira de Santana, Feira de Santana 44036-900, Brazil

Manuscript Published; 11 October 2025.