Scientific Reports

https://doi.org/10.1038/s41598-026-56879-7

Article in Press

An interpretable generative probabilistic framework for demand characterization and consistency checking in dock-based bicycle-sharing systems: the case of BiciMad

Received: 18 December 2025

Accepted: 3 June 2026

Published online: 11 June 2026

Cite this article as: Vallez C.M., Contreras D. & Castro M. An interpretable generative probabilistic framework for demand characterization and consistency checking in dock-based bicycle-sharing systems: the case of BiciMad. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-56879-7

Carlos M. Vallez, David Contreras & Mario Castro

An Interpretable Generative Probabilistic Framework for Demand Characterization and Consistency Checking in Dock-Based Bicycle-Sharing Systems: The Case of BiciMad

Carlos M. Vallez (ORCID: 0000-0002-8758-4664)[1,*], David Contreras (ORCID: 00000-0003-1405-3878)[1,+], and Mario Castro (ORCID: 0000-0003-3288-6144)[1,2,*,+]

[1]Institute for Research in Technology, ICAI School of Engineering, Universidad Pontificia Comillas, c\Rey Francisco 4, 28008 Madrid, Spain
[2]Grupo Interdisciplinar de Sistemas Complejos (GISC), Universidad Pontificia Comillas, Madrid, Spain.
[*]cmvallez@comillas.edu; marioc@comillas.edu
[+]these authors contributed equally to this work

ABSTRACT

Bicycle-sharing systems (BSS) have become an important component of sustainable urban mobility, but their demand remains difficult to model. Usage varies across hours, stations, weather conditions, and types of day, while the available data often provide only a partial view of the underlying demand process. This study proposes an interpretable probabilistic framework to characterize and generate synthetic demand for BiciMad, Madrid's dock-based BSS, using trip-level data from 2018 and 2019. The contribution is not the introduction of new probability distributions, but the calibrated integration of standard probabilistic components into a demand-side generative framework. Trip distances are modeled with Gamma distributions, and hourly trip counts are represented with Negative Binomial distributions conditioned on hour, day type, and precipitation. These components are combined with empirical station-popularity profiles to generate synthetic origin–destination demand under explicit contextual assumptions.

Validation against observed data shows that the framework provides calibrated uncertainty estimates, with empirical coverage of the 95% prediction intervals close to the nominal level across contextual scenarios. An external consistency check using 2019 data further shows the practical value of the approach, as it helped identify systematic timestamp misattributions that were later confirmed by the data provider. The proposed framework is not intended as a full capacity-aware operational simulator. Instead, it provides a simple, interpretable, and uncertainty-aware baseline for demand characterization, synthetic demand generation, exploratory disruption analysis, and data-quality consistency checking in dock-based bicycle-sharing systems.

Introduction

Bicycle-sharing systems (BSS) have become essential components of sustainable urban mobility, supporting multimodal transportation, reducing congestion, and contributing to environmental goals. As part of the broader Mobility-as-a-Service paradigm[1], BSS complement public transport networks and provide flexible, short-range mobility options. Their relevance is further underscored by the central role of human mobility in urban planning, epidemiological modeling, transportation forecasting, and location-based services[2]. As cities increasingly adopt sustainable transportation solutions, BSS have become both an important mobility service and an active field of research[3,4].

Despite their operational importance, modeling demand in BSS remains challenging. Bicycle-sharing demand varies substantially across users, stations, temporal patterns, and contextual conditions, while the availability of real-time feedback data is often limited. In dock-based systems, these challenges are particularly relevant because demand imbalances may lead to empty origin stations or full destination stations, creating the need for bicycle rebalancing operations. This problem is typically divided into two sub-problems: (1) supply and demand modeling, and (2) optimization of bike distribution[5,6]. The present study focuses on the first of these components, namely probabilistic demand characterization.

More broadly, modeling human mobility is a major research area, with numerous movement patterns described in the literature[7] and many mathematical and statistical models developed to analyze them[8]. A central tool for understanding mobility systems is the Origin–Destination (OD) matrix, which quantifies the number of trips from origin location i to destination j. Traditional approaches, such as the Gravity model[9], have long provided interpretable aggregate descriptions of mobility flows

and remain a reference framework. A prominent formulation is

$$T_{ij} = \frac{P_i \cdot P_j}{f(D_{ij})},$$

yet classical OD models may struggle with generalization, overfitting to specific stations, limited stochastic representation, low explainability, and sensitivity to imbalanced flows[8]. At the same time, human mobility research has expanded through AI and data-driven methodologies[10], enabling more flexible and precise modeling of spatiotemporal phenomena.

Mobility demand models are commonly classified into individual- and population-level approaches[8]. Individual-level models simulate mobility flows based on traveler behavior, using methods that range from random walks to sophisticated algorithms with many explanatory variables; however, they often involve high complexity, behavioral uncertainty, and limited generalizability across scenarios. In contrast, population-level models aggregate flows between origin–destination pairs and rely on theoretical, empirical, or machine-learning formulations. In the context of shared mobility systems[11], and BSS in particular[5], forecasting and demand modeling have traditionally relied on regression-based and machine-learning approaches, including deep recurrent neural networks[12–14] and probabilistic frameworks such as Hidden Markov Models[15]. Related modeling and simulation approaches have also been explored in other shared-mobility domains, including open-data-driven e-scooter system design[16] and agent-based simulation of vehicle-sharing systems[17], showing that generative and interpretable frameworks are relevant across multiple shared transport modes.

These approaches have clear strengths. Machine-learning models can provide accurate forecasts when enough data and covariates are available; probabilistic state models can represent latent regimes or system states; and agent-based or operational simulators can describe system dynamics in greater detail. However, these benefits come with trade-offs. High-performing forecasting models may require rich input data and substantial computational resources, and their outputs can be difficult to interpret. Detailed simulators, in turn, often depend on behavioral assumptions, station-capacity information, bicycle availability, and other supply-side data that are not always available in public trip-record datasets. Classical OD and gravity-style models are easier to interpret, but they usually provide a limited representation of uncertainty. Recent studies in complex systems modeling have highlighted the value of simple, interpretable probabilistic frameworks when robustness, transparency, and decision support are important objectives[18]. Similarly, prior work has shown that uncertainty and overdispersion should be explicitly represented in shared mobility systems, particularly when robustness, interpretability, and predictive performance are all relevant[19]. For applications such as redistribution planning, anomaly screening, and exploratory scenario analysis, this points to the need for models that balance interpretability, uncertainty representation, and computational cost.

To clarify the positioning of the present work, Table 1 compares the proposed framework with representative families of approaches used in BSS and shared-mobility demand modeling. The purpose of this comparison is not to suggest that one type of model is generally preferable to the others. Rather, it helps define the specific role of the proposed framework: a transparent probabilistic baseline for demand characterization, synthetic demand generation, and consistency checking when only limited operational data are available.

Table 1. Positioning of the proposed framework with respect to representative families of approaches used in BSS and shared-mobility demand modeling.

Approach family	Main objective	Strength	Limitation relative to this study	Position of this work
Gravity / aggregate OD models	Estimate aggregate flows between zones or stations	Interpretable and analytically tractable	Often deterministic or weakly probabilistic; limited representation of hourly uncertainty	Adds contextual stochastic hourly demand and trip-level synthetic generation
Regression / ML forecasting	Predict future station-level or system-level demand	High predictive accuracy when rich features are available	May require many covariates and can provide limited interpretability	Provides a transparent probabilistic baseline rather than a black-box predictor
Deep probabilistic forecasting	Forecast demand while representing uncertainty	Flexible and accurate with sufficient data	Higher model complexity and lower transparency	Offers a lower-complexity way to characterize uncertainty
HMM / probabilistic state models	Infer station/bike states or latent regimes	Captures stochastic transitions and hidden states	Usually focused on state detection rather than OD-demand generation	Generates synthetic trip-level demand from calibrated probabilistic components
Agent-based / operational simulation	Simulate users, fleet dynamics, and supply constraints	Behaviorally and operationally detailed	Requires capacity, occupancy, and behavioral data	Provides a demand-side baseline when such data are unavailable
Empirical resampling	Generate synthetic profiles from historical data	Simple and faithful to observed records	Limited parametric interpretation and limited control over scenarios	Uses interpretable distributions to summarize demand regularities and generate scenarios
This study	Characterize and generate BSS demand under contextual conditions	Interpretable, calibrated, low-complexity, and generative	Not capacity-aware; limited representation of behavioral substitution	Complements forecasting and operational simulation as a probabilistic demand-characterization baseline

This positioning motivates a complementary approach: a simple and interpretable probabilistic framework that can characterize demand, generate synthetic trip profiles, and support consistency checks using limited data. Such a framework is not intended to replace high-resolution forecasting models or full operational simulators. Its purpose is more specific: to

provide a transparent probabilistic baseline for understanding the main demand regularities that can be inferred from trip-level records.

In this study, we propose such a generative probabilistic framework for dock-based BSS. The framework focuses on two core aspects of demand: trip distances between stations and hourly trip counts under different contextual conditions. Using trip-level data from Madrid's BiciMad system, we show that Gamma distributions provide a suitable representation of trip-distance patterns, while Negative Binomial distributions capture hourly trip counts, including overdispersion and variation associated with hour, precipitation, and weekday/weekend conditions. These components are then combined with empirical station-popularity profiles to generate synthetic origin–destination demand under explicit contextual assumptions.

Novelty and contribution

The novelty of this work does not lie in the use of Gamma or Negative Binomial distributions in isolation. Both are standard probabilistic tools. They are used here because they offer interpretable and empirically adequate representations of two key features of BSS demand: trip-distance patterns and overdispersed hourly counts.

The contribution lies in their integration with empirical station-popularity profiles into a calibrated demand-side generative framework. This structure links contextual hourly demand, station-level origin heterogeneity, and distance-based OD allocation, allowing synthetic OD demand profiles to be generated under explicit assumptions.

The scope of the framework is deliberately focused. It is not intended to replace high-performance forecasting models or capacity-aware operational simulators. Instead, it provides a probabilistic baseline for demand characterization, uncertainty representation, synthetic data generation, and data-quality consistency checking.

The paper makes five specific contributions. First, it shows that BiciMad trip distances and contextual hourly demand variability can be represented by stable, low-dimensional probabilistic families during the selected operational period. Second, it combines these components into a demand-side generative framework that produces synthetic OD demand profiles under explicit contextual assumptions. Third, it validates the framework through probabilistic calibration, including the empirical coverage of 95% prediction intervals across weekday/weekend and rainy/non-rainy scenarios. Fourth, it shows how the framework can support data-quality consistency checking, including the identification of systematic timestamp misattributions in 2019 data that were later confirmed by the data provider. Fifth, it provides a spatial-sensitivity analysis of station-unavailability scenarios, showing that the proposed global-renormalization rule preserves meaningful local concentration without requiring an exogenous nearest-neighbor constraint.

The proposed approach should therefore not be read as a simple application of standard distributions to a new dataset. Rather, these distributions serve as building blocks within a calibrated generative framework that connects demand volume, contextual variability, station-level heterogeneity, and distance-based OD allocation.

Overall, this study provides an interpretable probabilistic framework for characterizing bicycle-sharing demand and generating synthetic demand profiles under explicit assumptions. It complements more complex forecasting and simulation approaches by offering a low-complexity baseline for uncertainty representation, exploratory analysis, and operational data-quality consistency checking in dock-based bicycle-sharing systems. Its transferability to other urban contexts or shared mobility systems should be assessed through further adaptation and validation.

.

Results

Empirical characterisation of trip distances

The empirical distribution of trip distances derived from BiciMad data reveals a highly consistent pattern between 2018 and 2019, with a pronounced peak at short-to-intermediate urban distances and a smooth, heavy tail (Fig. 1). Kernel density estimates confirm that the shape of the distribution remains remarkably stable across both years, even after normalizing counts to account for differences in total trip volume.

To ensure comparability across years, distance histograms were normalized by the number of active stations in each period, since BiciMad substantially expanded its infrastructure in 2020. This adjustment is essential: without it, the raw 2020 distribution would appear artificially inflated simply because users had more origin–destination combinations available.

Even after this station-adjusted normalization, 2020 still shows visible deviations from the 2018–2019 pattern, consistent with the profound changes in mobility behavior during the COVID-19 pandemic. These observations motivate the use of a parametric family capable of capturing both the skewness and variability of the empirical distribution, particularly for modeling purposes in later sections.

To evaluate candidate distributions, we fitted six commonly used models to the empirical 2018 distance distribution: the Gamma, Lognormal, Weibull, Rayleigh, Normal, and Exponential distributions. Visual comparison (Fig. 2) shows that the Gamma distribution provides the best agreement across short, medium, and long distances, whereas the Exponential and

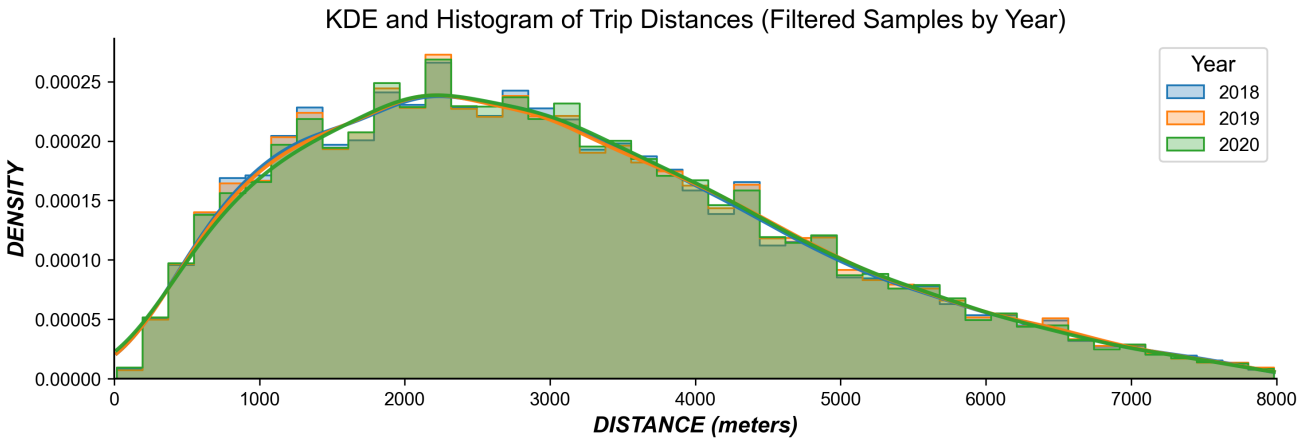


Figure 1. Trip distance distribution for 2018, 2019, and 2020 based on normalized histograms and kernel density estimates. Note how the distribution remains similar in 2018 and 2019, while the overall volume and pattern changed significantly in 2020 due to the impact of COVID-19.

Normal distributions fail to capture the distribution peak or the tail behavior. The Lognormal distribution performs reasonably well but systematically deviates at the extremes.

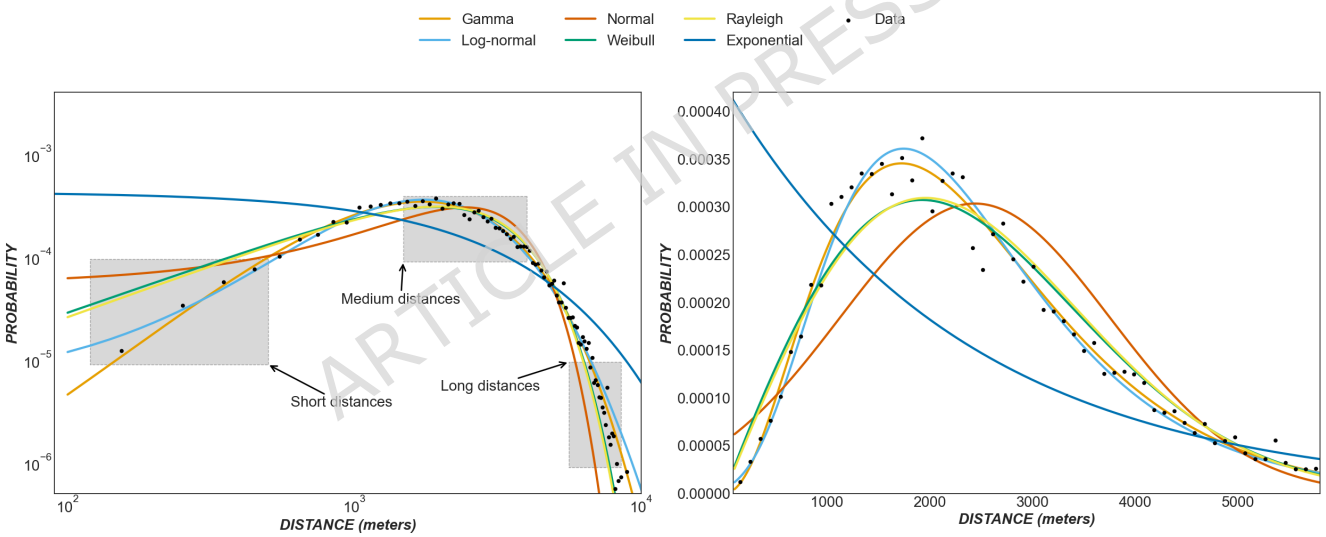


Figure 2. Empirical distribution of trips between stations at distance, d, for the 2018 Bicimad dataset. Left: fit to the most common probability distributions used in distances, on a doubly logarithmic scale; right: same plot in a linear scale. Note how this (typical) representation is misleading, as it does not provide enough resolution at the tails. Overall, the Gamma distribution is the one that best captures all the data points in the set, closely followed by the log-normal distribution. In Appendix A, we make this comparison more quantitative.

Quantitative comparison using KL divergence confirms the superiority of the Gamma distribution, yielding a divergence of 4.5×10^{-5}, substantially lower than that of alternative models. This result holds consistently across temporal subsets, including hourly aggregations, reinforcing the suitability of the Gamma distribution for generative modeling. Additional methodological details on the KL computation, parameter estimation, and distributional comparisons are provided in Appendix A for completeness.

A complementary goodness-of-fit assessment was conducted using the Anderson–Darling test[20] for the Gamma and Lognormal distributions, the two best-performing candidates according to the Kullback–Leibler divergence. This test, an extension of the Kolmogorov–Smirnov procedure with increased sensitivity to tail behavior[21], compares the empirical cumulative distribution function with the fitted theoretical one, assigning greater weight to discrepancies in the extremes. Gamma and

Lognormal each provided an excellent fit to the empirical distance data, and the identical Anderson–Darling statistic (5.715) suggests that the practical difference between them is small. Gamma was selected as a pragmatic choice because it consistently yielded a slightly lower KL divergence and offered a convenient form for the generative simulator. This decision is also consistent with the broader bike-sharing literature, where distance distributions have been reported as Gamma, Lognormal, or Weibull depending on the system and scale[22].

The Gamma distribution is not presented here as a methodological novelty in itself. Rather, it provides an empirically supported and interpretable distance component for the generative framework. Its role is to support a stable distance-based OD allocation mechanism, not to establish a universal law of bicycle-sharing trip distances.

Hourly variation in fitted Gamma parameters

Although the overall shape of the distance distribution remains similar across hours of the day, the fitted Gamma parameters exhibit meaningful temporal patterns (Fig. 3). The scale parameter increases during morning and afternoon peak periods, reflecting longer typical trip distances during commuting hours. Conversely, off-peak periods show more concentrated distributions with lower scale values. Despite these variations, the fundamental distributional form remains stable, allowing the model to capture hourly dynamics without requiring a complex feature set.

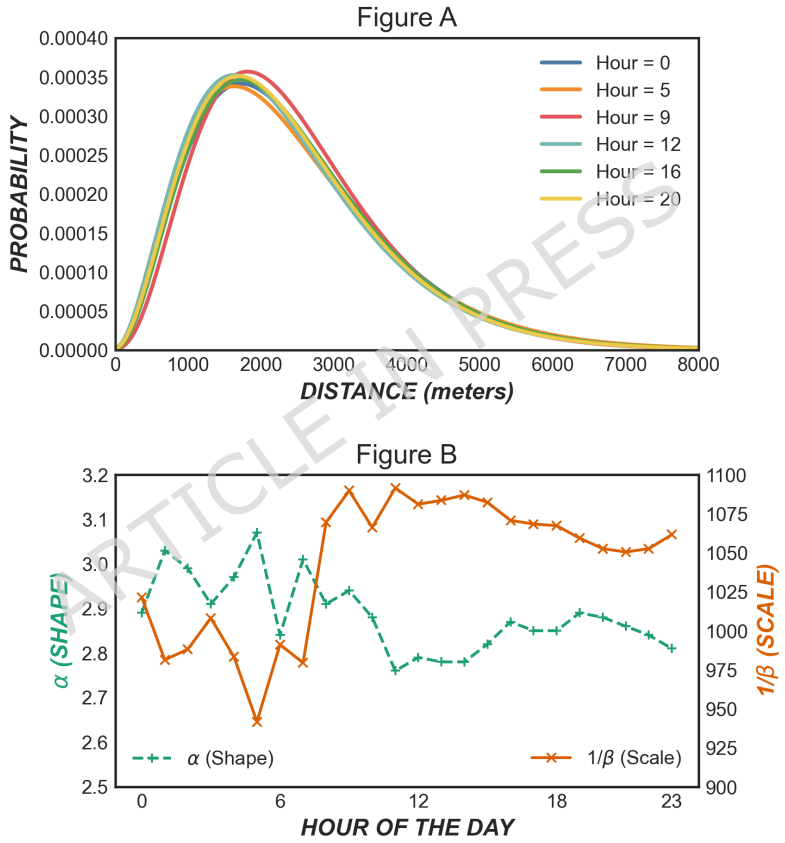


Figure 3. Top: Comparison of the best-fit gamma distributions for a sample of representative hours. Although the parameters in the bottom panel exhibit some variation throughout the day, the shape of the distribution remains largely unchanged. Bottom: Values of the parameters *shape* and *scale* of the Gamma distribution fitted for different hours during the day. The scale parameter is related to a *typical* mean travel distance, so the lower panel suggests that that distance is larger at rush hours.

Modeling hourly trip counts under contextual conditions

The empirical dataset provides trip timestamps only at an hourly resolution, with no minute-level information available. Consequently, trips beginning at 9:15 or 9:35 are both assigned to the 09:00–10:00 interval in our modeling framework. Our objective is to derive a theoretical distribution capable of predicting the number of trips occurring within each hourly window, which subsequently enables the probabilistic simulation of origin–destination flows using the Gamma-based distance model described above.

Trip counts aggregated at an hourly scale exhibit substantial variability that cannot be captured by Poisson models, which assume mean–variance equality. Instead, the empirical data display pronounced overdispersion, motivating the use of Negative Binomial (NegBin) distributions. When fitted separately under four contextual conditions—weekday versus weekend/holiday and rainy versus non-rainy days—negative binomial models accurately reproduce both the mean and variance of observed trip volumes.

Weather and day type exert strong influence over trip demand. Rainfall significantly reduces hourly usage, while weekends and holidays not only exhibit lower volumes overall but also show altered temporal patterns, including the absence of pronounced morning and afternoon commuting peaks. The fitted NegBin distributions capture these behavioral differences effectively, providing a flexible mechanism for scenario-dependent demand modeling.

The NegBin distributions fitted for each hour under non-rainy weekday conditions are illustrated in Appendix B Fig. 7, where the model closely reproduces the underlying empirical histograms. Additional discussion of why the Negative Binomial provides such an accurate representation—including its interpretation as a Poisson–Gamma mixture and its connection to maximum-entropy arguments—is provided in Discussion section.

The relevance of the Negative Binomial component is therefore not only descriptive. It provides an explicit probabilistic representation of overdispersion in hourly BSS demand, which is essential for generating realistic uncertainty intervals and for later consistency-checking applications.

Validation through probabilistic interval coverage

In addition to parameter estimation, we performed a series of probabilistic validation exercises to assess the fidelity of the fitted Gamma and NegBin models. Using the estimated distributions, we generated synthetic data samples—analogous to posterior predictive checks—and compared these with observed trip counts. For internal validation using 2018 data, hourly prediction intervals (2.5th–97.5th percentiles) were obtained from the fitted Negative Binomial distributions across all four contextual scenarios (weekday/weekend × rainy/non-rainy). Figure 4 shows the model-based 95% prediction intervals for hourly trip counts under these scenarios. Empirical interval coverage was computed separately by comparing the observed 2018 hourly counts against the corresponding intervals. Across all scenarios, 8,021 out of 8,456 observations fall within the corresponding 95% prediction intervals, yielding an overall coverage of 94.86%. Coverage by scenario was 96.23% for working sunny days, 93.20% for working rainy days, 94.18% for weekend/holiday sunny days, and 92.98% for weekend/holiday rainy days (see Table 2 and Appendix H for the full hour-by-hour breakdown).

Table 2 summarizes the empirical coverage of the 95% prediction intervals across the four contextual scenarios. These results show that coverage remains close to the nominal 95% level across all scenarios, although performance is slightly lower under rainy conditions, particularly on weekends/holidays.

Table 2. Empirical coverage of the 95% prediction intervals for hourly trip counts under the four contextual scenarios. Detailed hour-by-hour results are reported in Appendix H.

Scenario	Within interval / Total	Coverage (%)
Working sunny day	3956 / 4111	96.23
Working rainy day	1549 / 1662	93.20
Weekend/Holiday sunny day	1668 / 1771	94.18
Weekend/Holiday rainy day	848 / 912	92.98
Overall	8021 / 8456	94.86

This result is important for the intended role of the framework. The model is not assessed only by how well it reproduces average demand, but also by whether the uncertainty it generates is consistent with the variability observed in the data. This calibration is necessary for both synthetic demand generation and data-quality consistency checking, since both depend on reliable probabilistic bounds.

For each hour/scenario combination, the lower and upper bounds of the 95% prediction interval were computed from the fitted Negative Binomial distribution using the 2.5th and 97.5th percentiles. The number of empirical observations falling below or above these limits was then counted, and coverage was computed as the percentage of observed hourly counts lying within the corresponding interval:

$$\text{Coverage} = \frac{\text{total} - (\text{numhi} + \text{numlow})}{\text{total}} \times 100 \tag{1}$$

Figure 4 displays the model-based 95% prediction intervals for hourly trip counts under the four contextual scenarios, while Table 2 reports the corresponding empirical coverage.

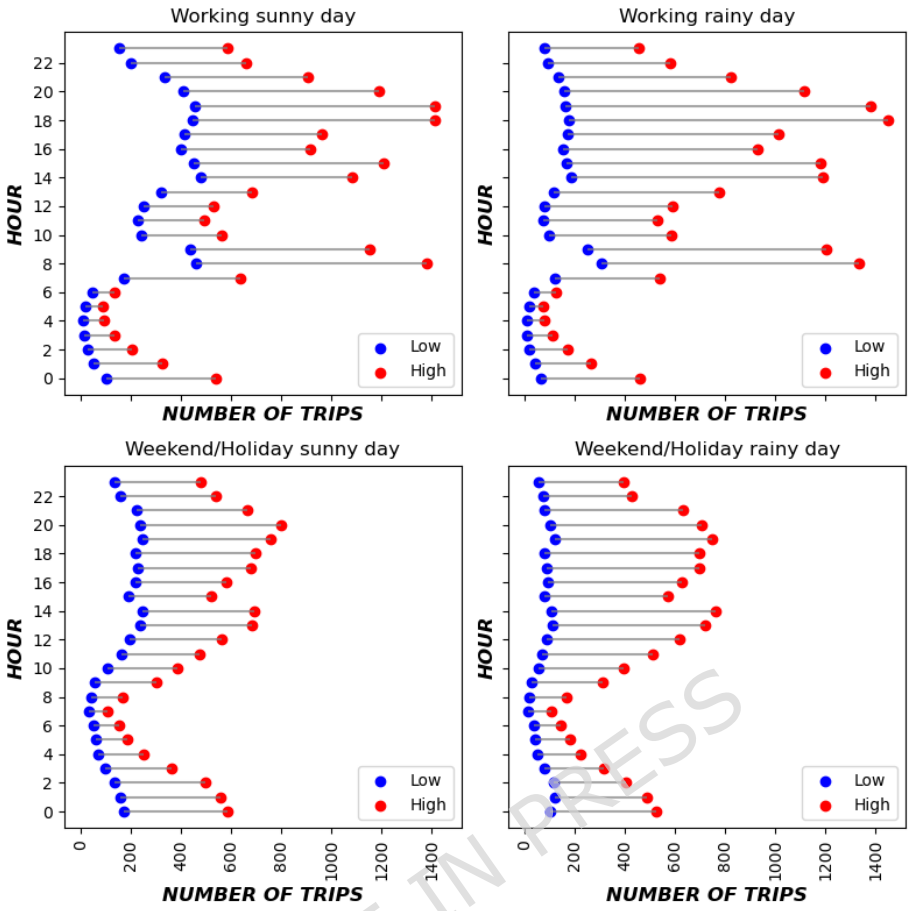


Figure 4. Model-based 95% prediction intervals (2.5th–97.5th percentiles) for hourly trip counts under the four contextual scenarios, obtained from the fitted Negative Binomial distributions. Empirical coverage is reported separately in Table 2 and Appendix H.

A detailed breakdown of these coverage statistics is provided in Appendix C.

For external validation, the 2019 dataset was compared with synthetic demand profiles generated from the 2018-fitted Gamma and Negative Binomial models. After harmonizing station availability between years, we detected systematic deviations during the early-morning period (6–8 AM), as detailed in Appendix D. These deviations were not isolated random errors; they formed a coherent temporal pattern that persisted across repeated checks. A uniform +2-hour correction brought the 2019 observations back into alignment with the expected demand patterns. This adjustment was later confirmed by the data provider, who attributed the issue to daylight-saving transitions and operational interventions. This result provides an external and operationally meaningful validation of the framework: beyond generating synthetic demand profiles, the model can also serve as a probabilistic baseline for detecting inconsistencies in real operational datasets.

Probabilistic framework for synthetic demand generation

By integrating the fitted Gamma and Negative Binomial distributions with empirically derived station-popularity profiles, we developed a demand-side generative framework for producing synthetic trip datasets under explicit contextual assumptions. In this sense, the framework can be used as a synthetic demand simulator, although not as a full operational simulator of the dock-based system. The simulation procedure operates in three stages: (1) sampling the number of trips per hour from the appropriate Negative Binomial distribution; (2) assigning origin stations based on popularity weights; and (3) generating destination stations by sampling from the hour-specific Gamma distance distribution and mapping the resulting distances to feasible OD pairs.

The framework simulates plausible demand profiles while preserving the uncertainty and heterogeneity represented by the fitted probabilistic components. The validation therefore supports a narrower interpretation: the generated datasets reproduce the main statistical features explicitly modeled in this study, namely trip-distance distributions and contextual hourly trip-count

variability.

We formalize the trip generation process in a dock-based bike-sharing system using a probabilistic graphical model (PGM) expressed in plate notation. The objective is to describe how contextual variables shape the joint distribution of hourly trip counts and origin–destination (OD) flows.

Mathematical specification

The hourly trip count is modeled as

$$N \sim \text{NegBin}(p, r), \tag{2}$$

where the parameters (p, r) depend on $HOUR$, $RAINY$, and $WEEKEND$. For each origin station $s \in \{1, \ldots, S\}$,

$$N_s \sim \text{Multinomial}\big(N, \mathbf{P}_{HOUR}\big), \qquad \mathbf{P}_{HOUR} = \big(P^{(1)}_{HOUR}, \ldots, P^{(S)}_{HOUR}\big), \tag{3}$$

so that N_s denotes the number of trips originating at station s during that hour.

For each hour/context combination, the origin-station popularity vector $\mathbf{P}_{\text{HOUR}}$ was estimated empirically as the normalized frequency of observed trip departures from each station in the filtered dataset:

$$P^{(s)}_{HOUR} = \frac{n^{(s)}_{HOUR}}{\sum_{j=1}^{S} n^{(j)}_{HOUR}} \tag{4}$$

where $n^{(s)}_{\text{HOUR}}$ denotes the number of departures from station s under the corresponding hour/context condition. These weights therefore represent marginal historical origin-station activity shares and do not explicitly account for station capacity, dock count, spatial dependence, station competition, catchment overlap, or dynamic substitution effects.

For each origin–destination pair (s, k), we introduce a latent score that is a deterministic function of the distance, $d_{s,k}$, between both stations. Namely,

$$\theta_{s,k} = \frac{\beta_s^{\alpha_s}}{\Gamma(\alpha_s)} d_{s,k}^{\alpha_s - 1} e^{-\beta_s d_{s,k}}, \quad d_{s,k} > 0 \tag{5}$$

where (α_s, β_s) are hour- and context-specific parameters (again depending on $HOUR$, $RAINY$, and $WEEKEND$). These scores are normalized to obtain OD assignment probabilities

$$\pi_{s,k} = \frac{\theta_{s,k}}{\sum_{k' \neq s} \theta_{s,k'}}. \tag{6}$$

Finally, OD-specific trip counts are drawn from a multinomial distribution

$$N_{s,k} \sim \text{Multinomial}\big(N_s, \pi_s\big), \tag{7}$$

where $\pi_s = (\pi_{s,1}, \ldots, \pi_{s,S})$ with $\pi_{s,s} = 0$.

Figure 5 summarizes this hierarchical structure using plate notation. Grey nodes denote model parameters, white nodes denote stochastic variables derived from the model, and plates (rectangles) indicate replication over stations and destinations.

The PGM formulation provides the structural backbone for the generative simulator described below, linking hourly trip volumes, station popularity, and distance-based OD allocation within a single probabilistic framework.

By integrating the fitted Gamma and Negative Binomial distributions with empirically derived station popularity profiles, we constructed a generative simulator capable of producing synthetic trip datasets under various scenarios. The simulator operates in three stages:

1. **Hourly trip counts.** For each hour and contextual condition (weekday/weekend, rainy/non-rainy), sample the total number of trips N from the corresponding $\text{NegBin}(p, r)$ distribution.

2. **Origin assignment.** Allocate these N trips across origin stations via a multinomial draw using the hourly popularity weights $\mathbf{P}_{HOUR}$, yielding station-level counts N_s.

3. **Destination assignment.** For each origin station s, generate candidate OD scores $\theta_{s,k}$ from the hour-specific Gamma distribution, normalize them to obtain π_s, and sample destination counts $N_{s,k}$ from $\text{Multinomial}(N_s, \pi_s)$. Distances implied by these OD pairs are consistent with the fitted Gamma distance model.

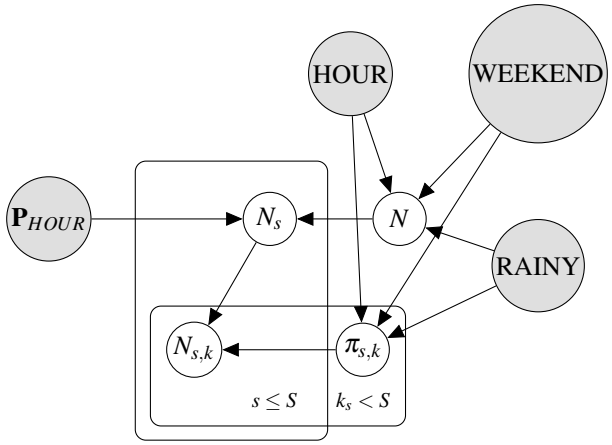


Figure 5. Probabilistic graphical model (plate notation) for hourly trip generation. Grey circles represent parameters or latent variables; white circles denote stochastic quantities derived from the model. Plates indicate replication over origin stations (s) and destination indices (k).

A more detailed schematic representation of the simulator workflow is provided in Appendix E.

This framework enables the generation of synthetic mobility patterns while explicitly preserving uncertainty and heterogeneity through the fitted probabilistic components. Rather than claiming full empirical equivalence between synthetic and observed data, the present validation supports a narrower interpretation: the probabilistic framework reproduces the two marginal features explicitly modeled in this study, namely trip-distance distributions and hourly trip-count variability under contextual conditioning.

Applications: redistribution scenarios and anomaly detection

Two practical applications demonstrate the utility of the proposed framework.

First, the synthetic demand simulator supports stylized counterfactual analysis of station unavailability, such as closures due to construction, maintenance, or temporary events. By removing a station from the feasible OD set and renormalizing the distance-based probabilities, the model estimates how demand would be redistributed under the assumption that affected trips are reassigned among feasible alternatives rather than canceled. This makes it possible to identify redistribution pressure induced by missing infrastructure and to highlight potentially vulnerable areas. However, because the procedure does not model user-level search behavior, station capacity, or local substitution dynamics, the experiment should be interpreted as a null redistribution scenario rather than as a full behavioral prediction. A detailed illustrative example is provided in Appendix F, where we also compare the original global-renormalization rule with a local K-nearest-neighbor redistribution heuristic to assess the sensitivity of this assumption. The results show that the proposed rule preserves a substantial degree of local spatial concentration without imposing a nearest-neighbor constraint, supporting its interpretation as a flexible null redistribution mechanism.

Second, the model detects anomalies in real operational data. By comparing observed hourly counts with probabilistic expectations, we uncovered systematic timestamp errors in late 2019 data—an inconsistency later confirmed by the system operator. This illustrates how probabilistic modeling can enhance data quality monitoring in mobility systems. A detailed account of this anomaly-detection process, together with supporting analyses and visualizations, is provided in Appendix D.

Discussion

The main contribution of this work is to show that a small set of interpretable probabilistic components can be integrated into a calibrated demand-side generative framework for dock-based BSS. This objective is different from high-resolution demand forecasting or capacity-aware operational simulation. The framework should therefore be evaluated as a probabilistic baseline for demand characterization, synthetic profile generation, and data-quality consistency checking, rather than as a replacement for machine-learning predictors or full operational simulators.

Although the individual components of the framework are standard, their integration addresses a specific need in BSS demand modeling. The proposed framework jointly represents contextual hourly demand uncertainty, station-level origin heterogeneity, and distance-based OD allocation while keeping the assumptions explicit and empirically calibrated. In this sense, the contribution is integrative and validation-oriented rather than algorithmic.

The results support this interpretation. The Gamma and Negative Binomial distributions capture two important statistical regularities of BSS usage: trip-distance patterns and hourly demand variability. Despite the complexity of individual route choices and the diversity of user motivations, aggregate mobility data exhibit recurrent distributional structure. This suggests that some demand regularities can be represented with relatively simple probabilistic models, provided that their scope is clearly defined.

A probabilistic interpretation of the distance component further supports these findings. In our formulation, trip distance is defined as the separation between an origin–destination pair, computed using an asymmetric distance metric that yields distinct values for $A \to B$ and $B \to A$. Among the classical probability families tested in the mobility literature, the Gamma distribution consistently provided the closest empirical match, with minimal KL divergence and stable behavior across temporal subsets. This is consistent with the widespread use of Gamma laws to describe travel distances under heterogeneous and context-dependent route conditions.

A theoretical justification for the emergence of the Gamma distribution arises from maximum entropy considerations. When only two coarse-grained constraints—the mean distance $\bar{d}$ and the mean logarithmic distance $\bar{L} = \langle \log d \rangle$—are imposed, the least-biased distribution consistent with these constraints is precisely the Gamma distribution[23]. This view aligns with long-standing interpretations of gravity-like mobility models[9], which posit that aggregate flows reflect entropic equilibria rather than detailed individual-level decision procedures. The use of a logarithmic constraint is further supported by Weber–Fechner laws of perceptual scaling, which state that humans perceive quantities such as time, money, and distance on approximately logarithmic scales[24]. Maximizing Shannon entropy under these assumptions yields

$$S = -\sum p_d \log p_d - \lambda \left(\sum p_d - 1\right) - \beta \left(\sum p_d d - \bar{d}\right) - (\alpha - 1)\left(\sum p_d \log d - \bar{L}\right),$$

whose solution is the Gamma density

$$p_d = \frac{\beta^{\alpha}}{\Gamma(\alpha)} d^{\alpha-1} e^{-\beta d}.$$

Alternative explanations—such as interpreting the Gamma distribution as the sum of multiple exponential components—are less compelling here, since the empirically estimated shape parameters are non-integer and consistently below 3. Overall, maximum entropy reasoning offers a plausible theoretical explanation for why aggregated urban trip distances may converge to a Gamma-like structure despite underlying behavioral heterogeneity.

The Negative Binomial distribution likewise provides a useful model for hourly trip counts, as it captures the substantial overdispersion present in the data. Its interpretation as a Poisson–Gamma mixture offers a natural explanation for why trip counts exhibit variance exceeding the mean: latent heterogeneity in mobility propensity—arising from weather, day type, and behavioral differences—induces variability that simple Poisson models cannot represent. This structure is also interpretable, since it separates contextual modulation from intrinsic demand variability more explicitly than black-box forecasting approaches.

By integrating these probabilistic components into a demand-side generative framework, the proposed approach moves beyond descriptive fitting. It enables the generation of synthetic demand profiles under explicit contextual assumptions and provides probabilistic bounds against which observed data can be compared. Within this scope, the framework can support exploratory analyses of altered or disrupted demand conditions, as well as consistency checks in operational datasets. The timestamp anomaly detected in late 2019 illustrates this second use. Rather than replacing visual inspection, the framework provides a probabilistic baseline against which deviations from expected hourly demand can be quantified. In this case, the observed deviations were later confirmed by the system operator as timestamp inconsistencies, showing that the framework can contribute to data-quality screening as well as demand characterization.

The framework assumes conditional stability within the modeled contexts, namely hour of day, day type, and precipitation regime. This assumption is appropriate for generating conditional demand profiles within a relatively stable operational period, but it limits the use of the model as a long-term forecasting tool. In particular, synthetic profiles generated from one year should not be interpreted as projections of future adoption, demographic change, tariff changes, or network evolution. This is why the analysis focuses on the 2018–2019 period and excludes later years affected by COVID-19, Storm Filomena, tariff changes, infrastructure expansion, and data inconsistencies.

Capacity constraints are not incorporated in the present framework. Because the model is fitted to completed trips, it characterizes realized demand patterns rather than unobserved suppressed demand. Ignoring bicycle and dock availability means that the generated profiles should not be used to estimate unmet demand directly. Instead, they provide a demand-side baseline that could be coupled with occupancy snapshots, dock-availability data, or rebalancing logs in future work to evaluate capacity-induced losses.

The station-popularity component also does not explicitly model spatial autocorrelation or behavioral substitution among neighboring stations. This limitation is particularly relevant for the station-unavailability experiment, because real users are more likely to substitute a closed or full station with nearby stations within the same catchment area[25,26]. To address this

issue without changing the scope of the framework, we added a spatial-sensitivity analysis comparing the original global-renormalization rule with a local K-nearest-neighbor redistribution heuristic. The results of this analysis are used to interpret the station-unavailability experiment as a stylized null redistribution scenario rather than as a full behavioral substitution model. Importantly, the comparison with KNN does not indicate that the global-renormalization rule is less informative. Rather, it shows that the proposed rule achieves a non-negligible level of local concentration without imposing local substitution mechanically. This is a relevant methodological property of the framework: local concentration emerges from the estimated probability structure, rather than being imposed through an exogenous choice of K.

Several additional limitations should also be acknowledged. Operational rebalancing activities, whereby operators redistribute bicycles among stations, are treated as exogenous and unobserved, even though these interventions directly affect the supply side and may confound demand patterns inferred from trip records alone. The absence of GPS trajectory data also precludes modeling within-trip heterogeneity or route-specific dynamics. Finally, while the binary precipitation indicator provides a useful first-order approximation of weather effects, finer-grained meteorological variables such as temperature, wind speed, and humidity, as well as their nonlinear interactions, would likely improve the representation of cycling propensity.

Future extensions could enrich the station-popularity component by incorporating socioeconomic indicators, temporal network features, or real-time operational signals. Time-varying components could also be added to capture evolving demand regimes more explicitly. Similarly, occupancy snapshots or other high-frequency operational data would allow the framework to be connected to supply constraints and capacity-induced losses. As systems expand, hierarchical Bayesian formulations or probabilistic programming frameworks may also increase flexibility while preserving interpretability.

Further research could also assess the transferability of the framework to other BSS networks and to dockless systems, where origin–destination constraints are weaker and spatial distributions are broader. Such settings may require alternative parametric families or spatial point processes. Another promising direction is the development of cross-system meta-distributions capable of capturing structural regularities across cities with different sizes and network topologies. More generally, integrating richer temporal, spatial, and operational information would help move from the present demand-side baseline toward more behaviorally realistic and capacity-aware extensions.

Overall, the proposed framework provides a focused and interpretable approach to probabilistic demand characterization in dock-based BSS. Its main contribution is not the use of Gamma or Negative Binomial distributions in isolation, but their calibrated integration with empirical station-popularity profiles into a demand-side generative structure. Applied to BiciMad, the framework reproduces key empirical regularities in trip distances and contextual hourly demand variability, provides calibrated prediction intervals, and supports the identification of data inconsistencies in operational records. It should not be interpreted as a full operational simulator. Rather, it offers a low-complexity probabilistic baseline for synthetic demand generation, exploratory analysis, and data-quality consistency checking under explicit assumptions.

Methods

Case Study

Madrid, the capital of Spain, with approximately 3.3 million residents [27] and a metropolitan area of 604.45 km^2, has faced increasing challenges related to congestion, air pollution, and mobility inefficiencies. In response, the city launched its public bicycle-sharing service, BiciMad, in 2014. After an initial phase of technical instability and vandalism, the service was municipalized in 2016, which improved its reliability and operational capacity. Since then, BiciMad has expanded significantly, growing from 175 stations and 2,000 bicycles to over 611 stations and 7,500 bikes as of 2024.

BiciMad provides an ideal case study for modeling demand in a dock-based bicycle-sharing system due to its open-access data policy, its large and heterogeneous user base, and the varying urban conditions across its service area.

Datasets and Data sources

The primary dataset comprises anonymized trip-level records from BiciMad, Madrid's dock-based bicycle-sharing system, covering the period from January 2018 to December 2019. Data were obtained from the open-data portal of Madrid's *Empresa Municipal de Transportes* (EMT)[28] and include trip identifiers, origin and destination station IDs, timestamps, user type, and total travel duration.

The analysis focuses on the 2018–2019 period to avoid atypical mobility patterns associated with the COVID-19 pandemic (2020–2021) and extreme weather events such as Storm Filomena. More recent data (2022–2025) were excluded due to multiple inconsistencies, including intermittent data unavailability, changes in station infrastructure and tariff structures (notably periods of free access), and regulatory updates affecting competing mobility services. Restricting the analysis to these two years ensures temporal stability and comparability across observations.

To maintain behavioral consistency, the analysis was restricted to trips made by annual-pass users (`user_type` = 1). The remaining user categories correspond either to operator-driven bicycle repositioning trips or to short-term users holding 1-, 3-, or 5-day tourist passes, both of which are associated with travel motivations and usage patterns that differ substantially

from those of regular city users. Including these groups would introduce additional behavioral heterogeneity into the fitted demand distributions and could therefore obscure the recurrent mobility patterns of interest; accordingly, they were excluded from the present analysis and would be more appropriately examined in a separate study. In addition, stations introduced after 2018 or removed before the end of 2019 were excluded, resulting in a stable network configuration across both years. After preprocessing and filtering, the final analytical dataset comprised 3,249,624 trips in 2018 and 3,648,755 trips in 2019.

Contextual information was incorporated from external sources. The official holiday calendar for the city of Madrid was retrieved from the Madrid City Council open-data portal, enabling the identification of weekdays, weekends, and public holidays. Meteorological data were obtained from the Spanish Meteorological Agency (AEMET), using daily precipitation measurements recorded at the Retiro weather station, which best represents conditions across the BiciMad service area. A binary precipitation indicator was defined as

$$\text{rain\_flag} = \begin{cases} 1, & \text{if daily precipitation} \geq 1 \text{ mm}, \\ 0, & \text{otherwise}. \end{cases}$$

A detailed description of all datasets and their original fields is provided in Appendix G. Also an explanatory pseudocode for the data extraction process of BiciMad trips is available in the same Appendix G.

Preprocessing and filtering

Several preprocessing steps were applied to ensure consistency:

- Removal of dummy testing stations (IDs 1000–1017, 2000–2017, 3000).
- Filtering of trips with identical origin and destination, which typically correspond to aborted rentals.
- Exclusion of trips with unrealistic duration:

$$\text{travel\_time} < 60 \text{ s} \quad \text{or} \quad \text{travel\_time} > 10800 \text{ s}.$$

- Aggregation of start times to hourly bins.

Two binary contextual variables were created:

$$\text{weekend\_holiday\_flag} = \begin{cases} 1, & \text{weekend or official holiday} \\ 0, & \text{weekday}, \end{cases} \qquad \text{rain\_flag as defined above.}$$

Transformations are summarized in Appendix H.

Distance computation between stations

Several methods exist for calculating the distance between two geographical points, which are defined by their latitude and longitude. A brief overview of the most commonly used approaches—including Euclidean, Manhattan, and Haversine distances is provided in Appendix I. Prior work has examined the suitability of these alternatives for geographic distance estimation in mobility studies (e.g.,[29]). In the context of bike-sharing systems, many studies ultimately adopt the Haversine distance after evaluating multiple candidates (see, for example,[30]), a choice that is also common in related transportation domains such as taxi mobility analysis[31].

In the absence of GPS trajectories, trip distances in this study were approximated using station coordinates. We adopted a hybrid Manhattan–Haversine method designed to better reflect urban street navigation while retaining the robustness of great-circle distance calculations. This approach captures the grid-like structure of city layouts—characteristic of Manhattan distance—while avoiding the distortions introduced by purely planar metrics.

Given an origin station with coordinates (x_1, y_1) and a destination station (x_2, y_2), the trip distance is computed as

$$d = d_H\big((x_1, y_1), (x_1, y_2)\big) + d_H\big((x_1, y_2), (x_2, y_2)\big),$$

where $d_H(\cdot, \cdot)$ denotes the Haversine distance. The Haversine formula is defined as

$$d_H(p, q) = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{y_1 - y_2}{2}\right) + \cos(y_1)\cos(y_2)\sin^2\left(\frac{x_1 - x_2}{2}\right)}\right),$$

with Earth radius $R = 6371$ km.

Conceptually, this method introduces an intermediate (imaginary) point that shares one coordinate with the origin and the other with the destination, effectively decomposing the trip into two orthogonal segments. Unlike a classical Manhattan metric, each segment is evaluated using the Haversine distance, thereby accounting for Earth curvature. A schematic illustration of this distance definition is shown in Fig. 6.

Figure 6. Distance calculation based on a mix of Haversine and Mahattan

A direct consequence of this distance definition is the presence of a negligible asymmetry between distances computed from $i \to j$ and $j \to i$. As illustrated in Table 3, these differences are on the order of one to two centimeters and are therefore irrelevant for quantitative analysis. However, they provide a convenient mechanism to uniquely label directional origin–destination (OD) pairs in large-scale datasets without resorting to artificial perturbations. Compared to ad hoc alternatives—such as adding a small numerical offset to distinguish directions—the proposed method more faithfully reflects the nature of real bicycle circulation in urban environments.

Table 3. Example of the proposed distance metric for a pair of stations. Differences between opposite directions are on the order of 1–2 cm and are therefore quantitatively negligible, while still providing distinct directional labels for each trip.

Origin ID	Origin station	Destination ID	Destination station	Distance (m)
99	Biblioteca Nacional	110	Serrano	699.799
110	Serrano	99	Biblioteca Nacional	699.784

Fitting distance distributions

We tested several candidate distributions commonly used in mobility modeling, including Exponential, Lognormal, Weibull, Rayleigh, Normal, and Gamma. Parameters were estimated using maximum likelihood estimation via `scipy.stats`.

Goodness of fit was quantified using the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_x (P(x) + \varepsilon) \log \left(\frac{P(x) + \varepsilon}{Q(x) + \varepsilon} \right),$$

with small constant $\varepsilon = 10^{-5}$ to avoid singularities.

The Gamma distribution yielded the lowest divergence,

$$D_{\mathrm{KL}} \approx 4.5 \times 10^{-5},$$

and consistently outperformed all alternatives across temporal subsets. Hourly Gamma distributions were fitted to capture diurnal variation in parameters.

Modeling hourly trip counts

Hourly trip counts exhibit strong overdispersion, for which the Poisson model is inadequate. We therefore used the Negative Binomial (NegBin) distribution. For empirical mean μ and variance σ^2, the parameters are:

$$p = \frac{\mu}{\sigma^2}, \qquad n = \frac{\mu^2}{\sigma^2 - \mu}.$$

The probability mass function (PMF) is:

$$P(K = k) = \binom{k+n-1}{k} p^n (1-p)^k, \qquad k = 0, 1, 2, \ldots$$

Separate distributions were fitted for each hour and for each combination of contextual conditions:

$$(\text{weekday}, \text{weekend/holiday}) \times (\text{rain}, \text{no rain}).$$

Probabilistic generative framework

The generative simulator integrates three probabilistic components:

1. **Hourly trip count generation**: For each hour h, sample
   $$T_h \sim \text{NegBin}(n_h, p_h).$$
2. **Origin assignment**: Origins are sampled from an empirical station popularity distribution
   $$P(\text{origin} = i) = \pi_i.$$
3. **Destination sampling**: Given origin i, we sample a distance
   $$d \sim \Gamma(\alpha_h, \beta_h),$$
   and map it to the nearest feasible OD pair (i, j) in the empirical set.

For the station-unavailability application, we also assessed the sensitivity of the redistribution rule by comparing the original global-renormalization procedure with a local K-nearest-neighbor heuristic. This additional check is not part of the core generative model, but is used to evaluate how strongly the redistribution results depend on the assumed spatial-substitution rule. In particular, it helps distinguish spatial concentration that emerges from the Gamma-derived probability structure from spatial concentration that is imposed mechanically through a fixed neighborhood size. The analysis is described in detail in Appendix F.

The simulator generates synthetic datasets that match the statistical structure of observed BiciMad usage, while allowing for the exploration of counterfactual scenarios.

Validation procedures

Internal validation. For each hour and contextual condition, we computed 95% prediction intervals using the fitted NegBin models and verified empirical coverage:

$$\text{Coverage} = \frac{\#\{T_{\text{obs}} \in [\text{CI}_{2.5}, \text{CI}_{97.5}]\}}{\text{Total observations}}.$$

Coverage exceeded 94% across all conditions.

External validation. To assess external consistency, 2019 data restricted to the set of stations already operating in 2018 were compared with the model predictions. This analysis revealed systematic deviations concentrated around 6–8 AM. Once a +2-hour correction was applied, these discrepancies largely disappeared; EMT Madrid later confirmed that they were caused by timestamp misassignments associated with system regularizations and daylight-saving adjustments. This example shows that the probabilistic framework can provide a useful baseline for consistency checking in operational datasets, complementing descriptive inspection of the data. However, deviations between simulated and observed demand should not be interpreted exclusively as anomalies, since they may also reflect structural differences between years that are not explicitly represented in the model. In particular, because the framework assumes conditional stability within the modeled contexts, it does not capture longer-term changes such as evolving system adoption, shifts in user composition, or network and operational evolution.

References

1. Ho, C. Q. & Tirachini, A. Mobility-as-a-service and the role of multimodality in the sustainability of urban mobility in developing and developed countries. *Transp. Policy* **145**, 161–176 (2024).
2. Zhao, K., Tarkoma, S., Liu, S. & Vo, H. Urban human mobility data mining: An overview. *Proc. - 2016 IEEE Int. Conf. on Big Data, Big Data 2016* 1911–1920, DOI: 10.1109/BigData.2016.7840811 (2016).
3. DeMaio, P. Bike-sharing: HDeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. Journal of Public Transportation, 12, 41–56. *J. Public Transp.* **12**, 41–56, DOI: 10.1016/0965-8564(93)90040-R (2009).
4. O'Brien, O. Web: BIKE SHARE MAP. Global Map of Bikeshare by OOMap. Accessed : September 2024. https://bikesharemap.com (2024).
5. Vallez, C. M., Castro, M. & Contreras, D. Challenges and opportunities in dock-based bike-sharing rebalancing: A systematic review. *Sustain. (Switzerland)* **13**, 1–26, DOI: 10.3390/su13041829 (2021).
6. Koç Ç, T. I., Laporte G. A review of vehicle routing with simultaneous pickup and delivery. *Comput. Oper. Res.* **122**, DOI: 10.1016/j.cor.2020.104987 (2020).
7. Louail, T. *et al.* Uncovering the spatial structure of mobility networks. *Nat. Commun.* **6**, DOI: 10.1038/ncomms7007 (2015). 1501.05269.
8. Barbosa, H. *et al.* Human mobility: Models and applications. *Phys. Reports* **734**, 1–74, DOI: 10.1016/j.physrep.2018.01.001 (2018). 1710.00004.
9. Anderson, J. E. The gravity model. *Annu. Rev. Econ.* **3**, 133–160 (2011).
10. Rong, C., Ding, J. & Li, Y. An interdisciplinary survey on origin-destination flows modeling: Theory and techniques. *ACM Comput. Surv.* **57**, 1–49 (2024).
11. Zhang, R., Kan, H., Wang, Z. & Liu, Z. Relocation-related problems in vehicle sharing systems: A literature review. *Comput. & Ind. Eng.* 109504 (2023).
12. Lim, H., Chung, K. & Lee, S. Probabilistic Forecasting for Demand of a Bike-Sharing Service Using a Deep-Learning Approach. *Sustain. (Switzerland)* **14**, 1–18, DOI: 10.3390/su142315889 (2022).
13. Albuquerque, V., Dias, M. S. & Bacao, F. Machine learning approaches to bike-sharing systems: A systematic literature review. *ISPRS Int. J. Geo-Information* **10**, DOI: 10.3390/ijgi10020062 (2021).
14. Liang, M., Yang, L., Li, K. & Zhai, H. Improved collaborative filtering for cross-store demand forecasting. *Comput. & Ind. Eng.* **190**, 110067 (2024).
15. Alhussam, M. I., Ren, J., Yan, P., Risha, O. A. & Alhussam, M. A. Data-driven approach based on hidden markov model for detecting the status of bikes in bike-sharing systems. *Comput. & Ind. Eng.* **196**, 110470 (2024).
16. Ciociola, A., Cocca, M., Giordano, D., Vassio, L. & Mellia, M. E-scooter sharing: Leveraging open data for system design. In *2020 IEEE/ACM 24th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*, 1–8 (IEEE, 2020).
17. Jiménez-Meroño, E. & Soriguera, F. Agent-based simulation of vehicle-sharing systems. *J. Simul.* **19**, 84–107 (2025).
18. Wang, F., Yin, C., Chang, X., Zhang, X. & He, Z. Exploring the relationship between built environment and bike-sharing demand: Does the trip length matter? *Transp. Plan. Technol.* **48**, 826–852, DOI: 10.1080/03081060.2024.2400281 (2025). https://doi.org/10.1080/03081060.2024.2400281.
19. Peng, T., Gao, J. & Cats, O. Uncertainty-aware probabilistic travel demand prediction for mobility-on-demand services. *Transp. Res. Part C: Emerg. Technol.* **181**, 105383, DOI: https://doi.org/10.1016/j.trc.2025.105383 (2025).
20. Anderson, T. W. & Darling, D. A. Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *Annals Math. Stat.* **23**, 193–212, DOI: 10.1214/aoms/1177729437 (1952).
21. Sprent, P. & Smeeton, N. C. *Applied nonparametric statistical methods* (CRC press, 2016).
22. Kou, Z. & Cai, H. Understanding bike sharing travel patterns: An analysis of trip data from eight cities. *Phys. A: Stat. Mech. its Appl.* **515**, 785–797 (2019).
23. Jaynes, E. T. *Probability theory: The logic of science* (Cambridge university press, 2003).
24. Dehaene, S. The neural basis of the weber–fechner law: a logarithmic mental number line. *Trends Cogn. Sci.* **7**, 145–147, DOI: https://doi.org/10.1016/S1364-6613(03)00055-X (2003).

25. Rudloff, C. & Lackner, B. Modeling demand for bikesharing systems: neighboring stations as source for demand and reason for structural breaks. *Transp. Res. Rec.* **2430**, 1–11 (2014).

26. Faghih-Imani, A. & Eluru, N. Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of new york citibike system. *J. transport geography* **54**, 218–227 (2016).

27. Functional urban areas. https://ec.europa.eu/eurostat/web/products-datasets/-/urb_ltran (2024). Accessed: 2024-04-24.

28. EMT Madrid . Bicimad. datos de los itinerarios de las bicicletas eléctricas. https://datos.emtmadrid.es/dataset/historicos-de-bicimad-2017_2023 (accessed on 10 Jan 2023).

29. Ridwan, M., Cahyono, A., Mariza, I. & Wirawan, S. Electric Bike Monitoring and Controlling System Based on Internet of Things. *J. Nasional Tek. Elektro dan Teknologi Informasi* | **11**, 53–60 (2022).

30. Li, Z. *et al.* Large-scale trip planning for bike-sharing systems. *Pervasive Mob. Comput.* **54**, 16–28, DOI: 10.1016/j.pmcj.2019.01.007 (2019).

31. Poongodi, M. *et al.* New York City taxi trip duration prediction using MLP and XGBoost. *Int. J. Syst. Assur. Eng. Manag.* **13**, 16–27, DOI: 10.1007/s13198-021-01130-x (2022).

32. Vallez, C. M., Castro, M. & Contreras, D. List of Negative binomial Params. https://github.com/cvallez/ProbabilisticMLModel.git (2025).

33. Government Area of Finance and Personnel of the Madrid City Council]. Working calendar. https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=9f710c96da3f9510VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default (accessed on accessed on 02 Feb 2023).

34. AEMET [Meteorological Spanish Agency]. Open data api. https://www.aemet.es/es/datos_abiertos/AEMET_OpenData (accessed on accessed on 13 Feb 2023).

Data availability

This study relies exclusively on publicly available third-party datasets. Trip-level data from Madrid's dock-based bicycle-sharing system (BiciMad) are available from the Madrid City Council open data portal at `https://datos.emtmadrid.es/dataset/historicos-de-bicimad-2017_2023`. Meteorological data were obtained from the official public meteorological service at `https://www.aemet.es/es/datos_abiertos/AEMET_OpenData`. Official public holidays data are available from Madrid City Council open data portal.. No new raw data were generated by the authors. Model parameters inferred during the study are available from the corresponding author upon reasonable request. A public GitHub repository `https://github.com/cvallez/ProbabilisticMLModel.git` is provided to centralize dataset references and document the data acquisition and preprocessing workflow.

Code availability

The code used for preprocessing, distribution fitting, probabilistic validation, and synthetic trip generation will be made publicly available upon acceptance at GitHub repository `https://github.com/cvallez/ProbabilisticMLModel.git`

Fundings

This work has been partially supported by Grant PID2022-140217NB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and, by "ERDF/EU A way of making Europe".

Author contributions statement

C.M.V. Conceptualization, Methodology, Software, Data curation, Writing – Original Draft , D.C. Conceptualization, Methodology, Writing – Review and Editing and M.C. Conceptualization, Methodology, Software, Data curation, Writing – Original Draft

Additional information

Appendix A: KL Divergence Computation and Distributional Comparison

Once the KL-divergence methodology was established, the fitted values enabled a quantitative comparison of how different theoretical distributions approximate the empirical trip–distance data. Continuous distributions were fitted using the scipy.stats

library, where models are typically parameterized by a shape and a scale parameter governing skewness and dispersion, respectively. Table 4 summarizes the estimated parameters and the corresponding KL divergence values for the 2018 BiciMad distance distribution, computed without additional temporal stratification.

Distribution	Fitted parameters	KL value
Gamma	Shape = 3.44; Scale = 714.58	**4.5e-05**
Lognormal	Shape = 0.40; Scale = 2975.81	7.9e-05
Weibull	Shape = 1.97; Scale = 2761.65	1.5e-04
Rayleigh	Shape = 2 (fixed); Scale = 1960.28	1.5e-04
Normal	Mean = 2440.80; Std. Dev. = 1314.51	8.6e-04
Exponential	Scale = 2336.19; $\lambda = 4.3e04$	2.4e-03

Table 4. Fitted parameters and KL divergence values for the tested probability distributions, ordered from best (lowest KL) to poorest fit.

The Gamma distribution achieves the smallest KL divergence, indicating the closest agreement with the empirical data. The Lognormal distribution also provides a strong fit, differing mainly in the tails of the distribution. Weibull and Rayleigh models capture some features but with reduced accuracy, and Rayleigh is expectedly limited as a special case of Weibull. The Normal distribution fails to capture the strong right skew of the data, while the Exponential distribution—whose mode lies at zero—cannot reproduce the observed peak near 2000 m, resulting in substantially larger KL divergence.

Appendix B: NegBin Distributions

The NegBin Distributions produced for the different hours in the Non-rainy and working day scenarios are shown in Figure. 7. Note how the fitting follows very accurately the underlying empirical distributions. One potential reason behind this high accuracy lies in the fact that the NegBin distribution can be seen as a mixture (related, hypothetically, to individual preferences) of Poisson distributions, being that the latter is the maximum entropy distribution for counts.

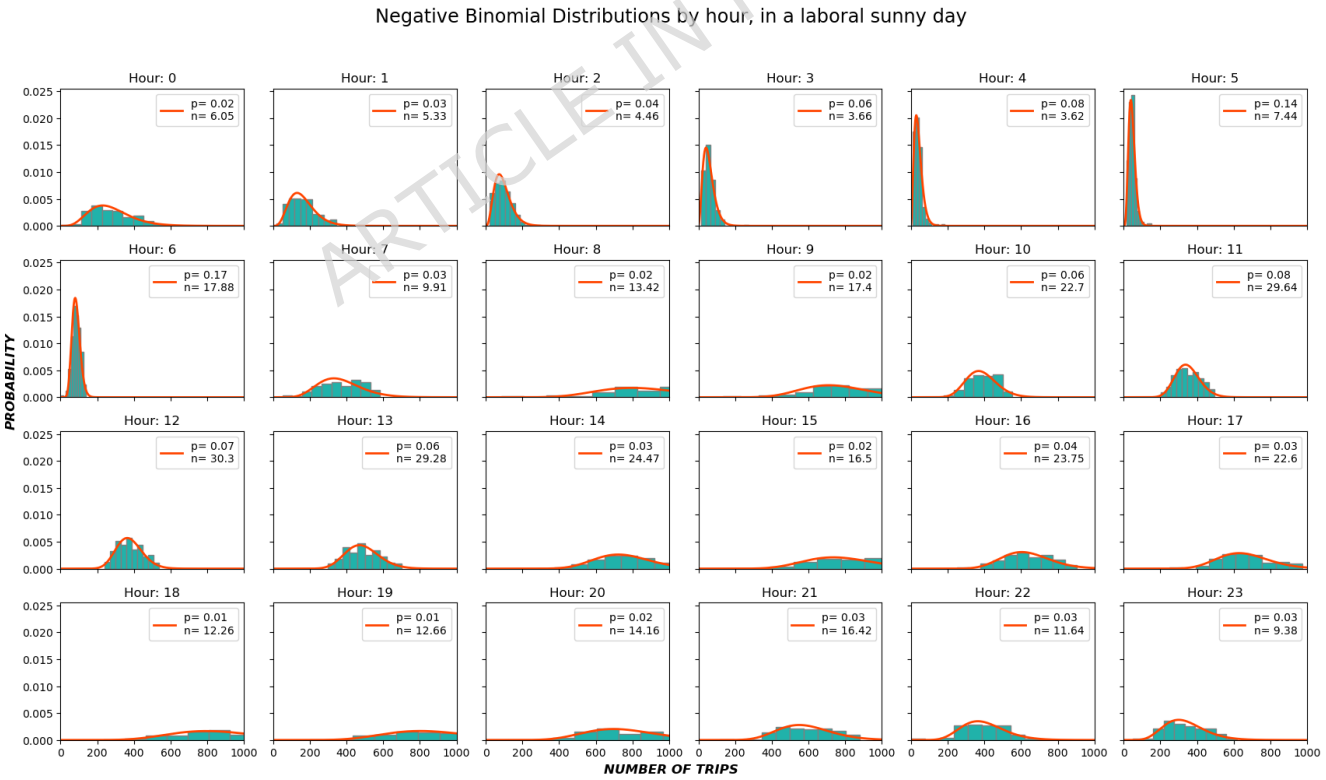


Figure 7. NegBin parameters and charts for sunny and working days by hour. Note how accurately the distribution fits the underlying empirical histograms.

Table 5 displays a reduced set of the different values for the p and n parameters of the NegBin distribution associated

with the particular scenario of time, day type, and precipitation indicator. For a detailed values table please refer to GitHub repository[32]

Table 5. Example subset of the fitted Negative Binomial parameters (p and n) across hours, stratified by day type and precipitation.

Weekend/Holiday	Rain	Hour	p	n	Mean	Var	Std
0	0	0	0.02	6.05	275.95	12864.55	113.42
0	0	1	0.03	5.33	161.05	5025.52	70.89
0	0	2	0.04	4.46	95.14	2126.61	46.12
			...				
0	1	6	0.14	12.23	74.00	521.63	22.84
0	1	7	0.02	7.41	291.02	11725.02	108.28
			...				
1	0	15	0.05	16.70	336.98	7137.17	84.48
1	0	16	0.04	17.09	377.62	8719.89	93.38
			...				
1	1	21	0.01	4.14	290.46	20668.82	143.77
1	1	22	0.03	5.70	215.43	8358.46	91.42
1	1	23	0.02	4.44	185.39	7932.24	89.06

Appendix C: CSV file with the intervals information

To create the Dumbells chart with the intervals, it was necessary to previously create the file `validationoutput.csv`, which can be found at the GitHub link[32].

Each row contains the following columns:

- hour: Hour of the day (0-23).
- weekend_holiday_flag: Indicates whether the day is a weekend/holiday (1) or not (0).
- rain_flag: Indicates whether it is raining (1) or not (0).
- percentage: Percentage of values that fall within the confidence interval of the negative binomial distribution.
- numhi: Number of values exceeding the upper limit of the interval.
- numlow: Number of values below the lower limit of the interval.
- numtotal: Total number of values considered.
- rangehi: Upper limit of the confidence interval.
- rangelow: Lower limit of the confidence interval.

Appendix D: Boxplot of distribution of the number of trips grouped by hour of the day

The internal validation of the probabilistic framework was followed by an external evaluation using 2019 data. Because BiciMad expanded its infrastructure in 2019, the dataset was filtered to retain only trips whose origin and destination corresponded to stations already present in 2018. A direct comparison revealed marked discrepancies during early morning hours (06:00–08:00), where 2019 observations frequently fell outside the empirical range established in 2018. A detailed trip-level and hour-level inspection identified 107 anomalous days, predominantly concentrated between 27 November and 20 December 2019.

Further investigation showed that applying a uniform +2-hour shift to the timestamps of these days brought the 2019 distributions into alignment with those of 2018, with all hourly counts falling within expected ranges.

This strongly suggested a systematic time misattribution. The anomaly was later confirmed by BiciMad's data administrators, who reported that docking-related adjustments and daylight-saving transitions can induce timestamp irregularities. As stated in their public response (9 September 2021):

> *"Sometimes, regularizations are applied in the BiciMad system due to customer service actions. As a result, if an adjustment is made during docking or undocking, these types of effects may occur. This can also happen during seasonal daylight saving time changes when the hour is shifted."*

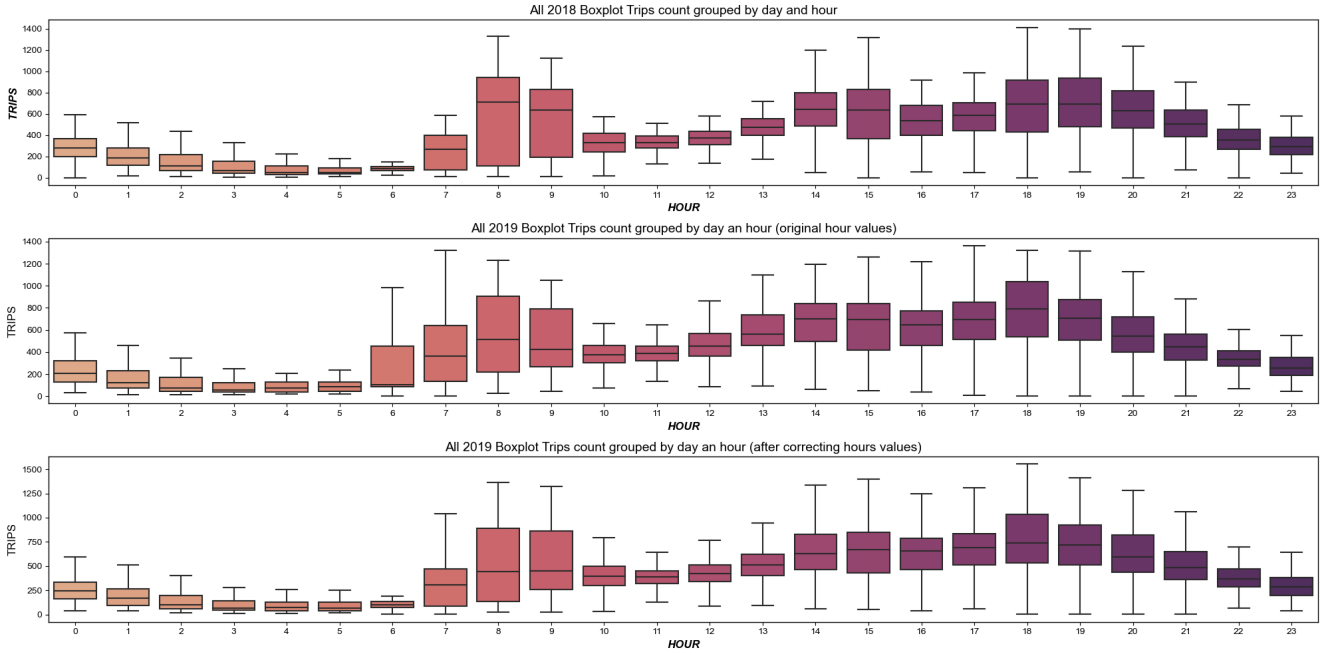


Figure 8. Distribution of the number of trips per hour in 2018 (top), 2019 (middle), and 2019 after correcting misattributed data (bottom). KL comparison of boxplots in the range of 6 and 8 a.m. helped us to identify huge changes in those hours. After contacting the data provider, we confirmed that part of the data (after the October daylight saving time), was wrongly labeled. In the bottom figure, we replot it after this correction, and the distribution matches closely that of 2018, in the same fashion as the distribution of distances in 2018 and 2019 are almost identical.

This unanticipated finding highlights an additional practical utility of our probabilistic framework: beyond demand modeling, it provides a systematic mechanism for detecting data inconsistencies when observed values fall outside the theoretically expected ranges.

Appendix E: Schematic representation of the simulator

The diagram in Fig. 9 describes the simulator's components in schematic form. The processes are then outlined and explained one by one.

Description of the simulator steps:

- Step 1: The initial step of the procedure is to iterate through each of the hours that make up a day.
- Step 2: For each hour of the day, we generate four possible scenarios that we used to calculate the parameters of the negative binomial distributions. These 4 scenarios are the combination of rainy or non-rainy days with working days or holidays/weekends.
- Step 3 Starting with the current time and scenario, we look at Table 5 to get the values of the parameters p and n. These parameters define the negative binomial distribution for that hour and scenario, which represents the number of trips. We request a random integer based on that distribution. We assume that the generated number represents the number of trips that would be made on a hypothetical day at the given time and under the provided situations. If we performed this process multiple times, we would get different numbers of trips, but they would all follow the same binomial distribution, resulting in distinct counterfactual days corresponding to hypothetical situations that may have happened.
- Step 4: The `id` of the origin station is calculated for each of the trips created in the previous stage, based on a popularity function derived from the empirical data.
- Step 5: We look for the gamma distribution linked with the trip's time and choose only the distances related to trips whose origin is the station `id` determined in the previous step using the popularity function. We normalize the probabilities of this set of possible distances (all combinations of trips starting at the origin station). generate a random number according to that normalized probability distribution, thus obtaining a distance from which the destination station is derived.

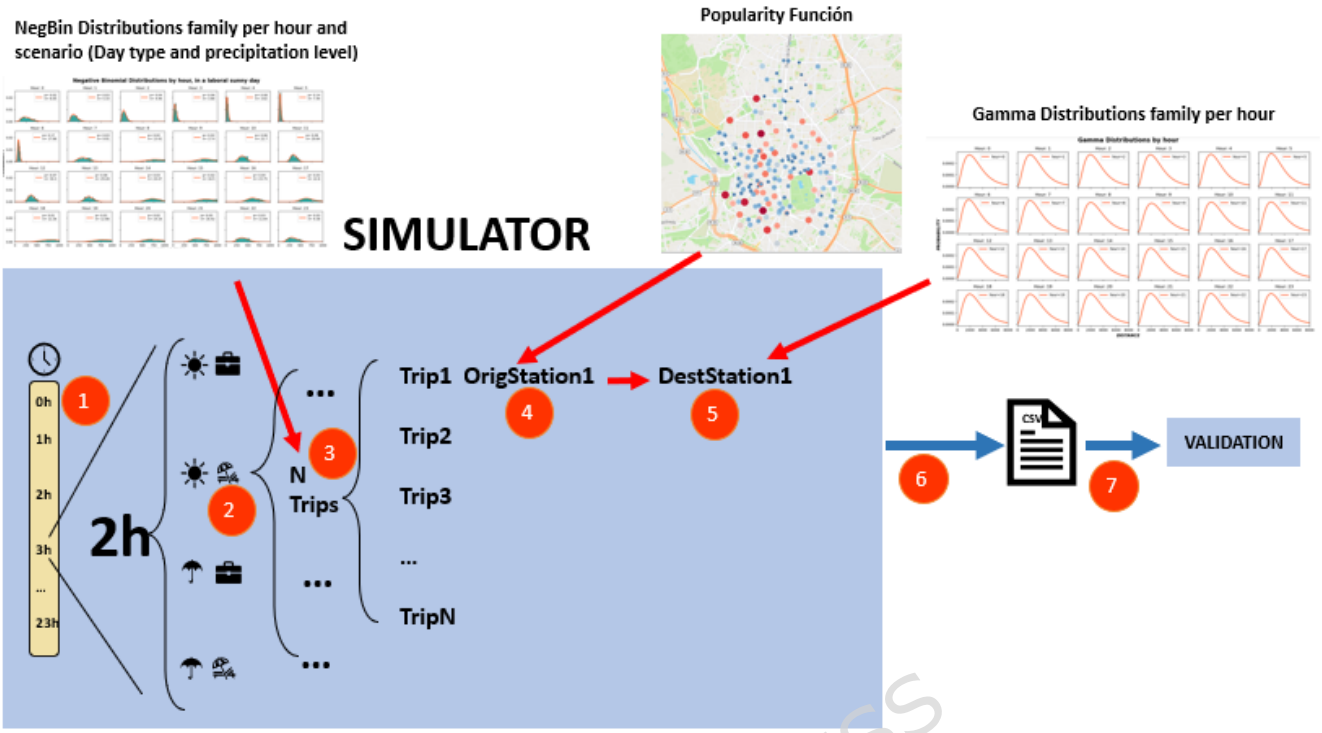


Figure 9. Diagram showing the constituent elements of our probabilistic simulator: given (1) the desired hour; (2) the weather and labor conditions; (3) we estimate the (stochastic) number of trips; (4) and generate different trips according to the station popularity to a random (5) destination. Finally, we (6) collect the data and (7) perform some validation tests.

- Step 6: By combining all of the data in a csv file, we can generate counterfactual scenarios that correspond to the 2018 empirical distribution.
- Step 7: Furthermore, we have verified internally the number of trips obtained each hour by comparing the generated data with the empirical data from 2018.

Appendix F: Detailed illustrative example of counterfactual analysis

Temporary station closures are operationally relevant scenarios in bike-sharing systems. They may occur because of roadworks, multi-day events, strikes, maintenance operations, or planned station relocations. To illustrate how the proposed probabilistic simulator can be used in this type of situation, we present a simplified station-unavailability use case.

We begin by selecting a subset of BiciMad stations (Step 1) and computing pairwise trip probabilities using the fitted Gamma distance distribution, excluding self-originating destinations. Because the Gamma distribution is continuous, probabilities are normalized so that only the selected stations form the feasible destination set.

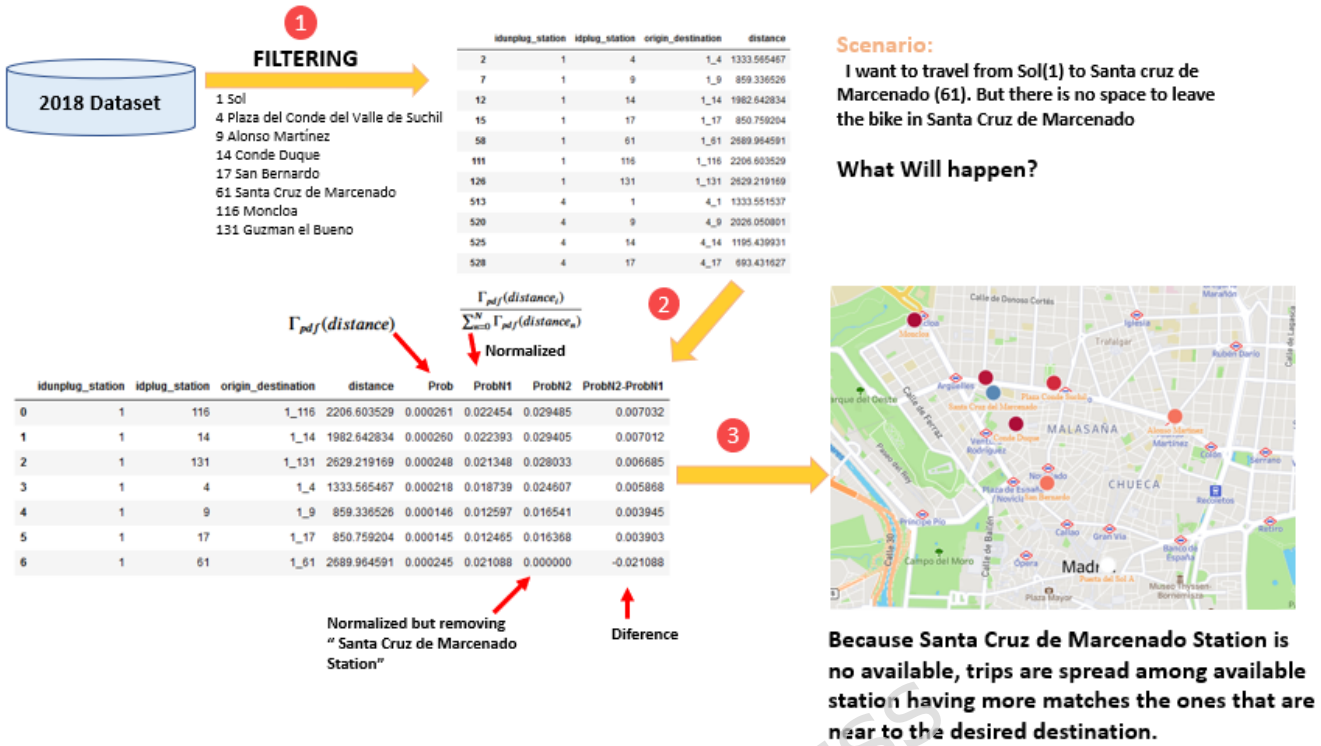
Next, we simulate the closure of station Santa Cruz de Marcenado by removing it from the OD set and re-normalizing the corresponding Gamma-derived probabilities (Step 2). Conceptually, this corresponds to a user departing from Sol station and attempting to end the trip at Santa Cruz de Marcenado, only to find no available docking points.

In Step 3, the simulator reallocates these trips according to the normalized Gamma-based probabilities, redistributing demand among the remaining stations.

Under this original rule, the simulator reallocates the affected trips according to the normalized Gamma-based probabilities among the remaining feasible stations. This provides a useful first approximation of the redistribution pressure induced by the unavailable station. However, the resulting pattern should be interpreted with care. The procedure does not model user-level search behavior, station capacity, bicycle or dock availability, or explicit local substitution among neighboring stations. It is therefore best understood as a null redistribution scenario rather than as a full behavioral prediction of how users would adapt in practice.

Spatial sensitivity analysis for station-unavailability scenarios

The station-unavailability experiment described above uses a global renormalization rule: when a destination station is removed from the feasible OD set, its probability is set to zero and the remaining Gamma-derived probabilities are renormalized across



	idunplug_station	idplug_station	origin_destination	distance
2	1	4	1_4	1333.565467
7	1	9	1_9	859.336526
12	1	14	1_14	1982.642834
15	1	17	1_17	850.759204
58	1	61	1_61	2689.964591
111	1	116	1_116	2206.603529
126	1	131	1_131	2629.219169
513	4	1	4_1	1333.551537
520	4	9	4_9	2026.050801
525	4	14	4_14	1195.439931
528	4	17	4_17	693.431627

	idunplug_station	idplug_station	origin_destination	distance	Prob	ProbN1	ProbN2	ProbN2-ProbN1
0	1	116	1_116	2206.603529	0.000261	0.022454	0.029485	0.007032
1	1	14	1_14	1982.642834	0.000260	0.022393	0.029405	0.007012
2	1	131	1_131	2629.219169	0.000248	0.021348	0.028033	0.006685
3	1	4	1_4	1333.565467	0.000218	0.018739	0.024607	0.005868
4	1	9	1_9	859.336526	0.000146	0.012597	0.016541	0.003945
5	1	17	1_17	850.759204	0.000145	0.012465	0.016368	0.003903
6	1	61	1_61	2689.964591	0.000245	0.021088	0.000000	-0.021088

Figure 10. Application 1 of the simulator: Distribution of trips after missing stations. In this scenario, we generate counterfactual situations in which a station is removed (or inaccessible for technical reasons) to the users. Using the simulator we can estimate the *pressure* induced to other stations that collect the trips normally allocated to that missing station.

all feasible destinations. This rule is simple and consistent with the distance-based structure of the simulator, but it may dilute local substitution effects. In a real dock-based BSS, users affected by an unavailable destination station are more likely to search for nearby alternatives within the same catchment area.

To assess how sensitive the results are to this assumption, we compare the original global-renormalization rule with a local K-nearest-neighbor (KNN) redistribution heuristic. In the KNN variant, the probability mass associated with the unavailable station is reassigned only to its K nearest neighboring stations. This represents a simple local substitution assumption. The procedure is not intended to model user behavior in full detail, and it does not incorporate station capacity or bicycle availability. Its purpose is more limited: to quantify how much the redistribution results depend on assuming global rather than local redistribution of displaced demand.

We evaluate the KNN rule for $K = 3$, $K = 5$, and $K = 10$, subject to the feasible destination set available in the selected station-unavailability scenario. When fewer than K feasible neighboring stations are available, we report the effective number of neighbors used, denoted as K_{eff}. For each case, we report three diagnostic quantities: the local absorption ratio, defined as the share of displaced probability assigned to the nearest feasible neighbors; the mean substitution distance, defined as the expected distance between the unavailable station and the stations receiving redistributed demand; and the set of top receiving stations.

Interpretation of the sensitivity analysis. The sensitivity analysis highlights an important difference between the proposed global-renormalization rule and the KNN-based alternatives. The global-renormalization rule allows the displaced probability mass to be redistributed over the full feasible destination set, rather than restricting it a priori to a fixed local neighborhood. Nevertheless, 56.8% of the displaced mass is assigned to the three nearest feasible stations. This is a relevant result, because it shows that the baseline redistribution mechanism preserves a meaningful degree of spatial concentration without imposing a nearest-neighbor constraint by construction.

By contrast, the KNN rules achieve local concentration through an explicit restriction of the redistribution support. For $K = 3$, the mass assigned to the three nearest stations is equal to 1.000 by construction, since all displaced probability is forced to remain within those three stations. When the KNN neighborhood is expanded, the share assigned to the same three-station reference set decreases to 0.600 for $K = 5$ and to 0.500 for $K = 10$ with $K_{\text{eff}} = 6$. At the same time, the mean substitution distance increases from 311.6 m to 527.0 m and 683.7 m, respectively.

These results suggest that the global-renormalization rule provides a useful and less restrictive null redistribution model.

Table 6. Sensitivity of the station-unavailability experiment to the redistribution rule.

Redistribution rule	Redistribution support	Mass assigned to 3 nearest stations	Mean substitution distance	Top receiving stations
Global renormalization	All feasible stations	0.568	611.3 m	116 – Moncloa; 14 – Conde Duque; 131 – Guzmán el Bueno
KNN, K=3	3 nearest feasible stations	1.000	311.6 m	4 – Plaza Conde Suchil; 14 – Conde Duque; 131 – Guzmán el Bueno
KNN, K=5	5 nearest feasible stations	0.600	527.0 m	14 – Conde Duque; 116 – Moncloa; 17 – San Bernardo
KNN, K=10 ($K_{\text{eff}} = 6$)	6 nearest feasible stations	0.500	683.7 m	4 – Plaza Conde Suchil; 14 – Conde Duque; 131 – Guzmán el Bueno

Notes: The column "Mass assigned to 3 nearest stations" reports a common local-concentration diagnostic across all redistribution rules: the share of displaced probability mass assigned to the 3 nearest feasible stations to the unavailable station. Under global renormalization, redistribution is allowed over the full feasible destination set; therefore, the reported value indicates how much of the displaced mass is nevertheless absorbed by the nearest feasible stations. For KNN with $K = 3$, this value is 1.000 by construction, since the redistribution support coincides exactly with the 3 nearest feasible stations. For KNN rules with $K > 3$, the value is recomputed over the same 3-station reference set and is therefore not mechanically equal to one. K_{eff} denotes the effective number of feasible neighbors when fewer than K destinations are available. This diagnostic should be interpreted as a measure of local concentration, not as a measure of predictive accuracy.

Unlike KNN, it does not require the analyst to impose an arbitrary neighborhood size, yet it still produces a spatially plausible redistribution pattern. Therefore, its main advantage is not that it maximizes local absorption, but that it balances flexibility and spatial coherence: displaced demand can be reallocated across the feasible network while remaining substantially concentrated around geographically plausible substitute stations.

Appendix G: Datasets

In this study we have used datasets from several sources:

1. TRIPS DATASET: The following text outlines trips made using the Bicimad service, obtained from the Open Data website of the "Empresa Municipal de Transportes de Madrid" (EMT Madrid). This organization is responsible for planning public urban transportation in Madrid. For our model training, we will use data from 2018 and validate it using data from 2019. We chose these years because they are the only ones with complete information available from 2018 to 2022. We opted not to use data from the other years due to various disruptions. The COVID-19 pandemic significantly impacted 2020 data, as the service was suspended during periods of confinement and experienced reduced mobility. Although service levels were gradually restored, this situation resulted in data anomalies that could not be effectively modeled without risking overfitting. In 2021, the effects of COVID-19 from the previous year continued to affect the data, compounded by the disruptions caused by the snowstorm "Filomena." Lastly, we decided against using data beyond 2022 due to changes in service management, including structural adjustments, format alterations, and issues with data file availability. It is important to extract information about Bicimad stations in addition to trip information. Specifically, that which connects their unique identifiers to the latitude and longitude of their geographical position.

   The following Table 7 provides an overview of the qualitative and quantitative properties of the original files available in the data source.

Table 7. Qualitative and quantitative properties of the original files.

Field concept	Value
Available data	April 2017 to June 2021
Number of files	17 files (monthly; periods 07–12/2019 and 01–05/2020 are each consolidated into a single file)
File format	ZIP archive containing a JSON file
Max JSON file size (monthly split)	1,094,440 KB (09/2018)
Min JSON file size (monthly split)	3,363 KB (04/2020)
Average JSON file size (monthly split)	336,906 KB (across 51 files)

   Table 8 shows a summarizing of available fields in initial files.

2. HOLIDAYS DATASET: This dataset contains the working calendar for the city of Madrid (extracted from the Madrid City Council's open data portal[33]), which includes municipal, regional, and national holidays affecting the city.

Table 8. Original source data fields from the BiciMad trip dataset.

Field name	Description	Field type
_id	Unique trip identifier.	Alphanumeric string
user_day_code	Identifier shared by all trips made by the same user on a given date, enabling analysis of daily user activity patterns.	Alphanumeric string
idunplug_station	Identifier of the station where the bicycle is released and the trip begins.	Integer
idunplug_base	Identifier of the docking point at the origin station.	Integer
idplug_station	Identifier of the station where the bicycle is docked and the trip ends.	Integer
idplug_base	Identifier of the docking point at the destination station.	Integer
user_type	Encodes the type of user performing the trip: 0 (unknown), 1 (annual subscriber), 2 (occasional/tourist), 3 (company employee), 6 (occasional non-subscriber), or 7 (BiciMAD GO dockless user).	Integer
ageRange	Encodes the user age range: 0 (unknown), 1 (0–16), 2 (17–18), 3 (19–26), 4 (27–40), 5 (41–65), 6 (66+).	Integer
travel_time	Total trip duration in seconds between bicycle release and docking.	Integer
zip_code	Postal code of the user associated with the trip.	Integer

Table 9. Original source data fields from the holidays dataset.

Field name	Description	Field type
Dia	Calendar date (day, month, year)	Date (dd/MM/YYYY)
Dia semana	Name of the weekday	Alphanumeric string
Laborable / festivo / domingo festivo	Indicates weekday, holiday, or holiday weekend	Alphanumeric string
Tipo de Festivo	National or regional holiday indicator	Alphanumeric string
Festividad	Name or reason for the holiday	Alphanumeric string

3. WEATHER DATASET: To include weather conditions in our model, we retrieved weather data from the Spanish Meteorological Agency's API (AEMET)[34]. We chose the measurements produced by the "Retiro" station from among various measurement stations available in Madrid's metropolitan region. Given the placement of Bicimad stations in 2018 and 2019, it is the weather station that best depicts the weather circumstances experienced by service customers. We use the date ("Fecha") and rainfall ("Prec" expressed in mm of rainfall) information from the api's accessible fields.

Algorithm 1 Data extraction process

$sourcehtml \leftarrow https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)$
for $i \leftarrow line0, linen$ **do**
  if $i.labeltype$ is 'a' and i.href=True **then**
    if "getattachment" and ("usage" or "movements" in i.hrefcontent **then**
      listlinks.add(i) ▷ list of url liks that references to movements data files
    end if
  end if
end for
for $movementdatalink \leftarrow listlinks[]$ **do**
  $zipfile \leftarrow request.geturl(movementdatalink)$
  $jsonfile \leftarrow unzip(zipfile)$
  $dataframe \leftarrow read_json(jsonfile)$
  csvfile=ProcessDataFrameToCSV(dataframe)
  WriteCSV(csvfile)
end for

Appendix H: Transformations

The table below summarizes the transformations made to the raw data available in original extracted csv files. .

Table 10. Summary of data transformations.

Id	Transformation	Fields involved	Description / comments
1	Remove dummy stations	idunplug_station, idplug_station	According to EMT documentation, the released dataset includes records generated during system testing and network expansion. Trips involving "laboratory" stations (IDs 1000, 1001, 2000–2017, and 3000) were removed.
2	Filter by user type	user_type	Only annual-pass users (user_type = 1) were retained to ensure behavioral consistency with the scope of this study.
3	Remove trips with identical origin and destination	idunplug_station, idplug_station	Trips with the same origin and destination were excluded. These typically correspond either to aborted rentals (e.g., undocking failures followed by immediate re-docking) or to recreational round trips returning to the starting station.
4	Remove trips with invalid duration	travel_time	Trips shorter than 60 s or longer than 3 h were filtered out as implausible operational records.
5	Compute trip distance	idunplug_station, idplug_station, station coordinates	Trip distances were computed using origin and destination station coordinates. As a hybrid distance metric is employed, the procedure is described in detail in Methods section.

Appendix I: Different methods to calculate distance between two geographical points

The following table 11 describes the most commonly used distance calculation methods.

Table 11. Distance metrics commonly used to compute geographic separation between two points. In this work, a hybrid Manhattan–Haversine distance is adopted to better approximate urban street layouts while introducing a negligible asymmetry between trajectories $i \to j$ and $j \to i$.

Method	Description and formulation
Euclidean	Straight-line distance between two points in Euclidean space, computed using the Pythagorean theorem. For two-dimensional coordinates (longitude x, latitude y): $$d_E(p,q) = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}.$$
Manhattan	Distance measured along orthogonal grid paths, also known as the taxicab metric: $$d_M(p,q) = \lvert x_1-x_2\rvert + \lvert y_1-y_2\rvert.$$ This metric approximates movement along rectilinear street networks.
Haversine	Great-circle distance between two points on the surface of a sphere: $$d_H(p,q) = 2R\arcsin\left(\sqrt{\sin^2\left(\frac{y_1-y_2}{2}\right)+\cos(y_1)\cos(y_2)\sin^2\left(\frac{x_1-x_2}{2}\right)}\right),$$ where $R = 6371$ km is the Earth's radius.