

Identifying and ranking high-demand areas in a bike-sharing system based on extreme value theory

Lixuan An^{a,*}, Bernard De Baets^a, Lisheng Jiang^{a,b}, Stijn Luca^a

^a*Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure links
653, B-9000, Ghent, Belgium*

^b*College of Management Science, Chengdu University of Technology, 610059, Chengdu, China*

Abstract

Bike-sharing systems provide a convenient and eco-friendly transportation mode in cities worldwide. The continued expansion of urban regions introduces significant challenges in planning docking infrastructure adaptively, which requires accurate identification of emerging high-demand areas. While recent studies have advanced mathematical modeling of demand patterns, they still face limitations in spatial partitioning and computational efficiency. In this work, we formulate the detection of irregular high-demand areas as a group anomaly detection task, and not only address these limitations but also achieve a key advancement over prior studies by enabling the quantitative ranking of candidate areas. We develop a probabilistic framework where candidate areas are characterized by a likelihood model that accounts for their intensity, spatial coverage, and the configuration of their surrounding areas. Unlike existing studies, our spatial partitioning leverages prior knowledge of the operational system, eliminating the need for fine-grained hyperparameter tuning. Furthermore, the model's analytical form enhances its interpretability and practical applicability, and its design ensures low computational complexity. We further conduct a high-level comparison with representative recent methods, demonstrating that our method achieves superior performance. Our approach is applied to nine years (2015–2023) of rental records from the Münchner Verkehrsgesellschaft bike-sharing system. The anomaly extent is quantified across candidate areas, supporting the identification and prioritization of sites for future docking stations, thereby contributing to more data-driven urban mobility planning.

Keywords: Bike-sharing systems, Extreme value theory, Group anomaly detection, Anomaly ranking, Docking station deployment

1. Introduction

Driven by environmental benefits and the evident enhancement of urban mobility, urban bike-sharing systems (BSSs) have rapidly expanded worldwide over the past decade (Reiss and Bogenberger, 2015; O'Brien et al., 2014; Zhu et al., 2024). BSSs have now entered their fourth generation,

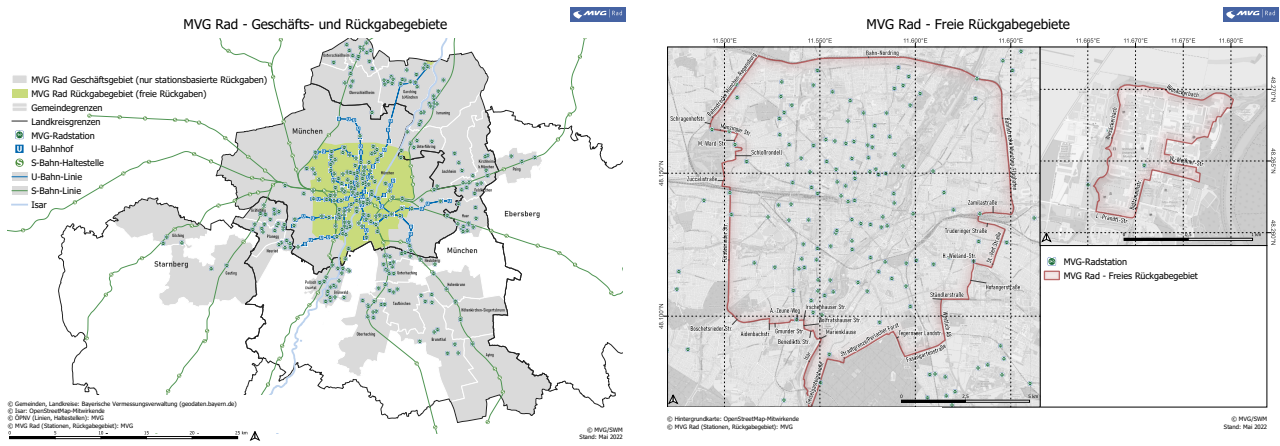
*Corresponding author. Email: lixuan.an@ugent.be

integrating mobile application support, real-time data, and interoperability with broader mobility networks (Duran-Rodas et al., 2020). These advancements build upon earlier stages, where the first generation introduced free-use bicycles (DeMaio, 2009), the second generation added coin-based access and fixed return stations to improve control (Shen et al., 2018), and the third generation incorporated digital systems enabling user identification and station-based tracking (Fishman, 2016; Du et al., 2019).

Building upon the aforementioned evolution of BSSs, two dominant operational systems have emerged: docking-based systems and free-floating (or dockless) systems (Zhuang et al., 2025). Docking-based systems were introduced during the early stages of BSSs, with users picking up and returning bikes at designated docking stations. With the development of technology, free-floating systems bring more flexibility by allowing bikes to be parked at arbitrary locations within service regions, without requiring physical docking infrastructure. Despite their convenience, free-floating systems can lead to overoccupation in high-demand areas, as well as higher operational costs due to rebalancing bikes dispersed to remote areas (Du et al., 2019; Shen et al., 2018). Given these limitations, hybrid systems have emerged in recent years and are regarded as a promising alternative for future mobility (Duran-Rodas et al., 2020). These systems are characterized either by integrating designated docking infrastructure into otherwise free-floating models, or by adopting spatially differentiated parking policies across city regions—some allowing for free parking, others requiring docking at specific stations (Zhao and Ong, 2021; Liu and Tian, 2021; Lohrer et al., 2022). Some Chinese cities that were originally dominated by free-floating systems have begun adopting hybrid models by adding docking sites or electronic fences (Wei et al., 2024; Zhao and Ong, 2021). In Europe, a representative hybrid system is the MVG Rad in Munich, operated by the Münchner Verkehrsgesellschaft, which also serves as the use case in this study. As shown in Figure 1, bicycles can be picked up and returned freely within the designated green regions (Figure 1(a)). Outside these zones, however, users must return bikes to official MVG docking stations (marked by green dots), otherwise they face a penalty.

The effectiveness of a hybrid bike-sharing system is determined by whether newly emerging high-demand areas are served by designated docking stations, rather than being simply incorporated into free-floating regions. Evidence shows that expanding free-floating coverage in such areas often leads to unmanaged bike distribution and operational inefficiencies (Luo et al., 2019; Nikitas, 2019), whereas deploying docking stations in response to emerging demand patterns provides a more sustainable solution (Cheng et al., 2023). These findings motivate our modeling efforts to support strategic planning for station deployment in BSSs.

In this study, we aim to develop an efficient approach not only to identify but also to rank high-demand areas for guiding docking station deployment. This requires modeling demand-driven spatial usage patterns with sufficient granularity and accuracy. When external covariates are not considered,



(a) Service area and return regions in the Munich area, with free returns in the inner city and docking station returns elsewhere.

(b) Free-return regions (detail view) with MVG docking stations.

Figure 1: Operating areas of the MVG Rad bike-sharing system in Munich (Münchner Verkehrsgesellschaft (MVG), 2022).

modeling demand-driven spatial patterns typically is an unsupervised learning task. Existing studies have mainly adopted clustering-based approaches to address it (Zhao and Ong, 2021; Wei et al., 2024; Liu and Tian, 2021; Hua et al., 2020; Wang et al., 2022; Du et al., 2019; Zhang et al., 2019), as will be discussed in detail in the literature review (Section 2). However, these approaches exhibit multiple critical limitations that reduce their effectiveness in practice:

- First, as a preprocessing step, the macro spatial partitioning strategies employed in existing studies—such as uniform grids (Hua et al., 2020; Wei et al., 2024; Chen et al., 2025), administrative boundaries (Du et al., 2019), or the application of k-means clustering to return locations (Wei et al., 2024; Liu and Tian, 2021; Zhao and Ong, 2021)—overlook prior knowledge of existing transportation structures and often rely on hyperparameter tuning (e.g., grid size or the number of clusters).
- Second, clustering methods such as k-means, DBSCAN, and graph-based algorithms have been widely employed to model spatial demand within partitioned areas, but each comes with notable drawbacks. Among them, DBSCAN—a density-based method—is capable of identifying high-demand areas, but it entails relatively high computational complexity and requires extra hyperparameter tuning (Wei et al., 2024; Liu and Tian, 2021; Zhao and Ong, 2021). Graph-based clustering approaches are also computationally intensive, especially when applied to large-scale spatial networks. This issue is further amplified when such algorithms are integrated with optimization into hybrid frameworks (Wang et al., 2022; Chen et al., 2025). K-means clustering is computationally faster, but its distance-based nature tends to produce clusters of similar size and shape, making it more suitable for spatial partitioning (as previously noted) than for accurate modeling of localized demand patterns.

- c) Third, to the best of our knowledge, existing studies focus on detecting anomalous or high-demand areas but do not provide a ranking across candidate areas. In the context of continued urban expansion and constrained resources, the ability to globally rank irregular candidate areas is critical for prioritizing infrastructure investment.

In contrast to existing methods, we propose an analytical framework to model the demand patterns of irregular candidate areas. This study utilizes the geographical location of return records (latitude and longitude) as the primary data source. In contrast to pickup locations, which are heavily affected by operational factors such as bike availability or user proximity to stations, return locations more faithfully reflect users' intended destinations and thus provide a more reliable indication of spatial demand, supporting decisions on permanent infrastructure. Our approach is grounded in Extreme Value Theory (EVT), which provides a principled way to characterize the stochastic behavior of a univariate process at unusually large (or small) levels. The main contributions of this study are as follows:

1. We introduce an effective partitioning strategy that leverages the prior knowledge of existing docking stations and requires no hyperparameter tuning. Based on this, we formulate the modeling of demand patterns in irregular areas—an unsupervised learning task—within a group anomaly detection framework.
2. We develop a likelihood-based model using EVT that enables the evaluation of how globally anomalous a candidate area is, considering its joint profile of intensity, spatial coverage, and the configuration of surrounding areas.
3. Our approach achieves not only the identification but also the ranking of irregular candidate areas for new docking station deployment. At the same time, it maintains a low computational complexity, making it suitable for large-scale urban applications.
4. We demonstrate the practical utility of our approach in a case study of the MVG Rad system, a representative hybrid bike-sharing network. The results uncover data-driven patterns in unusual high-demand areas and support infrastructure planning.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. In Section 3, the modeling framework is explained in detail, including an overview of EVT, the construction of the proposed model, and a comparison with representative approaches. Section 4 presents the application to the MVG Rad system in Munich. Section 5 concludes and outlines directions for future work.

2. Literature review

Existing studies on BSSs can be broadly categorized into three directions. The first direction focuses on descriptive analyses of specific bike-sharing systems and associated policy implications. [Jensen et al. \(2010\)](#) provided insights into cycle flows in Lyon. [Sarkar et al. \(2015\)](#) discuss the similarity of bike usage patterns across ten cities. [Jonkeren and Kager \(2021\)](#) conducted a statistical investigation of the parking behavior of bicycle–train commuters. The second direction investigates the relationship between demand patterns and various explanatory variables, such as socio-demographic variables and weather conditions. [Shen et al. \(2018\)](#) employed spatial autoregressive models to analyse the impact of the surrounding built environment, access to public transportation, and weather conditions on demand patterns. [Du et al. \(2019\)](#) used random forests to explore the contribution of land-use variables, population, point-of-interest variables and cycling environment variables to demand patterns. [Ashqar et al. \(2017\)](#) predicted the number of available bikes through regression models from neighborhood availability, temporal factors, and weather conditions. The third research direction, to which our study contributes, emphasizes the development of approaches that directly model demand patterns—without relying on external explanatory variables—to guide future decisions regarding infrastructure design.

Before reviewing specific studies within the third direction, we begin by discussing the role of clustering algorithms, which are among the most widely used techniques in demand pattern modeling. K-means clustering is particularly prevalent due to its low computational complexity and ease of implementation. By assigning each point to the nearest centroid and updating centroids to the mean, it yields compact, roughly convex clusters with similar spread. However, as a distance-based iterative method, k-means clustering can produce clusters with heterogeneous internal densities, which limits its ability to capture spatial density structures. Therefore, k-means clustering has been mainly used to partition the studied area into subregions ([Wei et al., 2024](#); [Liu and Tian, 2021](#); [Zhao and Ong, 2021](#)), and, in rare cases, to determine virtual station locations for specific datasets ([Hua et al., 2020](#)). In contrast, as a density-based clustering algorithm, DBSCAN is capable of identifying clusters with arbitrary shapes and consistent internal densities. This makes it well suited for detecting high-demand areas. However, due to its higher computational complexity, DBSCAN is generally employed in the final stage of modeling pipelines—after segmentation, preprocessing, and optimization—to identify high-demand areas at a finer spatial scale ([Wei et al., 2024](#); [Liu and Tian, 2021](#); [Zhao and Ong, 2021](#)). Other representative methods use graph-based clustering, including community detection, to identify cohesive spatial structures, with computational cost rising on large spatial networks ([Liu and Tian, 2021](#); [Wang et al., 2022](#)).

For the third direction, representative studies have modeled demand patterns through unsupervised

learning, typically in hybrid frameworks built around clustering algorithms. Caggiani et al. (2018) proposed a spatiotemporal clustering method in which temporal clusters are built first, and subsequently k-means clustering is combined with bilevel optimization to form spatiotemporal clusters. Zhao and Ong (2021) first applied DBSCAN to detect dense clusters, then used k-means clustering to subdivide large clusters under a service-radius constraint to derive geo-fenced parking locations and capacities. Liu and Tian (2021) combined k-means clustering and complex network analysis to conduct macro- and micro-level spatial partitioning, followed by DBSCAN for final demand detection. Wang et al. (2022) constructed a graph representation of the bike-sharing system, where vertices represent bike stations and arcs represent bike flows between stations. A modified fast unfolding algorithm was then applied to this graph to detect communities. Wei et al. (2024) modeled demand by first partitioning the space with k-means clustering and then applying DBSCAN within each part, and demonstrated the advantage of this approach over using either algorithm independently.

3. Model

This section presents the proposed model, which identifies and ranks high-demand candidate areas by quantifying the anomalousness of each area in terms of demand intensity, spatial coverage, and the configuration of its surrounding areas. We begin by introducing key concepts from EVT in Section 3.1, which form the statistical basis of our approach. Building on this foundation, the subsequent subsections develop the proposed modeling framework. In Section 3.2, we formulate this demand modeling task as a group anomaly detection problem through a two-stage partitioning strategy. Then Section 3.3 outlines the detailed modeling procedure, including feature extraction, modeling, and anomaly scoring. Finally, in Section 3.4, we analyse representative approaches and demonstrate our method's advantage in comprehensively capturing demand characteristics. For illustrative purposes, we use the MVG bike-sharing data from Munich to support the explanation of our modeling framework. Nonetheless, the proposed model is general and not specific to this system. A summary of the key mathematical symbols used in this study is provided in Table 1 for ease of reference.

3.1. Extreme value theory

This subsection introduces the core EVT concepts that provide the foundation for our approach. It first introduces the two classical EVT models, block-maxima (BM) and peaks-over-threshold (POT), followed by the point process model (PPM), which is an approach that unifies the two classical models.

EVT is a statistical discipline with the objective of modeling the stochastic behavior of a univariate process at unusually large (or small) levels, which is widely applied in fields such as reliability analysis, structural health monitoring, meteorology, and financial risk assessment (Li and Xiao, 2023). The

Table 1: Summary of key notations and abbreviations

Symbol / Abbreviation	Definition
X, x	A univariate random variable and a realization thereof; in our case it is the magnitude feature of a subregion, where $X = \log(C) \cdot \log(E)$
$\tilde{x} = \{x_1, x_2, \dots, x_n\}$	A group realization consisting of n points
u_n	Threshold used to determine exceedances for X
c_n, d_n, ξ	Location, scale, and shape parameter for the right tail of X
u	Normalized u_n where $u = \frac{u_n - c_n}{d_n}$
C, E	Number of returns in a subregion and spatial extent (convex hull area)
$\tilde{x}^{\text{exc}} = \{x_1^{\text{exc}}, x_2^{\text{exc}}, \dots, x_k^{\text{exc}}\}$	Set of exceedances within the group $\tilde{x}$
$l_n(\tilde{x}^{\text{exc}})$	Likelihood of the exceedance pattern in the group $\tilde{x}$
V_n	Likelihood variable where $V_n = l_n(\tilde{x}^{\text{exc}})$
$G(v)$	Analytical cumulative distribution function of the likelihood where $G(v) = \lim_{n \rightarrow +\infty} P(V_n \leq v)$
$\mathbb{A} = \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m\}$	A collection of m group realizations
$S(\tilde{x})$	Anomaly score for group realization $\tilde{x}$
BSSs	Bike-sharing systems
EVT	Extreme value theory
GEV	Generalized extreme value family of distributions
GPD	Generalized Pareto family of distributions
PPM	Point process model
BM	Block-maxima model
POT	Peaks-over-threshold model

central result in EVT is the Fisher–Tippett theorem concerning the limiting distribution of the block maxima

$$M_n = \max\{X_1, X_2, \dots, X_n\}, \text{ as } n \rightarrow +\infty,$$

where $X_1, X_2, \dots, X_n$ are independent and identically distributed (i.i.d.) univariate random variables with a common distribution function F_X .

Theorem 1. [Fisher–Tippett–Gnedenko] *If there exist sequences of normalizing constants c_n and d_n such that the following convergence in distribution holds:*

$$P\left(\frac{M_n - c_n}{d_n} \leq x\right) \rightarrow G_\xi(x), \text{ as } n \rightarrow +\infty, \quad (1)$$

then the non-degenerate limiting distribution $G_\xi(x)$ is a member of the so-called Generalized Extreme

Value (GEV) family of distributions:

$$G_{\xi}(x) = \begin{cases} \exp\left(-(1 + \xi x)^{-\frac{1}{\xi}}\right) & , \text{ if } \xi \neq 0 \\ \exp(-\exp(-x)) & , \text{ if } \xi = 0. \end{cases} \quad (2)$$

If $\xi \neq 0$, then the domain of the distribution is restricted to the set $\{x \mid 1 + \xi x > 0\}$.

If the shape parameter ξ is negative, zero, or positive, respectively, then the corresponding distribution G_{ξ} belongs to the Weibull, Gumbel and Fréchet family, respectively. The shape parameter ξ thus determines the behavior in the tails of the distribution F_X .

The GEV family provides a model for block-maxima, *i.e.*, maxima of sequences $\{X_1, X_2, \dots, X_n\}$. However, for relatively large n , there will be few block-maxima available from the data which can lead to unreliable estimates of ξ (Coles, 2001). To overcome this problem, the peaks-over-threshold method can be used, which avoids the need for blocking. This method infers a model from all observations that exceed some high threshold.

Theorem 2. [Pickands–Balkema–De Haan] For a given $u \in \mathbb{R}$, the sequence of thresholds $u_n = c_n + u d_n$ is considered, where c_n and d_n denote the normalizing constants as defined in (1). If (1) holds for some member $G_{\xi}(x)$ of the GEV family, then the distribution of the exceedances $X - u_n$, conditional on $X > u_n$, satisfies the limiting property:

$$\lim_{n \rightarrow +\infty} P\left(\frac{X - u_n}{d_n} < x \mid X > u_n\right) = H_{\xi}(x). \quad (3)$$

The distribution

$$H_{\xi}(x) = \begin{cases} 1 - \left(1 + \frac{\xi}{1 + \xi u} x\right)^{-\frac{1}{\xi}} & , \text{ if } \xi \neq 0 \\ 1 - \exp(-x) & , \text{ if } \xi = 0 \end{cases} \quad (4)$$

denotes the family of Generalized Pareto Distributions (GPD) with $x \geq 0$ in case $\xi \geq 0$, and $0 \leq x \leq -\frac{1}{\xi}$ in case $\xi < 0$.

These two classical models represent the cornerstone of EVT. The three parameters, *i.e.*, the shape parameter ξ and the normalizing constants c_n and d_n , are typically estimated using maximum likelihood methods. For a range of commonly encountered distributions, such as the gamma, exponential, Gaussian, and beta distributions, Embrechts et al. (2013) derive closed-form expressions for estimating these parameters. For instance, in the Gumbel case, exceedances over a high threshold can be used to estimate:

$$c_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right), \quad d_n = E(X - u_n \mid X > u_n), \quad (5)$$

where $F^{\leftarrow}$ denotes the quantile function corresponding to the distribution of X .

A (spatial) PPM in Euclidean space can be viewed as a random variable $\tilde{X}$ defined over all possible point patterns in some subspace A of $\mathbb{R}^d$ ($d \in \mathbb{N}_0$ represents the dimension of the multivariate variable X), where a realized point pattern is given by $\tilde{x} = \{x_1, x_2, \dots\}$, $x_i \in \mathbb{R}^d$ (Illian et al., 2008; Vo et al., 2018). A PPM integrates stochastic information about the number of points and their individual locations within the subspace A .

A PPM can be used to elegantly characterize the exceedances in a way that unifies the EVT models presented in Theorems 1 (block-maxima) and 2 (peaks-over-threshold) (Coles, 2001; Moins et al., 2023).

Theorem 3. [Point Process Characterization] *When the limiting distribution (3) of the block maxima of the univariate random variable X holds, the PPM of patterns consisting of exceedances of the block $\{X_1, \dots, X_n\}$ w.r.t. the threshold u_n , i.e.:*

$$\tilde{X}_n^{exc} = \{X_1, X_2, \dots, X_n\} \cap [u_n, +\infty),$$

converges to a Poisson point process as $n \rightarrow +\infty$. The corresponding counting measure $N_A(\tilde{X}_n^{exc})$ converges in distribution to a Poisson distribution:

$$\lim_{n \rightarrow +\infty} P(N_A(\tilde{X}_n^{exc}) = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad (6)$$

where $\lambda = E(N_A(\tilde{X}_n^{exc}))$ is the expectation of the counting measure.

In the Gumbel case, where $\xi = 0$, it can be shown that $\lambda = \lim_{n \rightarrow +\infty} nP(X > u_n) = e^{-u}$ (Embrechts et al., 2013).

3.2. Formulation of the group anomaly detection task

In this subsection, we formulate the problem of detecting areas with unusually high demand in BSSs as a group anomaly detection task. Since direct modeling of the entire citywide area is computationally infeasible, a macro-level partition is required. The resulting partitioned areas are treated as structurally equivalent modeling units so that subsequent modeling can be applied consistently across them. Common partitioning strategies include uniform grids (for example, 50 m $\times$ 50 m), administrative boundaries, and k-means clustering. However, these strategies are affected by the modifiable areal unit problem: results can vary with scale and configuration. In practice, the grid size, administrative level, and clustering hyperparameters must be determined through optimization or expert judgment, which increases the computational cost.

In our framework, we propose a data-driven spatial partitioning strategy. The study area is divided into regions, where each region is defined by a unique docking station and contains all bike return

locations closest to it. This strategy is motivated by the station deployment logic. Existing docking stations were deployed to match local demand and to balance service load across their catchments, which means the layout of stations is expected to reflect the spatial distribution of demand. As a result, the partition yields structurally equivalent modeling units for downstream modeling and avoids complicated hyperparameter tuning required by other approaches. This process is illustrated using the MVG bike-sharing data during the period 2015–2023. The studied area is the non-free return zone. Figure 2(a) shows the return activity heatmap within the studied area, with blue points indicating docking stations installed in the past. Figure 2(b) presents the partitioning result. Many existing stations remain aligned with local return patterns, for example, those within the red boxes. But in other areas, the current station layout leaves clear gaps in coverage due to the expansion of demand with ongoing urban development.

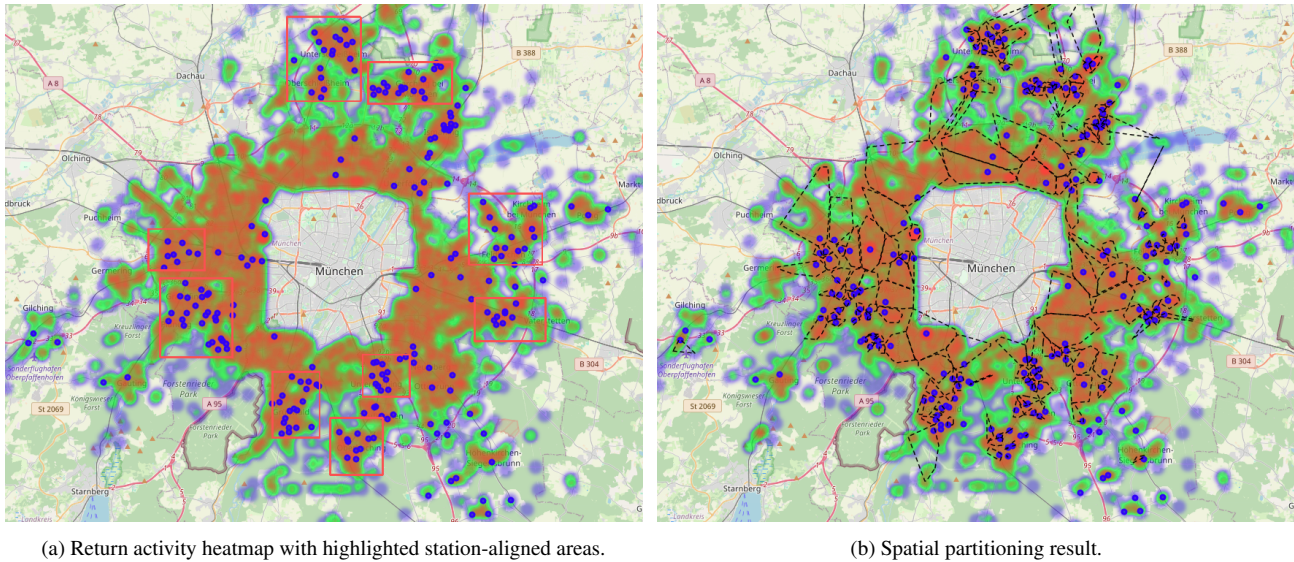


Figure 2: Return activity heatmaps and spatial partitioning within non-free return regions of Munich (2015–2023). Blue points denote existing docking stations.

We now develop a group anomaly detection framework to model demand patterns in candidate areas. In this framework, each region—anchored by a single docking station—is treated as a distinct group. Within each group, bike return locations are further clustered into localized subregions using k-means clustering. This produces spatially coherent subregions that reflect local variation in demand. K-means clustering is adopted for its computational efficiency and its tendency to produce compact, approximately balanced subregions. For each group, the value of k is selected from a candidate set using the elbow criterion. Figure 3(a) shows the spatial partitioning of the studied area into groups. Figure 3(b) presents an example of a group (Am Hart), with distinct colors denoting the localized subregions.

The regions resulting from the spatial partitioning are treated as groups of smaller subregions that are obtained by clustering the geographical coordinates of bike return records. In the next section,

the distribution of records across these subregions is modeled to determine whether a region displays anomalous demand patterns and which subregions require additional docking stations.

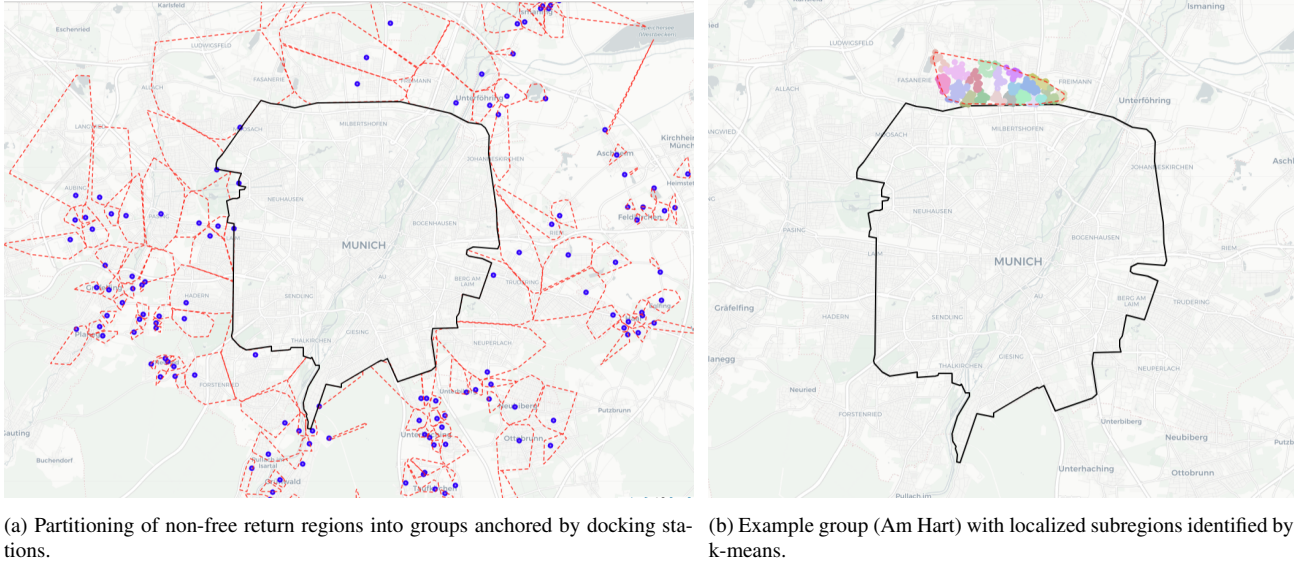


Figure 3: Spatial partitioning of non-free return regions into groups and localized subregions. Each group is anchored by a docking station, and within groups, k-means clustering produces subregions.

3.3. Subregion-level feature and group anomaly modeling

To describe the magnitude of demand within each subregion, we define a subregion feature. This feature captures both how many return records the subregion contains and how broadly that demand extends in space. These two facets are considered because an unusually high number of return records within a subregion indicates higher demand intensity, and a large spread implies a broad catchment. The feature is defined as:

$$X = \log(C) \cdot \log(E), \quad (7)$$

where C denotes the number of returns within a subregion, and E denotes the area of its convex hull. The logarithmic transformation of both intensity and area mitigates disparities in scale and stabilizes variance, preventing either attribute from dominating. The magnitude $X = \log(C) \cdot \log(E)$ amplifies when both intensity and area are high, and it diminishes when either component is small (diminishing marginal effect).

While the feature X summarizes the magnitude of a subregion, assessing a subregion's anomalousness cannot be determined from X alone, as it also depends on how it is embedded in the spatial configuration of surrounding subregions. We illustrate this with an example in Figure 4. Suppose the two red-boxed subregions have identical magnitude values X . Their priority for setting new docking stations differs. The upper subregion lies in an area surrounded by many neighboring subregions with high demand but few stations nearby. The lower subregion is surrounded by fewer high-demand

neighbors, whereas more stations are located nearby. Therefore, from an urban planning perspective, although their X values are the same, the upper subregion appears more anomalous and would receive higher priority for setting stations.



Figure 4: Illustration of two subregions with identical magnitude but differing anomalousness due to group-level spatial context. Red boxes indicate their approximate locations.

Next, we will model the configuration within groups using a point process model grounded in EVT to assess subregions' anomalousness. Formally, a group (realization) is represented by a set:

$$\tilde{\mathbf{x}} = \{x_1, x_2, \dots, x_n\}, \quad (8)$$

where x_i denotes the magnitude of the i -th subregion within the group. Subregions with magnitude values exceeding a high threshold u_n are considered as areas with an unusually high demand that need further attention. In EVT, these subregions are referred to as exceedances that capture events in the upper tail of X . The set of exceedances—representing a collective of unusual high-demand subregions within a group—constitutes a realization of a point process on the upper tail of X (see Theorem 3). Our model captures the spatial structure of exceedances within a group and enables the computation of the likelihood associated with any observed realization. We further derive the asymptotic distribution of this likelihood, which serves to quantify the anomalousness of each group.

According to the peaks-over-threshold approximation of EVT, the distribution of these exceedances can be approximated by a generalized Pareto distribution when the parent distribution of X has block maxima satisfying the GEV law, as established in Theorem 2. For a group realization $\tilde{\mathbf{x}}$, let $\tilde{\mathbf{x}}^{\text{exc}}$ denote the set of exceedances above u_n :

$$\tilde{\mathbf{x}}^{\text{exc}} = \{x_1^{\text{exc}}, x_2^{\text{exc}}, \dots, x_k^{\text{exc}}\}, \quad (9)$$

where x_j^{exc} denotes a realization in $\tilde{\mathbf{x}}$ that exceeds the threshold u_n . According to Theorem 2, if the tail distribution of X has a Gumbel limit (i.e., $\xi = 0$), the probability density of the j -th exceedance takes the form:

$$p(x_j^{\text{exc}}) = P(X_j^{\text{exc}} = x_j^{\text{exc}}) \approx \frac{1}{d_n} e^{-\frac{x_j^{\text{exc}} - u_n}{d_n}}. \quad (10)$$

Indeed, the Gumbel case covers many widely used parent distributions, such as the Gaussian, exponential, Weibull, gamma, and log-normal distributions.

Next, we consider the probability of observing $K = k$ exceedances. According to Theorem 3, as $n \rightarrow \infty$, the exceedances on the upper tail of X converge to a Poisson point process. The number of exceedances K is approximately Poisson with rate λ . Let μ_k denote the probability mass at k :

$$\mu_k = P(K = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}.$$

In the Gumbel case (i.e., $\xi = 0$), the Poisson rate parameter simplifies to $\lambda = e^{-u}$, with u being the normalized threshold.

Based on the above components, we derive the likelihood in Janossy representation for observing a group with the aggregate $\tilde{\mathbf{x}}^{\text{exc}}$ of k exceedances as follows:

$$l_n(\tilde{\mathbf{x}}^{\text{exc}}) := \mu_k k! \prod_{j=1}^k p(x_j^{\text{exc}}). \quad (11)$$

In this expression, μ_k is the probability of observing k exceedances; the factorial term $k!$ corrects for permutations among the k exceedances, ensuring the likelihood is invariant to their ordering; and the product $\prod_{j=1}^k p(x_j^{\text{exc}})$ corresponds to the joint likelihood of their observed exceedances under the exponential approximation. The likelihood $l_n(\tilde{\mathbf{x}}^{\text{exc}})$ serves as a unified measure to reflect the anomalousness of a group in terms of the exceedances as a whole. To probabilistically quantify this anomalousness measure, we derive the asymptotic distribution of the likelihood $l_n(\tilde{\mathbf{x}}^{\text{exc}})$, which forms the theoretical foundation of our framework. The formal statement of this limiting distribution is presented in the following theorem, with the proof given in Appendix A.

Theorem 4. *Consider a random variable X whose distribution F_X has block maxima satisfying the Gumbel law. Then, the exceedances over a high threshold u_n satisfy the following result:*

$$\lim_{n \rightarrow +\infty} P\left(\frac{X - u_n}{d_n} < x \mid X > u_n\right) = 1 - e^{-x}, \quad (12)$$

with $u_n = c_n + u d_n$, for some $u \in \mathbb{R}$ and c_n, d_n normalizing constants. The likelihoods $V_n = l_n(\tilde{\mathbf{x}}^{\text{exc}})$ of patterns of exceedances $\tilde{\mathbf{x}}^{\text{exc}} = \{x_1^{\text{exc}}, x_2^{\text{exc}}, \dots, x_k^{\text{exc}}\}$, $x_i^{\text{exc}} > u_n$, will converge in distribution to a random

variable V with CDF:

$$\begin{aligned} G(v) &= \lim_{n \rightarrow +\infty} P(V_n \leq v) \\ &= \mu_0 H(v - \mu_0) + \sum_{k \geq 1} \mu_k \left[1 - \mathcal{E}_{k,1} \left(-\log \frac{v \cdot (d_n)^k}{\mu_k \cdot k!} \right) \right], \end{aligned} \quad (13)$$

where $\mathcal{E}_{k,1}$ denotes the CDF of an Erlang distribution with shape parameter k and scale 1, and

$$\mu_k = \frac{\lambda^k}{k!} e^{-\lambda}. \quad (14)$$

Hence, H is the Heaviside step function, i.e. $H(v) = 1$ for $v \geq 0$ and $H(v) = 0$ for $v < 0$.

In summary, the anomalousness of a group is quantified by the likelihood derived in (11), where lower values indicate greater deviation from expected configurations of exceedances and thus higher anomalousness. Theorem 4 provides the asymptotic distribution of the likelihood, enabling us to probabilistically evaluate how rare a given group's likelihood is. Accordingly, we define the anomaly score for a given group $\tilde{x}$ with likelihood value v as the upper-tail probability

$$S(\tilde{x}) = P(V \geq v) = 1 - G(v), \quad (15)$$

where $G(v)$ is the limiting cumulative distribution function of the likelihood. The construction of $S(\tilde{x})$ ensures consistency with standard anomaly detection practices, where anomaly scores fall within the range $[0, 1]$, and higher scores correspond to more anomalous behavior. For implementation, the scoring procedure is formalized in Algorithm 1.

Algorithm 1 Computing group anomaly scores from the likelihood of the exceedance set

Input: A collection of m group realizations $\mathbb{A} = \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m\}$, where each $\tilde{x}_t$ consists of subregion-level magnitude vectors represented by X . The block size n is set as the maximum number of clusters across groups.

Output: Anomaly scores $\{S(\tilde{x}_1), S(\tilde{x}_2), \dots, S(\tilde{x}_m)\}$ for all groups.

- 1: Determine the exceedance threshold u_n using the mean residual life plot¹ constructed from the set of all subregion magnitudes x .
 - 2: Collect all exceedances that satisfy $x > u_n$.
 - 3: Estimate the normalization constants c_n and d_n in Eq. (5) by fitting an extreme value model to the exceedances.
 - 4: **for** each group $\tilde{x}_t \in \mathbb{A}$ **do**
 - 5: Compute the likelihood $v_t = l_n(\tilde{x}_t^{\text{exc}})$ based on Eq. (11), and evaluate its cumulative probability $G(v_t)$.
 - 6: Compute the anomaly score $S(\tilde{x}_t) = 1 - G(v_t)$ using Eq. (15), which yields a probabilistic measure of anomalousness.
 - 7: **end for**
-

Once the group-level anomaly scores $S(\tilde{x})$ are computed, the proposed framework proceeds to its

final stage. Groups are first ranked in descending order of anomalousness to prioritize those most in need of infrastructure improvement. Within each prioritized group, the subregions whose magnitudes exceed the threshold u_n are further sorted based on their magnitudes, allowing for the identification of localized high-demand areas. For each exceedance i in group t , we define the candidate

$$C_{ti} = (S(\tilde{\mathbf{x}}_t), x_{ti}).$$

A global ranking of subregions is obtained by ordering the candidates $C_{ti} = (S(\tilde{\mathbf{x}}_t), x_{ti})$ in descending lexicographical order, comparing group scores first and, if equal, subregion magnitudes:

$$C_{ti} \succ_{\text{lex}} C_{t'i'} \iff (S(\tilde{\mathbf{x}}_t) > S(\tilde{\mathbf{x}}_{t'})) \vee (S(\tilde{\mathbf{x}}_t) = S(\tilde{\mathbf{x}}_{t'}) \wedge x_{ti} > x_{t'i'}). \quad (16)$$

The top R candidates under this order are collected in a set D_R . The overall workflow is illustrated in Figure 5, which summarizes the full pipeline from raw return records to actionable deployment sites.

3.4. Comparison with representative recent methods

We compare our approach with four recent studies on the planning of docking stations. They cover the representative partitioning strategies and typical modeling choices used in practice. Since a direct comparison on a common dataset is not practical, as each study uses a different design of hyperparameters for its own data, we therefore methodologically summarize them in Table 2 with respect to four criteria (partitioning strategy, algorithmic structure, ranking capability, and computational complexity).

Table 2: Comparison of representative recent methods

Literature	Partitioning	Method	Supports Ranking	Computational Complexity
Zhang et al. (2019)	Uniform grids	Maximum coverage selection + DBSCAN	No	$O(m_g E \log V) + p\mathcal{A} + N_p^2)$
Zhao and Ong (2021)	DBSCAN	K-means + Pareto-based tuning	No	$O(GN^2)$
Liu and Tian (2021)	K-means clusters	Community detection + DBSCAN	No	$O(N^2)$
Wei et al. (2024)	K-means clusters	DBSCAN	No	$O(N^2)$
This study	Groups (via stations)	EVT-based model	Yes	$O(k^* \log(k^*))$

We first examine how recent studies delineate spatial partitions. [Zhang et al. \(2019\)](#) applies a uniform grid to divide the urban area into predefined spatial units. While this approach is simple, it is

¹The mean residual life plot is a diagnostic graph used in extreme value analysis. Its curve is constructed by plotting the average of the excess values over every possible threshold, as shown in Figure 8(b). The threshold u_n is typically chosen as the smallest value where the final, stable linear trend begins.

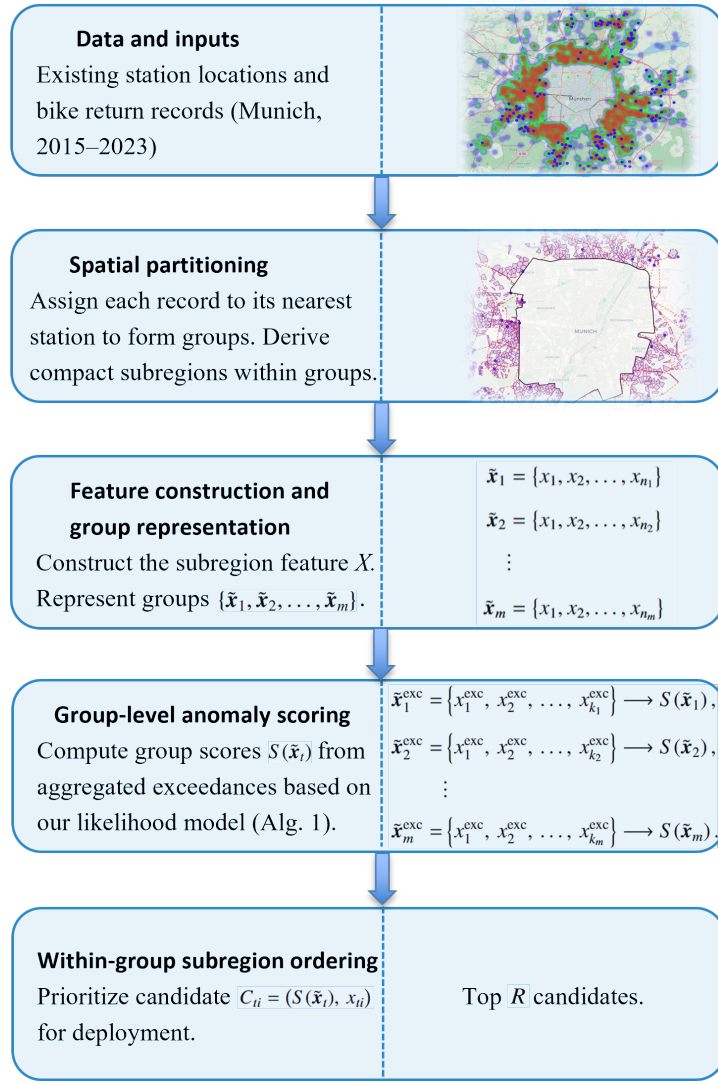


Figure 5: Overall workflow of the proposed framework.

fundamentally constrained by the modifiable areal unit problem. The outcome of the model can vary significantly depending on the grid size. In this study, the selected grid resolution ($50 \text{ m} \times 50 \text{ m}$) is determined empirically rather than through a formal optimization process. The other three studies employ data-driven partitioning schemes learned from observations. Zhao and Ong (2021) implemented partitioning using the DBSCAN algorithm, grouping contiguous and demand-similar locations into clusters while treating isolated points as noise. The two DBSCAN hyperparameters—neighborhood radius (ϵ) and minimum points (minPts)—are selected by solving a demand-based optimization problem. Although this enables the partition to reflect local density variations, the exhaustive parameter search entails substantial computational cost. Liu and Tian (2021) and Wei et al. (2024) employ k-means clustering for macro partitioning. In their reported cases, the number of clusters k is fixed *a priori*. However, it is not guaranteed that this way of choosing k generalizes to other applications. Our approach, on the other hand, is data-driven and informed by existing docking stations, leading to a partition that adapts to demand patterns rather than relying on fixed grids. This station-centered strategy

eliminates the need to algorithmically derive partitions from data and is computationally efficient.

We next analyse the computational complexity and ranking support of the methods. Zhang et al. (2019) considered fence-site selection as a maximum-coverage problem. The overall complexity is $O(m_g|E|\log|V|) + p\mathcal{A} + N_p^2$). Specifically, they divided the uniform grids into n_g demand cells and m_g candidate cells ($n_g > m_g$). The authors provide a road network $G = (V, E)$, where $|V|$ and $|E|$ are the numbers of nodes and edges. The coverage network is constructed by adding an edge between a candidate cell and a demand cell when their network distance (computed on the road network G) is within D . This construction process costs $O(m_g|E|\log|V|)$. Then selecting p cells from candidate cells by greedy maximum coverage takes $O(p\mathcal{A})$, where $\mathcal{A}$ is the number of edges in the coverage network. Finally, DBSCAN runs on all return records inside the p selected cells. With N_p records in total, this step costs $O(N_p^2)$. None of the three terms is guaranteed to be the leading term, so we keep all three. This approach outputs a binary selection of candidate cells, which does not support ranking. Zhao and Ong (2021) applied DBSCAN to all return records to make a macro partition, then ran k-means clustering inside each part. This approach selects ϵ and minPts via a grid search over G settings. Let N be the number of all records. One DBSCAN run costs $O(N^2)$, trying G settings costs $O(GN^2)$. The k-means clustering has bounded iterations and is lower-order, so it does not affect the leading $O(GN^2)$ term. Therefore, the overall complexity is $O(GN^2)$. This approach compares different settings (ϵ -minPts) rather than ranking areas for a given choice. Although Liu and Tian (2021) and Wei et al. (2024) differ in their intermediate steps, they use the same macro-micro pipeline: k-means clustering partitions the study area into coarse regions, and DBSCAN refines candidates within each region. K-means clustering is of lower-order cost here; the dominant cost is DBSCAN at the micro level. With N total records spread across regions, summing per-region costs gives $\sum_j O(N_j^2)$, and is upper-bounded by $O(N^2)$. Neither approach yields a global ranking of candidate areas.

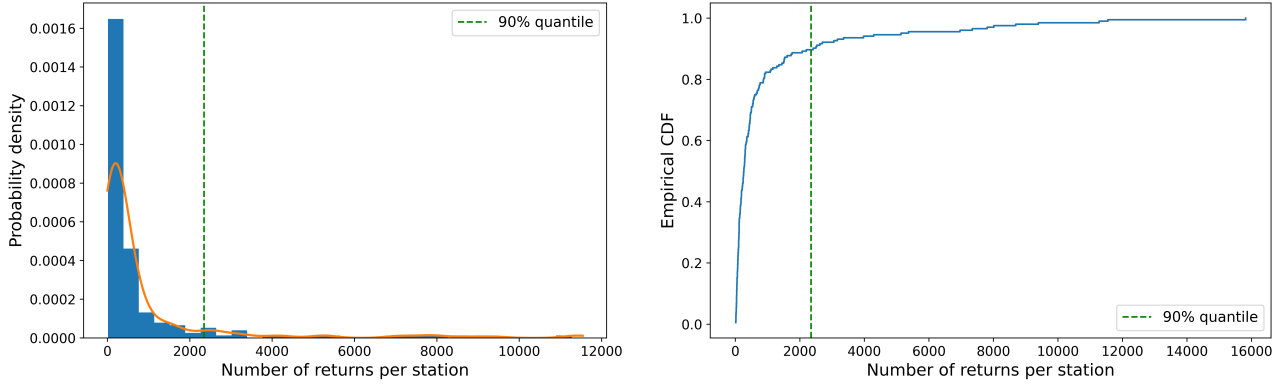
Our method provides a theoretically grounded mechanism for selecting and ranking docking sites with low computational cost. For each group $\tilde{x}_t$ with k_t exceedances, computing the likelihood (Eq. (11)) only requires scanning those k_t values once, having $O(k_t)$ complexity. Summed over all groups gives $O(k^*)$, where $k^* = \sum k_t$ denotes the total number of exceedances. The computational complexity is effectively linear in the number of exceedances. The subsequent global ranking of all k^* exceedances by the lexicographical order $(S(\tilde{x}_t), x_{ti})$ requires $O(k^* \log(k^*))$, which is the dominant term of our method. Compared with other approaches, this cost is substantially lower.

4. Case study

In this section the case study of the MVG Rad bike-sharing system in Munich is analysed. Section 4.1 describes the dataset. Section 4.2 applies the proposed framework to identify and rank anoma-

lous subregions for docking station deployment. Section 4.3 examines seasonal demand variation and takes into account season-specific overloads in deployment recommendations.

4.1. Dataset and preprocessing



(a) Histogram and kernel density estimate, with dashed line marking the 90-th percentile. (b) Empirical cumulative distribution function with dashed line marking the 90-th percentile.

Figure 6: Distribution of return records per station in the MVG Rad system.

This study uses official trip records from MVG Rad, the public bike-sharing system in Munich operated by the Münchner Verkehrsgesellschaft, accessible at <https://www.mvg.de/ueber-die-mvg/unsere-fahrzeuge/mvg-rad.html>. The dataset covers rides from 2015 to 2023 and contains 4.76 million trip records in total. Annual counts are 49.8 thousand (2015), 342.4 thousand (2016), 419.0 thousand (2017), 435.9 thousand (2018), 753.7 thousand (2019), 721.8 thousand (2020), 619.6 thousand (2021), 709.1 thousand (2022), and 710.1 thousand (2023). Time stamps are recorded in local time (CET/CEST), and locations are given as latitude and longitude in WGS 84. Each trip record includes a unique ride identifier, start and end time stamps, origin and destination coordinates, and flags indicating whether the ride began or ended at a docking station. The MVG Rad system follows an operational model where bikes can be freely picked up and returned within designated areas, while returns outside these areas are restricted to docking stations, as detailed in Section 1.

Preprocessing consists of steps that ensure data integrity and a consistent format for subsequent analyses. Records with missing return coordinates were excluded. Outliers with unrealistic coordinates (e.g., coordinates in polar regions) were removed. The indicators of start and return at docking stations were converted to a consistent binary (0/1) format. Extra spaces in some records were removed to maintain consistent formatting and prevent parsing issues during programming. All analyses were performed in Python 3.10.13 using JupyterLab on Ubuntu Linux. Computations were executed on an Intel Xeon w9-3475X with 36 cores and 72 threads.

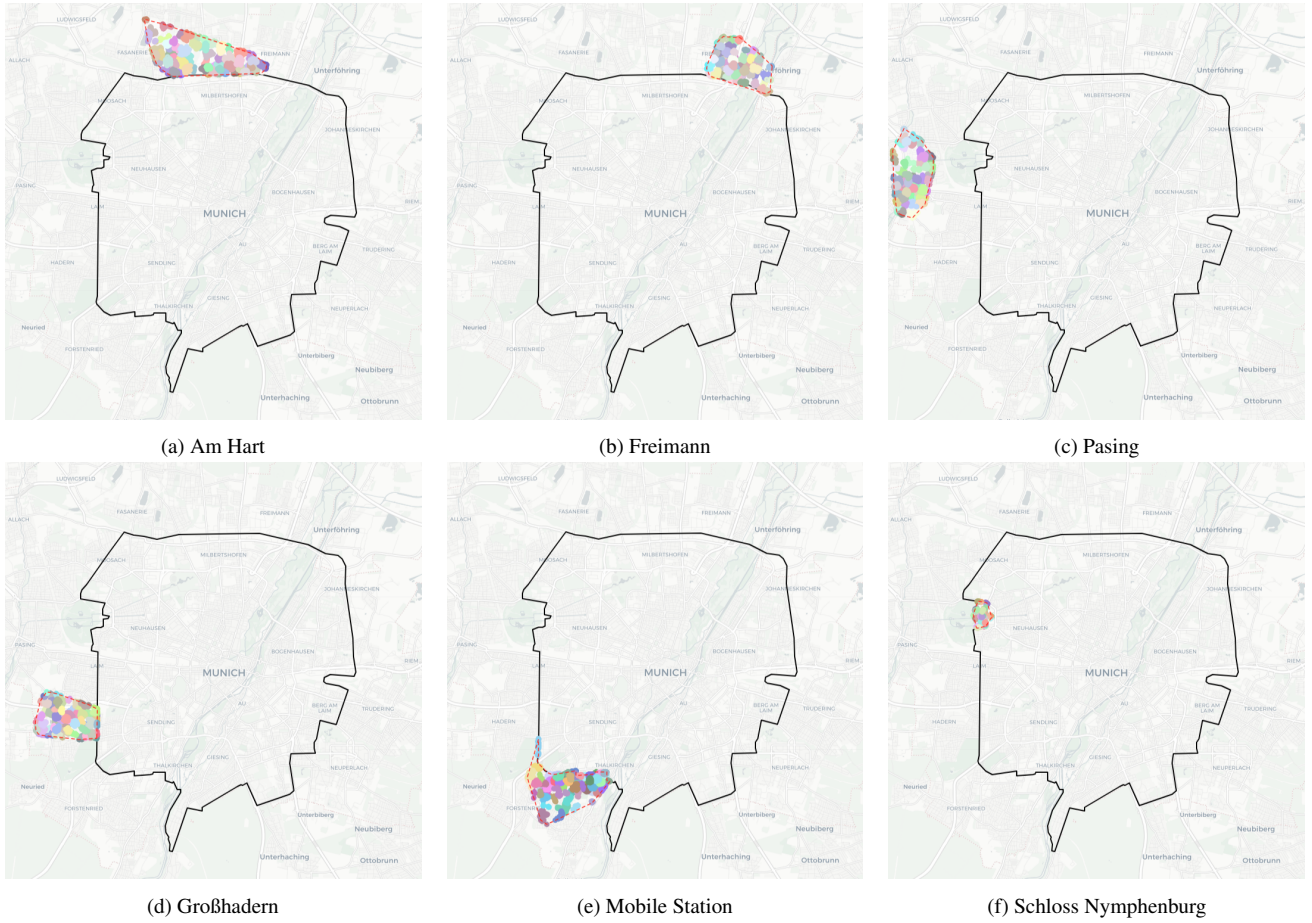


Figure 7: Examples of station groups and their subregions obtained by k-means clustering.

4.2. Detection of high-demand subregions and spatial ranking

In our experiments, the studied area is the non-free return region of the MVG Rad system. This system contains 180 docking stations that serve as valid group anchors, with the resulting partition shown in Figure 3. Figure 6 shows the density and cumulative distribution of the number of return records, obtained across the subregions. The distribution is strongly right-skewed, with a small subset of stations above the 90-th percentile accounting for those with an unusually high number of return records. Given the high demand for this subset of stations, an increased need for docking infrastructure in their neighborhood is expected.

We apply the approach outlined in Section 3. To this end, the study area is first partitioned into regions as shown in Figure 3, where each existing station belongs to a unique region. Within each region, k-means clustering is performed on the return record data, yielding a set of subregions. We use the elbow method to select the optimal number of clusters. An upper bound on the number of clusters is imposed via a minimum cluster size to prevent over-segmentation. This constraint is an application-level safeguard and is optional, not part of the core method. We use the 10-th percentile of the distribution of returns per station, which equals 45, as the reference minimum cluster size. Limiting the number of clusters in this way leverages the k-means objective to naturally discourage

unstable small clusters. Examples of six groups and their subregions are shown in Figure 7.

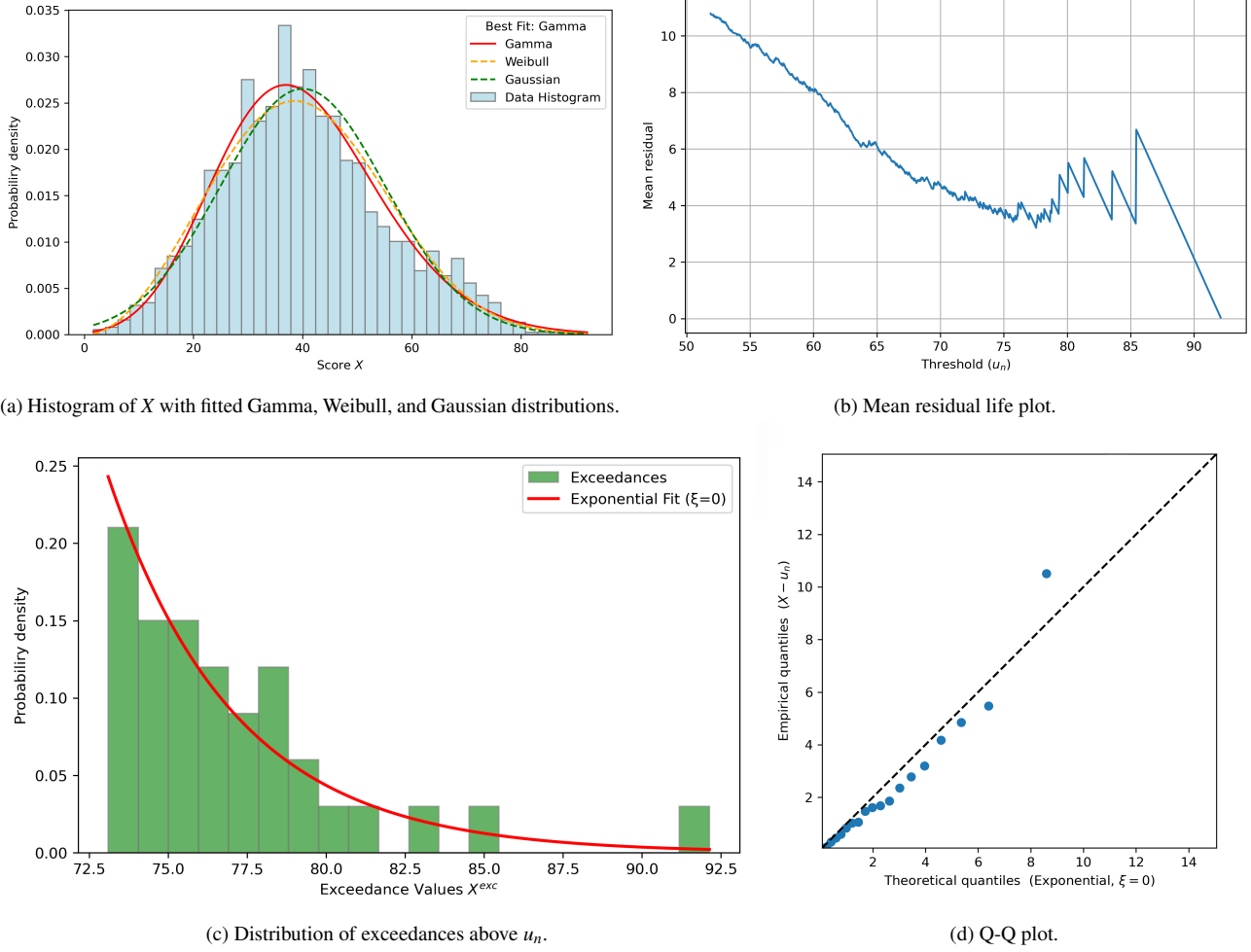


Figure 8: Diagnostic plots for assessing the Gumbel extreme value distribution of X .

The magnitude feature X (Eq. (7)) is constructed to characterize each subregion by combining intensity and areal coverage, and it serves as the univariate input for the subsequent modeling. Figure 8(a) presents the histogram of X across all subregions. Rather than aiming for a perfect parametric fit, here the goal is to assess whether the tail of X has a Gumbel limit, an assumption underlying our extreme value analysis. To validate this assumption, we examine the distribution of exceedances over threshold u_n using the mean residual life plot (Figure 8(b)). For a suitable high threshold, the plot should exhibit an approximately linear trend, which supports the robust estimation of the parameters of the GPD. In Figure 8(b), the final linear behaviour begins near 72, so we set $u_n = 72$. Figure 8(c) shows that the exceedances agree with a Gumbel fit. As a further check, the Q-Q plot in Figure 8(d) compares the empirical quantiles of $X - u_n$ with the theoretical exponential quantiles. The points lie close to the diagonal, which supports the Gumbel assumption.

Based on the results presented above, we apply the proposed approach to the case study. Figure 9 summarizes the ten most anomalous groups. Figure 9(a) shows anomalous groups in an orange color scale, where darker shades indicate higher group anomaly scores. To avoid confusion from similar

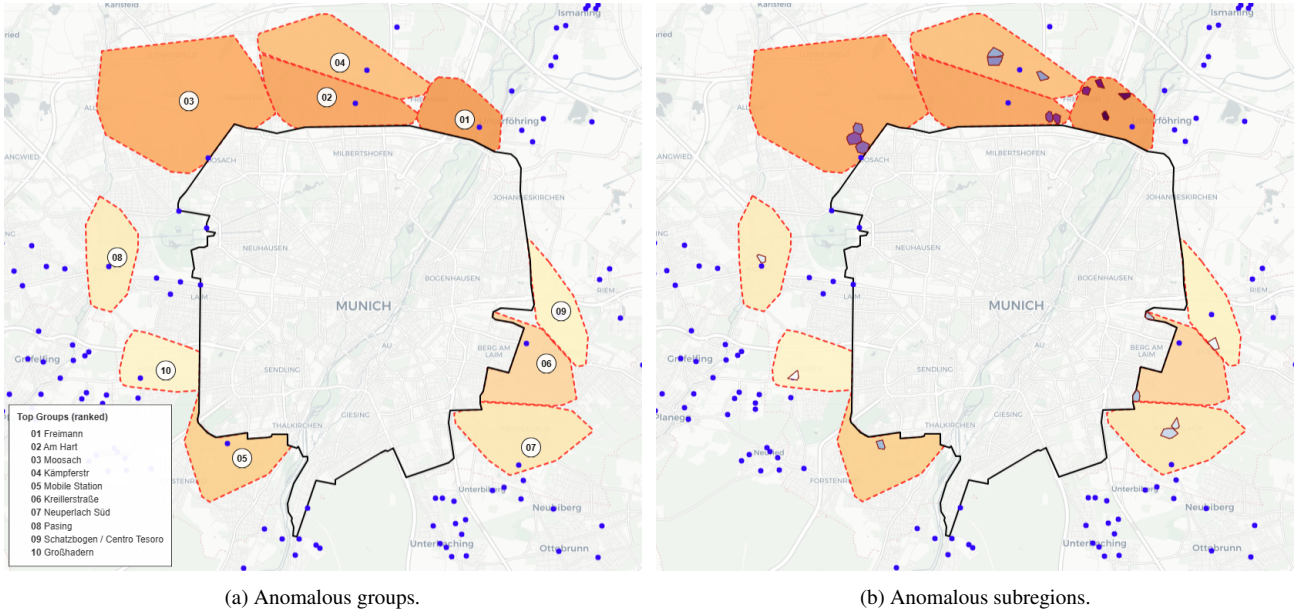


Figure 9: Detected anomalous groups and subregions prioritized for docking station deployment.

Table 3: Tabular summary of subregions (centroids and names) within the ten highest-ranked anomalous groups.

Rank	Group (Station)	Centroid (lat, lon)	Subregion Name
①	Freimann	48.19315, 11.61520	Freimann, A 9
		48.19917, 11.62391	Kleinlappen, Josef-Wirth-Weg 18
		48.20300, 11.61246	Kieferngarten, Kieferngartenstraße 11
		48.20038, 11.60639	Neufreimann, Heidemannstraße
②	Am Hart	48.19201, 11.59342	Freimann, Maria-Probst-Straße 14
		48.19309, 11.58999	Freimann, Maria-Probst-Straße 15
③	Moosach	48.18679, 11.50307	Moosach-Bahnhof, Teplitzer Weg 3
		48.18402, 11.50660	Alt Moosach, Dachauer Straße 312a
		48.18980, 11.50475	Alt Moosach, Wildermuthstraße 85
④	Kämpferstraße	48.21015, 11.56569	Nordhaide, Golddistelanger 8
		48.20538, 11.58679	Freimann, Ingolstädter Straße
		48.21222, 11.56623	Nordhaide, Graslilienanger 7
⑤	Mobile Station	48.09549, 11.51490	Freiland, Kistlerhofstraße 142
⑥	Kreillerstraße	48.11010, 11.62881	Ramersdorf, Leinbergerstraße
		48.13386, 11.63428	Josephsburg, Truderinger Straße 55
⑦	Neuperlach Süd	48.09877, 11.64357	Neuperlach, Von-Knoeringen-Straße 8
		48.10224, 11.64642	Neuperlach, Thomas-Dehler-Straße 9
⑧	Pasing	48.15105, 11.46202	Kolonie I, Wensauerplatz 10a
⑨	Schatzbogen/Centro Tesoro	48.12508, 11.66361	Trudering - Riem, Truderinger Straße
⑩	Großhadern	48.11583, 11.47643	Neuhadern, Würmtalstraße 136

colors on a static map, each group is labeled 1 to 10 to indicate its rank. Figure 9(b) displays the anomalous subregions ($X > u_n$) in purple within each orange group. Priority is global and two-level:

groups are ranked by the group anomaly score, then subregions (exceedances) are ordered by X within each group. Note that the purple intensity follows this two-level order: darker always means higher priority, and no subregion in a lower-ranked group appears darker than any subregion in a higher-ranked group. Table 3 provides the results in tabular form. It lists the group name, the centroid coordinates of each subregion, and the subregion name. Each subregion is given a readable geographic name constructed from the centroid's administrative context and nearest street. The numerical labels in Figure 9 correspond to the Rank column in the table.

Spatially, the highest-ranked groups cluster along Munich's northern axis (Freimann, Am Hart, and Moosach), with additional high-priority groups along the eastern corridor and a southern outlier at Mobile Station. This pattern is consistent with employment growth, residential expansion, and improved public transport access. The BMW Research and Innovation Center opened in September 2020 in Milbertshofen and Am Hart, strengthening a major commuter attractor (BMW Group, 2020). The area around Freimann has seen substantial residential growth, and the Neufreimann (Bayernkaserne) redevelopment is planned to add about 5,500 dwellings for up to 15,000 residents by 2030 (Landeshauptstadt München, 2025). Accessibility also improved with the opening of the U3 metro line extension to Moosach in 2010, which reinforced Moosach as a hub (MVG, 2024). To the east, the pattern aligns with the continued residential development in Messestadt Riem. After a 2021–2022 planning competition, construction of the fifth residential phase began in 2024 and has been progressing since. To the south, renewal of the LMU Klinikum Großhadern campus after the 2021 design competition aligns with the single high-priority group there. Taken together, these employment growth, new housing, and public transport improvements explain why demand concentrates in the north and east and why earlier placement of docking stations no longer matches emerging trip patterns.

At the finer scale, subregions highlight where return records actually cluster. In Freimann, high-ranked subregions lie along the A9 motorway corridor and the Kieferngarten–Heidemannstraße axis (an east–west arterial near Kieferngarten U-Bahn), where returns are more frequent near overpasses and station entrances. In Am Hart, the two Maria-Probst-Straße subregions are situated along a commercial street with shops and offices. In Moosach, Teplitzer Weg and Dachauer Straße are near the Moosach station, so ride endings are concentrated there. In Kämpferstraße, subregions at Golddistelanger and Graslilienanger are situated near exits towards Ingolstädter Straße, and a third subregion lies on the arterial. With limited crossings yet multiple bus and metro access points, ride endings cluster at these gateways. Along the eastern corridor, Ramersdorf–Leinbergerstraße and Josephsburg–Truderinger Straße follow the same arterial-frontage pattern. In the south-east, the two Neuperlach Süd subregions on Von-Knoeringen-Straße and Thomas-Dehler-Straße bridge offices and residential blocks. Schatzbogen/Centro Tesoro lies on Truderinger Straße in the eastern employment belt.

Großhadern falls on Würmtalstraße near the university hospital campus.

4.3. Seasonal analysis

In the previous sections, the MVG Rad records were analysed across the complete period of nine years to support stable long-term infrastructure planning. As a complement, we will extend our analysis by applying the proposed framework to seasonal data. Unlike divisions based on time of day or weekday cycles, seasonal divisions capture temporal shifts in mobility that are particularly meaningful for planning temporary infrastructure. This seasonal analysis serves two main purposes: (i) to examine whether the anomalous spatial patterns identified during the full period remain stable across different seasons; and (ii) to detect seasonal mismatches that can inform temporary docking stations or flexible geo-fencing measures.

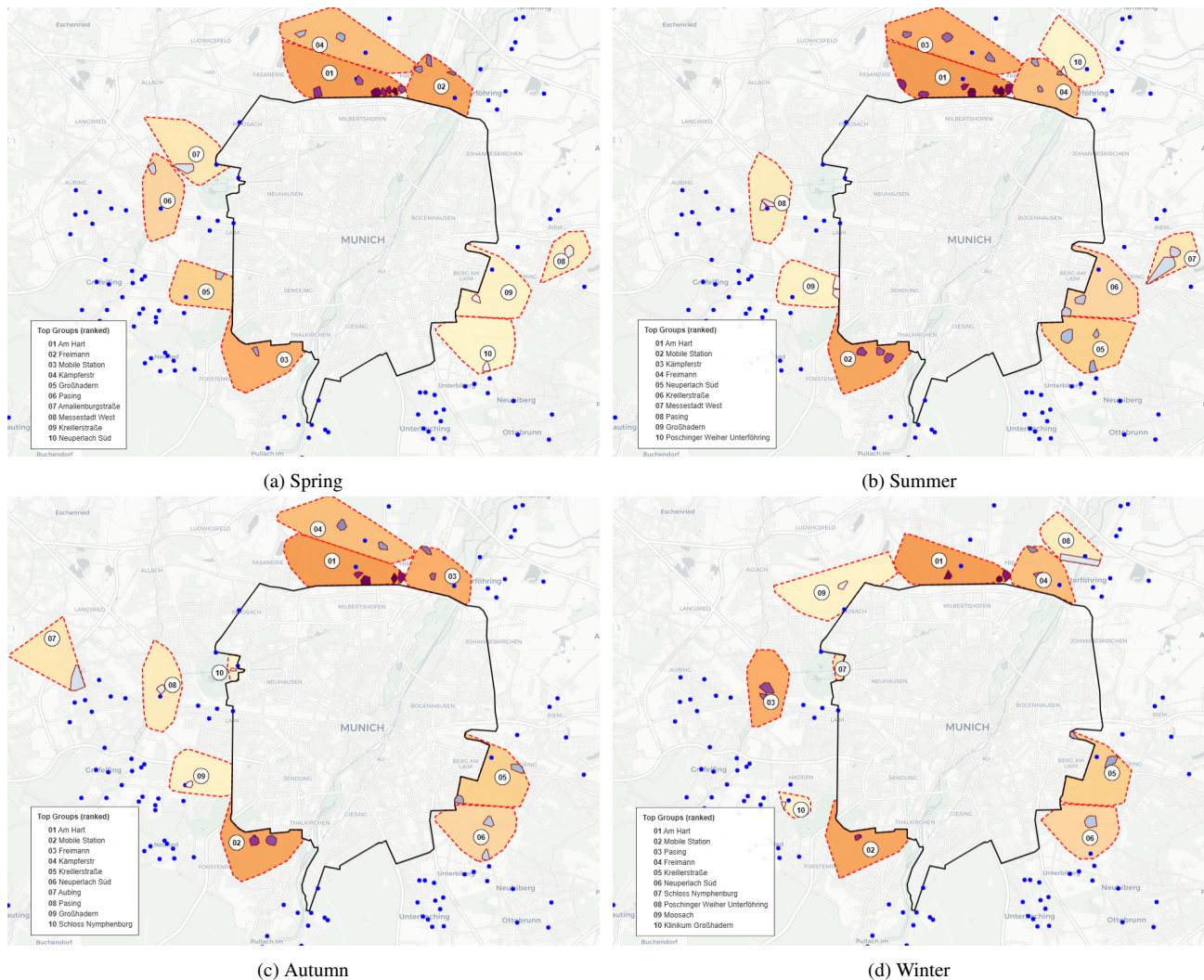


Figure 10: Seasonal variation in anomalous groups and subregions detected by the proposed framework across spring, summer, autumn, and winter.

Figure 10 illustrates the seasonal outcomes of our method, revealing both consistency and fluctuation across the four meteorological seasons. A stable belt in the north from Am Hart to Freimann

appears in every season with high ranks. The western fringe remains visible across seasons, although with lower intensity than the northern belt. South-eastern groups such as Kreillerstraße and Neuperlach Süd are present in most seasons but display larger variation in rank and subregion extent. The persistence of Am Hart and Freimann indicates structural demand, as these areas combine residential density, employment sites, and multiple transit access points. Their subregions are compact and recur in all four seasons, which is consistent with regular commuting and interchange traffic rather than season specific peaks or short-lived event surges.

Some detected subregions exhibit seasonal variability. In summer, subregions extend along the north and the northeast, consistent with longer leisure trips, better cycling conditions, and increased evening activity. Seasonal dock extensions and flexible geo-fencing are recommended near parks and river corridors. During spring and autumn, the overall pattern is similar to the full period baseline, where more western groups are present in the top rank. In winter, Pasing attains its highest rank and Moosach enters the top list. The detection of season-specific anomalies provides actionable guidance for deploying temporary docking stations or adjusting geo-fence boundaries in response to local demand shifts.

5. Conclusions and future work

This study presents a principled framework for identifying and ranking irregular high-demand areas for bike-sharing systems, with the aim of supporting docking station deployment. By leveraging EVT and incorporating prior knowledge of the operational system, our method addresses limitations of existing approaches. First, spatial partitioning strategies depend on hyperparameters that require tuning. Second, many detection models incur a high computational cost. Third, most methods do not provide a principled ranking across candidate areas. The proposed likelihood-based model considers demand intensity, spatial coverage, and neighborhood configuration within a unified framework, allowing for both detection and prioritization of candidate areas. Application to the MVG Rad system in Munich demonstrated the model's effectiveness in uncovering meaningful demand patterns and supporting strategic infrastructure planning. Future work could extend the framework to capacity-sizing of candidate sites in the high-priority groups, which involves determining the optimal number of docks and bikes for each station.

Appendix A.

The proof of Theorem 4 is provided here. The likelihood of a pattern of exceedances $\tilde{\mathbf{x}}^{\text{exc}} = \{x_1^{\text{exc}}, \dots, x_k^{\text{exc}}\}$ is introduced as follows:

$$\begin{aligned} l_n(\tilde{\mathbf{x}}^{\text{exc}}) &:= \mu_k k! \prod_{j=1}^k p(x_j^{\text{exc}}) \\ &= \mu_k k! \prod_{j=1}^k \frac{1}{d_n} \exp\left(-\frac{x_j^{\text{exc}} - u_n}{d_n}\right), \end{aligned} \quad (\text{A.1})$$

where $k \in \{1, \dots, D\}$ denote the size of $\tilde{\mathbf{x}}^{\text{exc}}$. The μ_k can be expressed as follows:

$$\mu_k = P(K = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}, \quad (\text{A.2})$$

where λ is the expected number of exceedances. For the Gumbel case, it can be shown that $\lambda = \lim_{n \rightarrow +\infty} nP(X > u_n) = e^{-u}$.

The likelihood function is derived from the Janossy representation to characterize the pattern of exceedances. In the theorem below, the limiting distribution of the likelihood $l_n(\tilde{\mathbf{x}}^{\text{exc}})$ is established, providing the probabilistic basis for the approach presented in the manuscript. The distribution of likelihoods $l_n(\tilde{\mathbf{x}}^{\text{exc}})$ is asymptotically given by a random variable V_n of mixed type. Random variables of mixed type are neither discrete nor continuous, but are a mixture of both. The discrete component results from the likelihoods $\mu_0 > 0$ that describe cases when there are no exceedances in $\tilde{\mathbf{x}}^{\text{exc}}$. This corresponds to a discontinuity in the CDF that can be described using the Heaviside step function:

$$H(v) = \begin{cases} 1 & , \text{ if } v \geq 0 \\ 0 & , \text{ if } v < 0. \end{cases} \quad (\text{A.3})$$

Theorem. Consider a random variable X with distribution F_X asymptotically having a Gumbel distribution for block maxima. Then there exist sequences $c_n \in \mathbb{R}$ and $d_n > 0$ such that, the following property holds:

$$\lim_{n \rightarrow +\infty} P\left(\frac{X - u_n}{d_n} < x \mid X > u_n\right) = 1 - e^{-x}, \quad (\text{A.4})$$

with $u_n = c_n + ud_n$, for $u \in \mathbb{R}$. V_n is the random variable defined as the Janossy likelihood of the

exceedance pattern $\tilde{\mathbf{x}}^{\text{exc}} = \{x_1^{\text{exc}}, x_2^{\text{exc}}, \dots, x_k^{\text{exc}}\}$ with $x_i^{\text{exc}} > u_n$. That is,

$$\begin{aligned} V_n &:= l_n(\tilde{\mathbf{x}}^{\text{exc}}) \\ &= \mu_k k! \prod_{j=1}^k p(x_j^{\text{exc}}). \end{aligned} \quad (\text{A.5})$$

Then V_n converges in distribution to a random variable V with CDF

$$\begin{aligned} G(v) &= \lim_{n \rightarrow +\infty} P(V_n \leq v) \\ &= \mu_0 H(v - \mu_0) + \sum_{k \geq 1} \mu_k \left[1 - \mathcal{E}_{k,1} \left(-\log \frac{v \cdot (d_n)^k}{\mu_k \cdot k!} \right) \right], \end{aligned} \quad (\text{A.6})$$

where $\mathcal{E}_{k,1}$ denotes the CDF of an Erlang distribution with shape parameter k and scale 1, and

$$\begin{aligned} \mu_0 &= e^{-\lambda} \\ \mu_k &= \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned} \quad (\text{A.7})$$

Proof. Firstly, we determine the distribution of $W_n = -\log(V_n)$ conditional on the number of exceedances k . We have

$$\begin{aligned} W_n &= -\log(l_n(\tilde{\mathbf{x}})) \\ &= -\log(\mu_k \cdot k!) + k \log(d_n) + \sum_{j=1}^k \frac{x_j^{\text{exc}} - u_n}{d_n}. \end{aligned} \quad (\text{A.8})$$

According to the classical EVT, the distribution of exceedances over a high threshold converges to the exponential distribution in the Gumbel case. By the limit result in (A.4), as $n \rightarrow +\infty$, the normalized exceedances $\frac{x_j^{\text{exc}} - u_n}{d_n}$ converge in distribution to a standard exponential distribution:

$$\frac{x_j^{\text{exc}} - u_n}{d_n} \xrightarrow{d} \text{Exp}(1). \quad (\text{A.9})$$

Therefore, the third term of W_n in (A.8), given by $\sum_{j=1}^k \frac{x_j^{\text{exc}} - u_n}{d_n}$, is the sum of k normalized exceedances. From the continuous mapping theorem (stating that convergence is preserved by a continuous transformation), the sum of k such independent exceedances converges to the distribution of a sum of k exponential random variables with scale $\sigma = 1$, which is an Erlang distribution with shape parameter k and rate parameter 1. Define the centered statistic

$$W_n^* := W_n + \log(\mu_k \cdot k!) - k \log(d_n), \quad (\text{A.10})$$

it follows that

$$W_n^* \xrightarrow{d} \text{Erlang}(k, 1). \quad (\text{A.11})$$

The cumulative distribution function of W_n can be derived using the law of total probability. Let $\bar{G}(w)$ denote the limiting cumulative distribution function of W_n :

$$\begin{aligned} \bar{G}(w) &= \lim_{n \rightarrow +\infty} P(W_n \leq w) \\ &= \lim_{n \rightarrow +\infty} \sum_{k \geq 0} \mu_k P(W_n < w \mid k) \\ &= \sum_{k \geq 1} \mu_k \lim_{n \rightarrow +\infty} P(W_n^* < w + \log(\mu_k \cdot k!) - k \log(d_n) \mid k) \\ &= \mu_0 H(w + \log(\mu_0)) + \sum_{k \geq 1} \mu_k \mathcal{E}_{k,1}(w + \log(\mu_k \cdot k!) - k \log(d_n)), \end{aligned} \quad (\text{A.12})$$

where $H(\cdot)$ is the Heaviside step function defined in (A.3). Transforming back to the original distribution by means of $V_n = e^{-W_n}$, one obtains the desired result for $G(v)$:

$$\begin{aligned} G(v) &= \lim_{n \rightarrow +\infty} P(V_n \leq v) \\ &= 1 - \bar{G}(-\log(v)) \\ &= \sum_{k \geq 1} \mu_k - \left[\mu_0 H(-\log(v) + \log(\mu_0)) + \sum_{k \geq 1} \mu_k \mathcal{E}_{k,1}(-\log(v) + \log(\mu_k \cdot k!) - k \log(d_n)) \right] \\ &= \mu_0 (1 - H(-\log(v) + \log(\mu_0))) + \sum_{k \geq 1} \mu_k \{1 - \mathcal{E}_{k,1}(-\log(v) + \log(k! \mu_k) - k \log(d_n))\} \\ &= \mu_0 H(v - \mu_0) + \sum_{k \geq 1} \mu_k \left[1 - \mathcal{E}_{k,1} \left(-\log \frac{v \cdot (d_n)^k}{\mu_k \cdot k!} \right) \right]. \end{aligned} \quad (\text{A.13})$$

□

References

- Ashqar, H. I., Elhenawy, M., Almannaa, M. H., Ghanem, A., Rakha, H. A., and House, L. (2017). Modeling bike availability in a bike-sharing system using machine learning. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 374–378. IEEE.
- BMW Group (2020). Building for the future: BMW group enters a new era of R&D with FIZ nord. <https://www.press.bmwgroup.com/global/article/detail/T0317727EN/building-for-the-future%3A-bmw-group-enters-a-new-era-of-r-d-with-fiz-nord?language=en>. Press release.
- Caggiani, L., Camporeale, R., Ottomanelli, M., and Szeto, W. Y. (2018). “A modeling framework for the dynamic management of free-floating bike-sharing systems”. *Transportation Research Part C: Emerging Technologies*, 87:159–182.
- Chen, W., Liu, X., Chen, X., Cheng, L., and Chen, J. (2025). “Deciphering flow clusters from large-scale free-floating bike sharing journey data: a two-stage flow clustering method”. *Transportation*, 52(1):155–184.

- Cheng, L., Huang, J., Jin, T., Chen, W., Li, A., and Witlox, F. (2023). “Comparison of station-based and free-floating bikeshare systems as feeder modes to the metro”. *Journal of Transport Geography*, 107:103545.
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.
- DeMaio, P. (2009). “Bike-sharing: History, impacts, models of provision, and future”. *Journal of Public Transportation*, 12(4):41–56.
- Du, Y., Deng, F., and Liao, F. (2019). “A model framework for discovering the spatio-temporal usage patterns of public free-floating bike-sharing system”. *Transportation Research Part C: Emerging Technologies*, 103:39–55.
- Duran-Rodas, D., Villeneuve, D., and Wulfhorst, G. (2020). “Bike-sharing: the good, the bad, and the future—an analysis of the public discussion on Twitter”. *European Journal of Transport & Infrastructure Research*, 20(4).
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling Extremal Events: for Insurance and Finance*. Springer, New York.
- Fishman, E. (2016). “Bikeshare: A review of recent literature”. *Transport Reviews*, 36(1):92–113.
- Hua, M., Chen, X., Zheng, S., Cheng, L., and Chen, J. (2020). “Estimating the parking demand of free-floating bike sharing: A journey-data-based study of nanjing, china”. *Journal of Cleaner Production*, 244:118764.
- Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Chichester.
- Jensen, P., Rouquier, J.-B., Ovtracht, N., and Robardet, C. (2010). “Characterizing the speed and paths of shared bicycle use in lyon”. *Transportation Research Part D: Transport and Environment*, 15(8):522–524.
- Jonkeren, O. and Kager, R. (2021). “Bicycle parking at train stations in the netherlands: Travellers’ behaviour and policy options”. *Research in Transportation Business & Management*, 40:100581.
- Landeshauptstadt München (2025). Neufreimann (ehemalige bayernkaserne). <https://stadt.muenchen.de/infos/neufreimann.html>. City planning project page.
- Li, C. and Xiao, X. (2023). “On censoring time in statistical monitoring of lifetime data”. *Technometrics*, 65(3):418–431.
- Liu, Y. and Tian, L. (2021). “A graded cluster system to mine virtual stations in free-floating bike-sharing system on multi-scale geographic view”. *Journal of Cleaner Production*, 281:124692.
- Lohrer, A., Binder, J. J., and Kröger, P. (2022). Group anomaly detection for spatio-temporal collective behaviour scenarios in smart cities. In *Proceedings of the 15th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, pages 1–4.
- Luo, H., Kou, Z., Zhao, F., and Cai, H. (2019). “Comparative life cycle assessment of station-based and dock-less bike sharing systems”. *Resources, Conservation and Recycling*, 146:180–189.
- Moins, T., Arbel, J., Girard, S., and Dutfoy, A. (2023). “Reparameterization of extreme value framework for improved Bayesian workflow”. *Computational Statistics & Data Analysis*, 187:107807.
- Münchner Verkehrsgesellschaft (MVG) (2022). Mvg rad — geschäfts- und rückgabegebiete / freies rückgabegebiet. <https://www.mvg.de/dam/jcr:1fd3dfde-19cb-4be0-ab54-9011d256b984/geschaeftsgebiet-mvg-rad.pdf>. Map; © MVG/SWM (Accessed on 2025-08-19).
- MVG (2024). U-umgebungsplan moosach. https://www.mvg.de/aushangfahrplan/P8_H_M0_0.pdf. Version date: 2024-02-12 (PDF).
- Nikitas, A. (2019). “How to save bike-sharing: An evidence-based survival toolkit for policy-makers and mobility providers”. *Sustainability*, 11(11):3206.
- O’Brien, O., Cheshire, J., and Batty, M. (2014). “Mining bicycle sharing data for generating insights into sustainable transport systems”. *Journal of Transport Geography*, 34:262–273.

- Reiss, S. and Bogenberger, K. (2015). GPS-data analysis of munich's free-floating bike sharing system and application of an operator-based relocation strategy. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 584–589. IEEE.
- Sarkar, A., Lathia, N., and Mascolo, C. (2015). “Comparing cities’ cycling patterns using online shared bicycle maps”. *Transportation*, 42:541–559.
- Shen, Y., Zhang, X., and Zhao, J. (2018). “Understanding the usage of dockless bike sharing in singapore”. *International Journal of Sustainable Transportation*, 12(9):686–700.
- Vo, B., Dam, N., Phung, D., Tran, Q. N., and Vo, B. (2018). “Model-based learning for point pattern data”. *Pattern Recognition*, 84:136–151.
- Wang, Y.-J., Kuo, Y.-H., Huang, G. Q., Gu, W., and Hu, Y. (2022). “Dynamic demand-driven bike station clustering”. *Transportation Research Part E: Logistics and Transportation Review*, 160:102656.
- Wei, Z., Ma, H., and Li, Y. (2024). “A multiscale approach for free-float bike-sharing electronic fence location planning: A case study of shenzhen city”. *Journal of Advanced Transportation*, 2024(1):1783038.
- Zhang, Y., Lin, D., and Mi, Z. (2019). “Electric fence planning for dockless bike-sharing services”. *Journal of Cleaner Production*, 206:383–393.
- Zhao, D. and Ong, G. P. (2021). “Geo-fenced parking spaces identification for free-floating bicycle sharing system”. *Transportation Research Part A: Policy and Practice*, 148:49–63.
- Zhu, B., Hu, S., Kaparias, I., Zhou, W., Ochieng, W., and Lee, D.-H. (2024). “Revealing the driving factors and mobility patterns of bike-sharing commuting demands for integrated public transport systems”. *Sustainable Cities and Society*, 104:105323.
- Zhuang, C., Li, S., Zhuang, H., and Liu, X. (2025). “Mining individual daily commuting patterns of dockless bike-sharing users: A two-layer framework integrating spatiotemporal flow clustering and rule-based decision trees”. *Sustainable Cities and Society*, 118:105985.