Evolutionary Intelligence (2025) 18:73
https://doi.org/10.1007/s12065-025-01062-6

RESEARCH PAPER

Pearclustering: a novel clustering algorithm with an application to bike mobility

Francisco Marquez-Saldaña[1] · Gonzalo A. Aranda-Corral[2] · Joaquín Borrego-Díaz[1]

Received: 7 March 2025 / Revised: 5 May 2025 / Accepted: 3 June 2025 / Published online: 21 June 2025

Abstract
Bike Sharing Systems (BSS) have become a key solution for urban mobility, reducing traffic-related CO_2 emissions. However, managing BSS poses challenges that require data-driven solutions, particularly for understanding their global behavior and forecasting their evolution. These dynamics arise from the interaction among users, companies, dock stations, and city policies, influenced by sociological and infrastructure-based factors. This paper proposes a novel clustering methodology to analyze BSS data across multiple cities. By clustering station-day tuples instead of aggregating statistics, our approach captures seasonal patterns, special events, and weekday/weekend differences. Using Pearson Correlation as a distance metric, it remains robust across different station sizes and system scales. Trained on three European BSS and evaluated across six cities from 4 different countries, our model uncovers meaningful patterns such as work, residential, and leisure areas, as well as seasonal changes even in systems not used in the training process. These insights enhance BSS management, expansion, and decision-making, with applications in monitoring, anomaly detection, and demand prediction.

Keywords Clustering analysis · Machine-learning in mobility · Bike sharing platforms · Artificial intelligence in engineering

1 Introduction

Bike Sharing Systems (BSS) play a key role in modern urban mobility by offering a sustainable and efficient alternative for navigating congested cities. Beyond their environmental benefits, their adoption promotes healthier mobility while reducing carbon emissions and traffic congestion [1]. Therefore, BSS have been integrated into the sustainable transportation ecosystem of modern cities [2, 3], making the understanding of their impact on urban planning a strategic necessity [4, 5]. All of this, positions BSS as a crucial component in the transition towards more sustainable and livable cities [6, 7].

Due to their relevance, BSS have opened several research and development lines, attracting significant academic interest. Social interest in using such services has been analyzed [8], revealing that improving bike lanes and infrastructure could be a key factor for the successful implementation of BSS.

BSS are also considered complex systems with several challenges to be addressed. Improving system topology has been tackled using various data-driven techniques. On the one hand, greedy algorithms can be used to analyze optimal station network expansion options [9]. On the other hand, a Multinomial Logit Model based on bike trajectories can be applied for lane planning [10]. See [11] for a compendium of scientific methodologies in this field.

Another key governance issue is station bicycle rebalancing, a widely researched topic. Approaches range from multi-agent system simulations [12] to custom optimization problem solvers [13, 14], as well as user incentivization methods [15]. Furthermore, techniques and findings

✉ Francisco Marquez-Saldaña
framisal@alum.us.es

Gonzalo A. Aranda-Corral
gonzalo.aranda@dti.uhu.es

Joaquín Borrego-Díaz
jborrego@us.es

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla, Avda. Reina Mercedes s.n., 41012 Sevilla, Spain

[2] Departamento de Tecnologías de la Información, Universidad de Huelva, Avda. de las Fuerzas Armadas s/n, 21007 Huelva, Spain

from BSS-focused studies have been repurposed for other services [16].

The impressive growth of urban mobility data presents significant challenges in processing and extracting meaningful insights. A key feature of BSS lies in the massive generation of data, in some cases producing complex datasets that need to be linked with other urban data sources. Moreover, they require advanced (big) data science techniques to transform raw information into actionable knowledge.

The interest in the system, along with its operational characteristics, makes understanding its evolution and efficiently managing it a strategic challenge for municipalities and service providers. For instance, detecting new patterns emerging from the service's behavior as a complex system can be difficult for specialists (e.g., urban planners and smart city administrators) without clear analysis from a Data Science perspective . Such a gap between domain experts and data scientists is common in knowledge representation [17]. However, in this case, the optimal approach is to bridge the gap by customizing conventional Data Science procedures to visualize the detection of typical patterns (e.g., [18]) while simultaneously unveiling new or overlooked behaviors using the same customization. Dashboard-like solutions enable municipal mobility control rooms to gain insights into both short- and long-term evolution.

1.1 Previous work and limitations

Unsupervised Machine Learning has been established as one of the most widely used tools by researchers to identify common behaviors and patterns in BSS. Therefore, clustering techniques have been extensively applied to various BSS datasets.

First, the station occupancy behavior of the Paris BSS was analyzed by Feng et al. using K-Means and Hierarchical Clustering methods. This study uncovered common station patterns and their impact on the overall system [19]. Conversely, Borgnat et al. focused on applying these techniques to BSS trips rather than occupancy. Their work revealed new patterns, enhancing the overall understanding of BSS dynamics [20].

Furthermore, Vogel et al. employed Expectation-Maximization (EM) and Sequential Information Bottleneck (sIB) algorithms to analyze BSS demand (pickups and returns). Their study demonstrated a clear improvement in performance metrics using EM compared to other methods [21].

External variables and city's behavior impact on resulting clusters was also been studied by Etienne et al. [22]. Relations between city socio-economic variables such as population densitiy, employment, leisure amenities and geo-spatial surface of the station location has been pointed by this work to be linked with clusters.

Clustering techniques not only uncover new patterns but also complement and enhance other Machine Learning methods. The study of this type of behavior has also led to the development of techniques for classifying stations based on spatial usage patterns, which have been considered the foundation of an outlier detection tool by Rennie et al. In that work, the effectiveness of clustering methods in station pattern outlier detection was highlighted [23].

Moreover, Feng et al. addressed the use of clustering-based station classification as a means to generate new input features for machine learning models. Their study demonstrated an improvement in demand prediction accuracy using a gradient boosting model [24].

Another example of a customized clustering model, Compound Stations Clustering, was implemented by Dai et al. to enhance ensemble models [25]. This approach reduced the complexity of the destination prediction problem by locally aggregating stations based on the clustering results presented in their work.

Recent advancements in spatio-temporal clustering (graph-based) has also been applied in BSS field. First, a Distance-Constrained Clustering Algorithm (DCCA) is proposed by L. Chen et al. to group neighbour stations with similar behaviours [26]. Stations were clustered into larger groups based on geographical distance, demand trend and external events history. Then, previous methodology has been modified by H. Zhu, et al. by using demand curve point of interest (POI) features instead of external events history [27]. Both works remark the improvement of considering bigger areas (groups of near stations) rather than station level in BSS demand forecast.

Apart from that, time-series clustering approaches has also been developed by the academia. A dimension reduction methodology (Discrete Wavelet Transform) together with Dynamic Time Warping (DTW) distance metric has been proposed by L. Duo et al. [28]. A bettter classification of BSS stations has been arrived by this work turning DTW distance as probably the state-of-the-art in clustering techniques applied to BSS.

Despite the extensive range of clustering analyses developed for BSS in the literature, most station-level clustering studies have focused on a single system (i.e., one city). This is primarily due to two main factors. On the one hand, the lack of standardized conventions and historical data makes BSS information difficult to access, compare, and share [29]. On the other hand, socio-cultural differences cause similar patterns across systems to shift throughout the day, complicating direct comparisons. These limitations must be addressed to enable a multi-BSS approach and, consequently, a generalized knowledge extraction process.

Concerning information aggregation, existing studies have typically focused on the station level. Therefore, aggregating BSS usage indicators, such as station occupancy or

activity over a specific period, is essential. However, selecting an inappropriate time window for clustering may lead to misleading pattern identification (cf. [30]).

For instance, seasonal and occasional patterns may be overlooked when using a long aggregation period, as the resulting indicator tends to highlight only the most frequent behaviors. Additionally, unrealistic patterns may emerge if aggregation averages out two well-defined but distinct trends at the same station.

In our case—understanding BSS behavior patterns on a global scale rather than a city-specific one—it is crucial to acknowledge that the aforementioned aggregation requires training at least two distinct clustering models: one for weekdays and another for weekends. Otherwise, weekend-specific patterns would not be accurately captured. By separating the training process, the final solution avoids misapplying weekday-derived behaviors to weekends and vice versa, thereby enhancing model generalization.

Addressing existing limitations in clustering techniques within the BSS domain could lead to significant advancements—not only in uncovering new and more robust patterns but also in enhancing other machine learning solutions that depend on prior unsupervised station classification.

While existing studies have primarily focused on single-system analyses, urban mobility patterns are increasingly interconnected. A multi-system approach enables the identification of common usage behaviors across different urban contexts, strengthening the generalizability of clustering models. Furthermore, it facilitates knowledge transfer between cities, allowing insights gained in one location to be effectively applied to others.

1.2 Aim of the paper

This paper introduces a novel multi-system clustering approach designed to extract knowledge from different Bike Sharing Systems (BSS) across various cities and countries. Beyond supporting the classification of stations based on conventional parameters and features (e.g., [31, 32]), the proposed method addresses previously mentioned clustering limitations in BSS applications.

By developing a methodology that goes beyond local specificities, this approach seeks to reveal universal usage patterns while also capturing the distinctive characteristics of each urban setting. This contribution is expected to benefit both practitioners, such as urban planners and mobility companies, and researchers aiming for a more comprehensive understanding of BSS dynamics.

The key contributions of this paper include:

- A semantic day routine binning approach that enables cross-system analysis.
- A correlation-based distance metric that captures usage patterns independent of system scale.
- A novel multi-system clustering methodology for BSS daily patterns instead of overall station behaviors.

1.3 Structure of the paper

This paper is organized as follows. First, the formal specification of the problem and proposed solution is presented in Sect. 2. Then, Sect. 3 focuses on data used to develop this work from its understanding to final selected data passing through data processing. Next, the advantages of using correlation distance instead of other widely used distances in clustering are presented in Sect. 4. After that, the whole modelling methodology, including training dataset sampling, validation metrics definition and baseline model implementation are described in Sect. 5. Then, developed model and its validation are explained in Sect. 6. Next, developed model capabilities and overcome limitations compared to the baseline are explored in Sect. 7. Finally, conclusions and future work are discussed together in Sect. 8.

2 Motivation of the problem, specification and preparation

Regarding BSS knowledge extraction through clustering techniques, most existing studies classify stations within a specific time period based on hourly statistics of features such as occupancy or activity. This methodology requires conducting separate analyses for business days and weekends (Fig. 1, left). When examining long time series (e.g., an entire year), seasonal or day-specific station behavior tends to be significantly smoothed over extended periods and becomes undetectable in shorter ones. Consequently, implementing multiple models is necessary to accurately capture temporal variations.

Although station-level clustering provides a valuable approximation for specific studies, these limitations make a generalizable pattern recognition approach across multiple BSS both challenging and essential.

To address these challenges, BSS clustering must be approached at a deeper level. Daily routines across different systems should be standardized using common short time intervals (bins) throughout the day. Each system must adopt the same number of bins with equivalent semantic meaning, differing only in their specific start and end timestamps. Consequently, the daily routine unification U_i for system i can be represented as:

$$U_i = \{b^{i,1}_{t1,t2}, b^{i,2}_{t2,t3}, ..., b^{i,n}_{tn-1,tn}\} \quad (1)$$

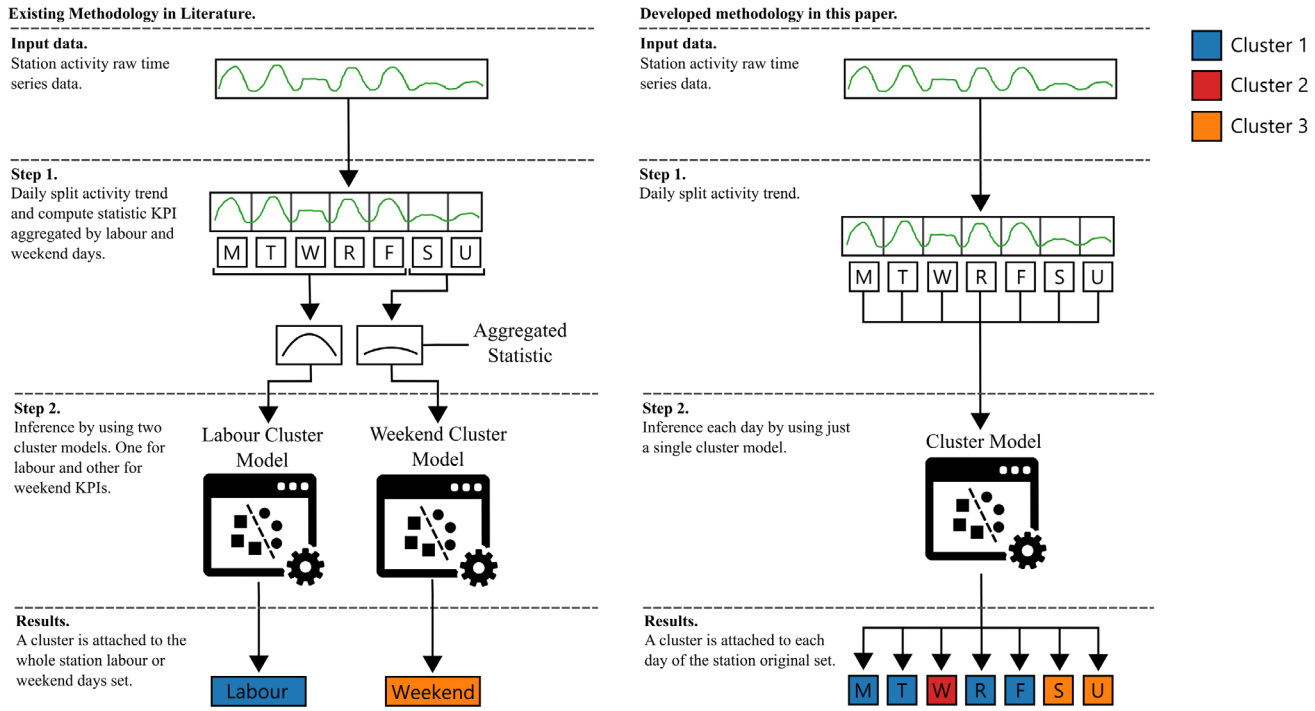


Fig. 1 Two clustering methodologies applied to BSS data. Most followed technique by literature is presented on the left. The new approach implemented in this work is shown on the right

where t represents the start and end times for each bin, and $b^{i,1} = b^{j,1}$ ensures alignment across systems.

Handling seasonality and special events cannot be effectively achieved using a single statistical trend per station. Instead, categorizing stations based on their daily activity trends, rather than relying on aggregated statistics over a given period—offers a more robust solution. This approach leverages the premise that, regardless of the system, station, or season, certain recurring daily behaviors persist. As a result, the proposed clustering method directly classifies daily activity patterns for each station:

$$\vec{st}_d^{ij} = (tk_{b^1}^{ij}, ..., tk_{b^n}^{ij}, rt_{b^1}^{ij}, ..., rt_{b^n}^{ij}) \tag{2}$$

where $tk_{b^n}^{ij}$ is the day median taken bikes for station j in system i at bin b^n, rt_{b^n} is the day median returned bikes for station j in system i at bin b^n and $\vec{st}_d^{j}$ is the vector of activity of station j in system i at day d.

Therefore, considering the set of daily activity during a certain period of n days for station i in system j as:

$$x_{ij} = \{\vec{st}_{d1}^{ij}, \vec{st}_{d2}^{ij}, ..., \vec{st}_{dn}^{ij}\}. \tag{3}$$

Let $X_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$ be the set of daily usage of a system i of n stations. Finally, the target set to be clustered could be defined as:

$$\mathcal{S} = \{X_1, X_2, ..., X_n\} \tag{4}$$

Formally, we hypothesize the existence of a clustering function $\mathcal{C}$ such that:

$$\mathcal{C}(\mathcal{S}) \rightarrow \Gamma \tag{5}$$

where Γ represents a set of generalized usage patterns that maximizes inter-cluster dissimilarity, minimizes intra-cluster variance, and ensures pattern transferability across systems.

As a result, different clusters can be assigned to each day of a station within a single developed model, depending on factors such as seasonality, weekdays vs. weekends, and even local festivities or events. This approach is illustrated in Fig. 1 (right).

Identifying shared patterns from a multi-system perspective is the primary objective of the proposed solution. To achieve this, one year of data from six cities across different countries, each with varying numbers of stations, has been analyzed. Additionally, information has been semantically unified across systems, and a non-conventional clustering distance measure, such as Pearson correlation instead of Euclidean distance, has been employed.

3 Studied data and its processing

In relation to BSS station evolution over time, two key metrics from the literature have been considered. On the one side, occupancy represents the normalized availability of bikes at a given moment in a station [33]. On the other side, taken and returned bikes describe station usage [34], also referred to as activity.

This study focuses on the latter, as occupancy alone cannot differentiate between stations that are used symmetrically, leading to the *Information Aggregation Problem*. Specifically, in stations where bikes are taken and returned at the same rate and proportion, occupancy appears constant, even if actual usage varies significantly throughout the day, as illustrated in Fig. 2.

It is important to note that, in this work, occupancy has been computed as the hourly mean of available occupancy records, whereas both taken and returned activity have been processed as the hourly sum.

3.1 Data preparation

In addition to selecting the most informative input metric, the sampling frequency must also be determined. At this stage, a wide range of intervals, from 5-minute to 6-hour periods, can be used depending on the required precision. As previously mentioned, differences in time schedules between locations or countries need to be minimized to develop a multi-system single model. These discrepancies, such as workday habits, may cause two stations from different systems to exhibit the same daily uptake or return evolution, but with a lateral shift in their curves.

To address this issue, a data processing method has been implemented to transform numerical time frequencies into *semantic intervals* with consistent meaning across different systems. These bins have been defined based on the analysis of daily usage patterns in the target systems.

Figure 3 shows that the considered systems exhibit very similar usage patterns, albeit shifted throughout the day, confirming the need for the daily unification developed in this work. A simple exploratory analysis reveals five well-defined periods (three peaks and two valleys) through graphical inspection. Consequently, after evaluating several approaches, the following semantic bins have been defined: "*night*", "*go to work*", "*morning*", "*lunch*", and "*evening*".

Each bin's metric value is calculated using the hourly median of taken or returned bikes. Once the time range for each system and bin is determined, daily evolution can be properly compared across systems, as shown in Fig. 4. This binning process helps reduce the well-known noise associated with this type of data [13] and even mitigates distortions caused by maintenance actions.

3.2 Processing static stations

A second challenge arises from the reduction of records after daily binarization, which prevents the application of commonly used techniques found in the literature for detecting constant and low-used trends. To address this, a simple yet precise statistical method based on data analysis has been implemented in our solution.

Focusing on the whole-day activity distribution could not make the difference between quite low usage days and those ones with just high usage in a single bin (see Fig. 5). Whereas the first one needs to be considered static, the second one might be interesting to understand station's behavior. Therefore, analyzing bin activity individually is required to face static days recognition needs.

The high prevalence of low-usage bins is evidenced by the prominent peak observed near zero activity (see Fig. 6A). Consequently, a bin interval is classified as having real activity when at least two bikes are taken or returned per hour. Based on this criterion, an instance is considered static if none of its bins exhibits real activity (see Fig. 6B).

Using this methodology for recognizing and filtering static instances before clustering provides several advantages. First, as shown in Fig. 6C, focusing on individual bins rather than overall daily activity ensures that instances with only a single bin of real activity (blue curve, left

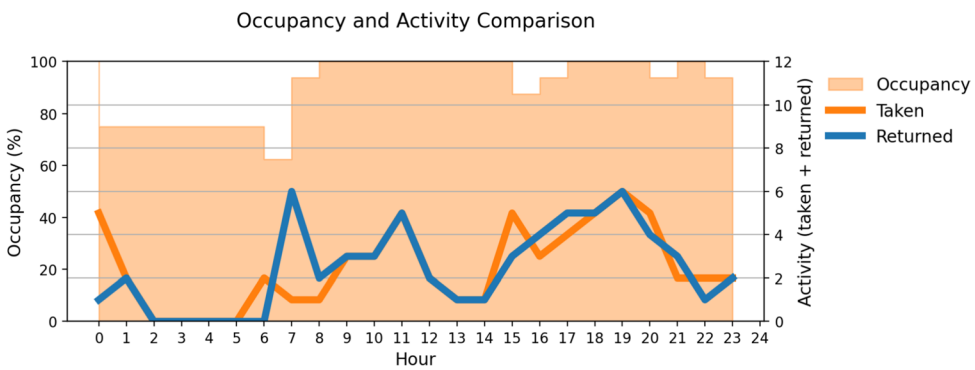


Fig. 2 Occupancy and activity comparison for Dublin's station 10 in July the 9th, 2016. When taken and returned bikes activity trend differ, occupancy change (between 6 and 8 h). However, when they present a similar curve, occupancy keeps constant despite station could be overused (between 9 and 14 h)

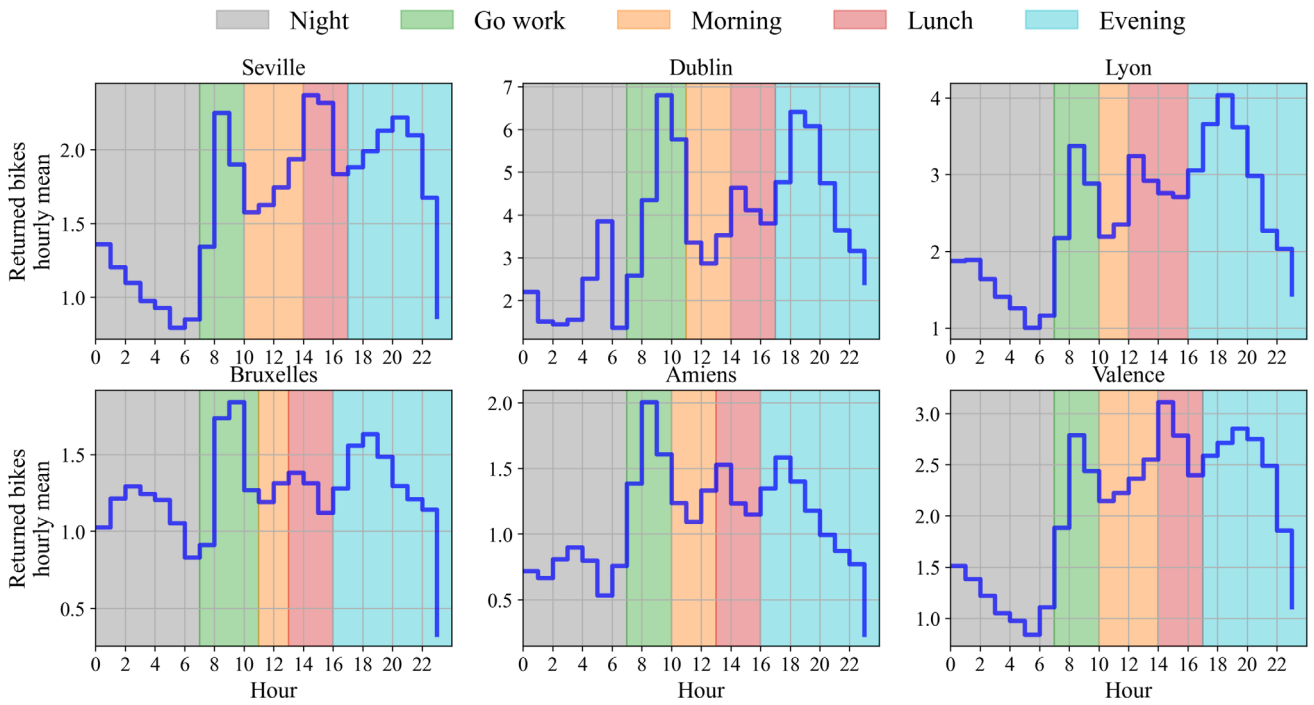


Fig. 3 Hourly median usage over defined intervals for each studied system. Despite all systems present almost the same returned bikes curve, it is displaced in time (X axis)

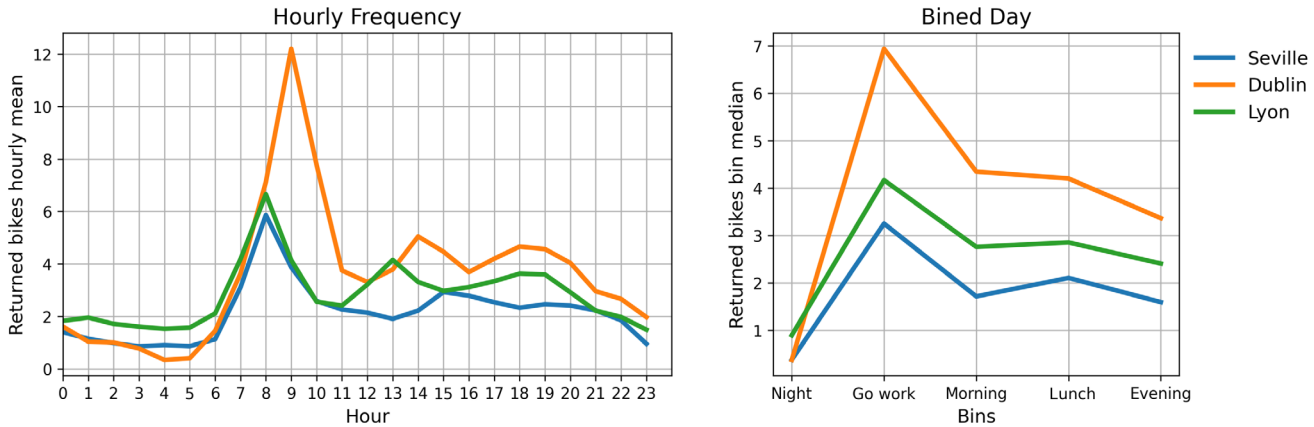


Fig. 4 Hourly frequency (left) comparison against binarization intervals (right) of system used in the training process. In relation to hourly frequency, notice that despite Seville and Lyon present same pattern in the morning, Dublin citizen tends to go work 1 h later. Moreover, having lunch seems to be different on each of them. Nevertheless, when binarizing hour, this horizontal displacement disappears and, therefore, enables system comparison

slope) are not overlooked. Additionally, the simplicity of this detection method reduces computational costs, as static instances do not need to be processed by the clustering model. Therefore, static days can be directly classified and excluded from both the training and inference processes.

3.3 Dataset splitting for training and testing

The input data consists of information from six different European BSS studied over an entire year, as the main objective of this work is to develop a multi-system clustering methodology. These systems are located in different countries, which introduces cultural and behavioral

Fig. 5 Trend difference between two instances with the same total daily activity. One considered as static (red line) and the other thought as no static (blue line)

differences as well as a wide range of weather conditions—factors that are highly relevant to the problem [35].

Moreover, system size, measured by the total number of stations, has also been considered, as daily bike usage trends may vary depending on the system's scale (see Table 1).

It is worth noting that, since the data collection methodology records only station status changes [36]—defined by the number of bikes available for rent, docks for returns, and total stands at a given time—two systems of similar size may exhibit significant differences in recorded data. For example, although Lyon and Brussels have comparable system sizes, the latter has recorded nearly half as many changes as the former, highlighting distinct usage patterns.

Deciding on the training and test sets is another crucial aspect. Selecting a small percentage of stations from all systems helps preserve the original variability and, consequently, enhances the knowledge extracted by the clustering algorithm. However, the developed model is expected to enable future analysis of any system, regardless of whether it was included in the training set or not. Therefore, three

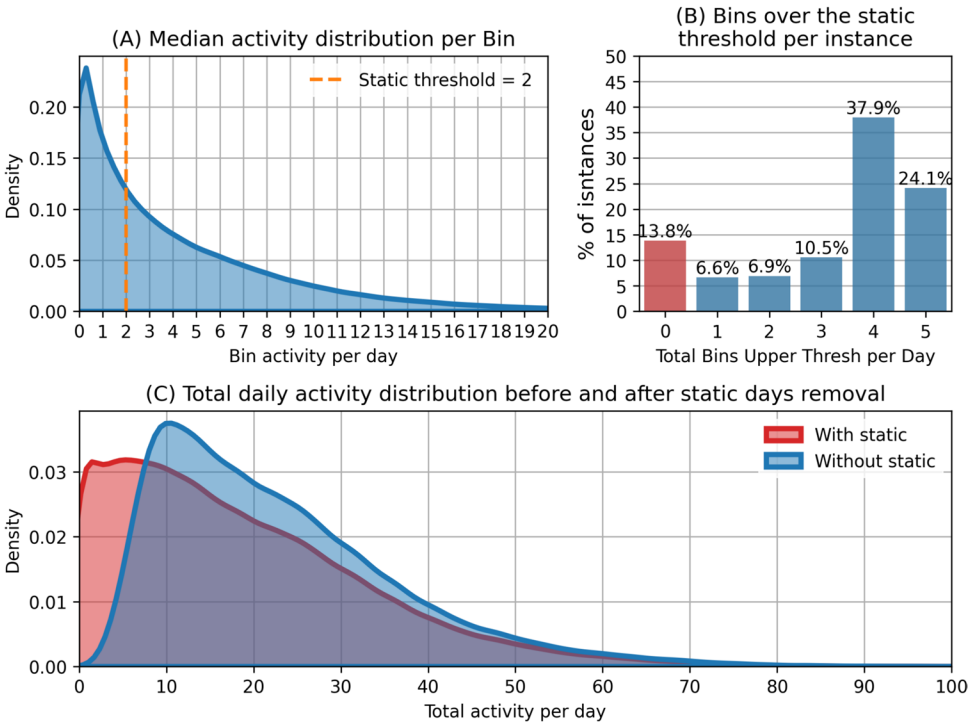


Fig. 6 Static days detection methodology summary. **A** Median activity distribution per bin shows high existence of no used stations at certain time. **B** Filtering bins with less than 2 movements shows that 13.8% of the records has no active bins in the whole day. **C** Total activity per day reflects a displacement to the right side after removing static days

Table 1 Selected BSS data summary

System city	City population	Total stations	Habitants per station	Total records	Data size (MB)	Purpose
Seville	690.566	260	2.656	1400513	167,41	Training
Lyon	1.661.000	348	4.772	2202558	282,56	Training
Dublin	544.107	101	5.387	641180	79,90	Training
Valence	790.201	276	2.863	1774073	218,51	Test
Amiens	135.429	26	5.208	105694	11,17	Test
Bruxelles	2.018.000	354	5.700	1213968	133,76	Test

systems (Valence, Amiens and Bruxelles) have been designated as the test set to evaluate clustering results and assess whether patterns learned from some cities can be applied to unseen ones as shown in Table 1.

Moreover, as each station is represented by the set of clusters assigned to its individual days (explained in Sect. 2), there is no need to predefine the time period to be analyzed or to distinguish between weekdays and weekends.

4 Pearson correlation as cluster distance

Widely used distance metrics in clustering, such as Euclidean or Manhattan distances [37, 38], evaluate similarity based on both trend and magnitude. However, when classifying station-day evolution patterns, it is essential to disregard station size, as two stations may exhibit the same behavior but with different magnitudes. In other words, the same trend may appear shifted along the Y-axis, as illustrated in Fig. 4.

To address this issue, Pearson Correlation [39] has been selected as the distance measure. This choice ensures that clusters exhibiting the same behavioral pattern but on different scales are not erroneously separated by the training algorithm.

Apart from that, working with station activity requires dealing with two signals: taken and returned bikes, which have been addressed in the literature using several techniques. For instance, these variables can be merged into a single one by simply adding their values, generating the total activity for the bin [24]. Alternatively, they can be combined within the same vector [22], as the distinction between taken and returned bikes carries significant meaning for the tackled problem.

Nevertheless, making a relationship between taken and returned bikes within the same bin is not supported by the aforementioned methodologies. Although this relationship could be established by concatenating both metrics into a single vector, the algorithm would not only associate taken and returned values within the same bin but also across different period bins, introducing noise into the problem.

To overcome these difficulties, a different approach has been developed in this work. First, the station's taken and returned bikes vector for a specific day has been defined as the target instance: each instance includes values for each calculated bin on a single day.

$$\vec{tk}_d^{ij} = (tk^{ij}_{night}, tk^{ij}_{gotowork}, tk^{ij}_{morning}, tk^{ij}_{lunch}, tk^{ij}_{evening}) \tag{6}$$

$$\vec{rt}_d^{ij} = (rt^{ij}_{night}, rt^{ij}_{gowork}, rt^{ij}_{morning}, rt^{ij}_{lunch}, rt^{ij}_{evening}) \tag{7}$$

Where $\vec{tk}_d^{ij}$ is the taken bikes vector of station i in system j on day d, and $\vec{rt}_d^{ij}$ is the returned bikes vector of station i in system j on the same day d. It must be mentioned that each vector tuple ($\vec{tk}_d^{ij}$ and $\vec{rt}_d^{ij}$) for the same day will be considered as an input instance.

In this approach, for each instance, the correlation with all other instances is first calculated separately for the taken and returned vectors. These correlation values are then combined into a single vector:

$$\vec{Total}^i_{corr} = \begin{bmatrix} tk^1_{corr} \\ tk^2_{corr} \\ \vdots \\ tk^n_{corr} \end{bmatrix} + \begin{bmatrix} rt^1_{corr} \\ rt^2_{corr} \\ \vdots \\ rt^n_{corr} \end{bmatrix} \tag{8}$$

Where $\vec{Total}^k_{corr}$ is the correlation vector of instance k with the other instances, tk^n_{corr} represents the correlation between the taken bikes vector of instance k and instance n, and rt^n_{corr} represents the correlation between the returned bikes vector of instance k and instance n.

In this representation, negative correlation needs to be addressed using this formulation. While a strong negative correlation may indicate an inverse statistical relationship, in the context of daily-trend similarity, it is interpreted as a completely different behavior. Since the goal of this methodology is to match similar daily curves rather than establish statistical dependence, negative correlations have been set to 0. Hence, this approach prevents cases where taken and returned vectors exhibit inverse relationships from resulting in a correlation of 0.

Finally, each correlation vector for each instance will serve as the input to the clustering model. As a result, day tendencies in which bikes are taken at the same bin but returned at different ones can be distinguished using this distance calculation. This highlights the advantage of Pearson Correlation over other distance metrics in this context.

It must be mentioned that, as this calculation is computationally complex and might require a lot of hardware and time resources, it is required to meet the balance between size and knowledge representation when selecting the training set.

5 Modelling methodology and validation

5.1 Training data selection methodology

Due to the computational complexity of correlation distance metric explained in Sect. 4, the use of 258.785 available instances (365 days times 709 stations, see Table 2 for more details) in the training results is impossible to handle. Thus, records sub-sampling is necessary to achieve this work.

Table 2 Available instances for training process

BSS city	Total stations	Total days	Total bins
Seville	260	365	474.500
Lyon	348	365	635.100
Dublin	101	365	184.325

The final training population dataset must include data from different seasons, weekdays and weekends, and special days to maximize extracted knowledge. However, given the significant number of static days and the fact that dynamic ones typically follow well-defined patterns, selecting 4,000 instances per system for training the model has been deemed sufficient to capture existing behaviors.

Nevertheless, training with 12,000 instances (4,000 per system) introduces the curse of dimensionality [40]. Thus, selection cannot be randomized, as this would increase the typical sparsity in high-dimensional data and, consequently, the impact of noise and lack of heterogeneity on the cluster training process.

Feature selection is one of the most commonly used techniques for addressing the aforementioned issue. However, variable pruning is not feasible due to the intrinsic nature of the clustered input. Since each column represents a correlation between two instances rather than regular features, pruning would effectively mean removing training instances, leading to information loss.

For that reason, a two-step solution has been implemented to tackle the problem of data selection. First, 40 stations per system have been chosen based on domain criteria. Then, a 100 days set for each of them has been considered to maintain station type representation. Concerning stations, location is one of the most influential factors in usage patterns [41]. Thus, the hours of use may vary depending on the distance to the city center. Stations in the city center might be used for a wide range of purposes, such as tourism, leisure, or commuting, whereas those located near the suburbs are more likely to serve large residential areas. Additionally, stations situated far from the city center are typically used less frequently or may even be considered static.

Station usage increases as the distance to the city center decreases (see Fig. 7), suggesting that central stations exhibit a richer variety of patterns compared to suburban ones. As a result, stations have been categorized into three rings: near, medium distance, and far from the city center. The final selection must include 40%, 40%, and 20% of stations from each category, respectively. To achieve this condition-based selection, Genetic Search [42] emerges as the ideal solution. The selected stations for each system are shown in Fig. 8.

Once target stations set has been selected, it is required to choose what day-trend to keep for each. 100 days randomly selected seems to be enough to maintain seasonality.

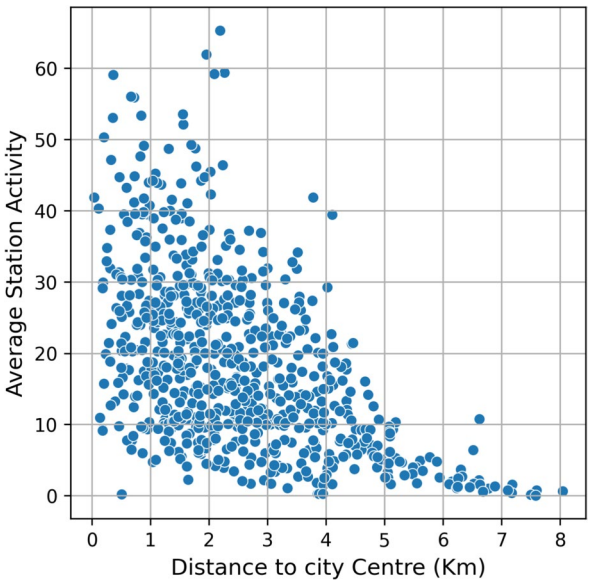


Fig. 7 Average activity relation with distance to city centre per station Seville, Lyon and Dublin (system used in training)

However, weekend days minority would be increased. Consequently, days have been chosen by a stratified sampling methodology [43] taking into account whether is weekend or not.

5.2 Cluster validation metrics

Selecting the right clustering validation methodology is considered one of the key aspects of the model training process. The chosen metric might not only directly influence what algorithm to be used, but also how many clusters (k) will drive to the best performance. Defining the correlation matrix as the input for clustering requires careful consideration when selecting the appropriate metric. Therefore, addressing the issue of high dimensionality is essential.

It is true that Silhouette [44] and Davies-Bouldin [45] scores are widely used and generally perform well in most cases. However, they tend to introduce stability issues when applied to high-dimensional data [46]. In contrast, the Calinski-Harabasz index [47] remains stable as dimensionality increases [46].

However, most of the original implementations of clustering metrics rely on the Euclidean distance between cluster centroids or points. In this case, interpretable results cannot be obtained when using a correlation matrix as input. Therefore, it is necessary to adapt the selected metric to work with the correlation matrix.

Silhouette score modification has been defined in this work as it provides a good balance between between-cluster and within-cluster dispersion. Hence, original implementation based on distance is presented next:

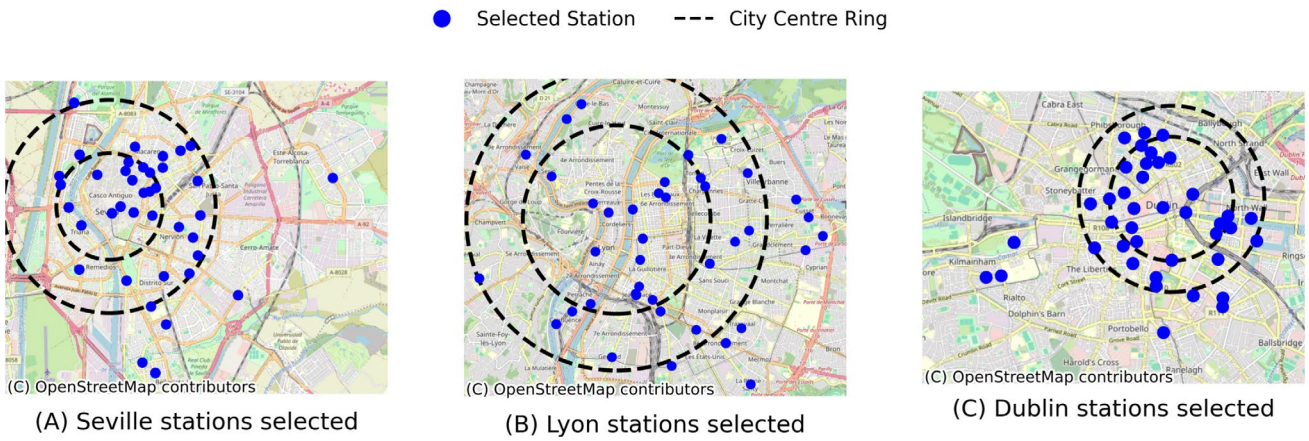


Fig. 8 Selected stations per system for training cluster. Each black ring marks distance to city centre from near, medium and far respectively. Notice that distance has been measured in Kilometers (Km). 1 km corresponds to 1000 m in international system of units (SI)

$$\mathrm{SIL}(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

$$\mathrm{SIL} = \frac{1}{N}\sum_{i=1}^{N} SIL(i) \tag{10}$$

Where $a(i)$ is the mean intra-cluster distance for sample i in cluster a, $b(i)$ is the mean distance to the nearest cluster $b \neq a$ for sample i, and $SIL(i)$ is the Silhouette score for sample i. Lastly, the overall SIL value is computed as the average Silhouette score across all samples.

The previous implementation based on distance has been modified as

$$\mathrm{SIL}_{corr}(i) = \frac{b_{corr}(i) - a_{corr}(i)}{\max\{a_{corr}(i), b_{corr}(i)\}} \tag{11}$$

$$\mathrm{SIL}_{corr} = \frac{1}{N}\sum_{i=1}^{N} SIL_{corr}(i) \tag{12}$$

Where $a_{\mathrm{corr}}(i)$ is the mean correlation between sample i from cluster a and all samples in cluster a, $b_{\mathrm{corr}}(i)$ is the mean correlation between sample i and all samples in the nearest cluster $b \neq a$, $SIL_{\mathrm{corr}}(i)$ is the Silhouette correlation score for sample i, and the overall SIL correlation value is computed as the average Silhouette correlation score across all samples.

Finally, using Silhouette score with correlation matrix as input has been enabled by the presented modification. This keeps rationale behind original implementation. Score values are in range $[-1, 1]$ in which higher values involve a better clusterization.

5.3 Baseline clustering model

Demonstrating how existing limitations will be overcome by the proposed methodology in this work has been considered necessary. To this end, two baselines clustering models have also been developed based on the best previous results found in the literature.

Regarding the input data for the baseline models, hourly activity (taken and returned bikes) is considered. The taken and returned vectors have been concatenated to enable the baseline model to learn taken-returned patterns.

It is expected that the baseline model will provide a station characterization. Hence, the input will be the median taken and returned bikes per hour of the day during the analyzed period. The baseline clustering input would be:

$$\vec{A}_i = (\tilde{tk}_{00}, \tilde{tk}_{01}, ..., \tilde{tk}_{23}, \tilde{rt}_{00}, \tilde{rt}_{01}, ..., \tilde{rt}_{23}) \tag{13}$$

Where $\vec{A}_i$ is the activity input vector for station i, $\tilde{tk}_h$ is the median taken bikes for hour h, and $\tilde{rt}_h$ os the median returned bikes for hour h.

Concerning clustering algorithm selection, several approaches, from K-Means to Hierarchical and Expectation Maximization, have been explored in academia. Results may vary depending on the input data and evaluation metrics used, but similar patterns tend to emerge across most methods. As a result, several approaches have been selected for implementing the baseline model to ensure most used solutions in the literature to be covered.

Despite defining daily curves based on a statistic has been the most used approach by the academia, used distance may vary depending on the work. On the one hand, Euclidean distance has been used the most by the academia arriving

to good results. On the other hand, Dynamic Time Warping (DTW) [48] has become the state-of-the-art for time-series clustering in recent years being also applied to BSS analysis. Therefore, a K-Means with DTW had been developed and compared with presented approach in this work.

Moreover, since aforementioned baseline model is based on a statistic for each station, two clustering approaches must be trained separately for working and weekend days.

Apart from that, correlation techniques have also been explored inside time-series clustering field. K-Shape algorithm and its use of cross-correlation [49] have become the state-of-the-art in recent years. Hence, a comparison between the developed method in this work and K-Shape seems mandatory.

Notice that no relevant application of previous cluster technique to BSS has been found. Furthermore, measuring differences between the both presented correlation approaches (developed in this work and cross-correlation in K-Shape) has been considered the main target of this comparison. Thus, K-Shape has been implemented over daily series rather than daily statistic curves to measure both solutions capabilities.

Finally, the cluster names identified by the baseline models have been unified with those of the new methodology presented in this work, based on the average hourly taken and returned pattern similarity. This allows for a direct comparison between both solutions.

6 Developed clustering model

The cluster training process has been developed in two steps. First, selecting the most suitable clustering algorithm from the wide range of available options is required. Then, defining the appropriate number of clusters (K) may be necessary depending on the chosen method.

The choice of an appropriate clustering technique for a given problem has been extensively discussed in the literature. Several approaches, ranging from plotting projections to employing complex frameworks as discussed in [50], may be applied. Among them, evaluating with an index is one of the most accepted methodologies, particularly for problems where ground truth is not available. Therefore, the previously defined modification of the Silhouette Score has been used for both selecting the clustering algorithm and determining the optimal number of clusters.

Dimensionality reduction methods, such as PCA, t-SNE, or UMAP [51], are commonly used techniques for clustering high-dimensional data. These methods focus on aggregating characteristics into a smaller set of features, allowing most clustering algorithms and evaluation indices to be applied effectively. However, in this problem, each dimension represents an instance rather than a feature. This fundamental difference in meaning renders such tools inapplicable without leading to information loss and increasing the risk of overfitting.

Due to the fact that correlation distance has been defined as the distance metric, as well as the requirement of dealing with the course of dimensionality, clustering algorithm type has to be taken into consideration. Despite Density-based methods reduce training parameters search time as there is no need to define the number of clusters (K), the methods perform badly in both evaluation and run-time when using high-dimensional data without dimensionality reduction [52].

Consequently, centroid-based and hierarchical clustering techniques have been explored to address the problem. More specifically, K-Means, Bisecting K-Means, PAM, and the WARD algorithm have been studied using the Scikit-Learn implementation (see https://scikit-learn.org/stable/modules/clustering.html).

To determine the most suitable tool for classifying station behavior, the target number of clusters, K, has been set from 2 to 10. A model has been trained for each algorithm and each possible value of K. Finally, the results of each execution have been evaluated using the modified Silhouette score. Based on the results, K-Means has been chosen as the final clustering method, as it achieves the best score (see Fig. 9).

Once the clustering technique has been selected, it is necessary to determine the optimal number of clusters. Using the Silhouette score to decide among different values of K, higher scores indicate greater inter-cluster divergence and stronger intra-cluster cohesion. In this case, exploring different values of K for K-Means reveals that the best peak occurs at K=4, as shown in Fig. 9.

Beyond identifying the best K, an additional sensitivity analysis has been conducted for values near the peak Silhouette score. This double-check ensures that no significant patterns are overlooked due to evaluation index limitations. As illustrated in Fig. 10, choosing $K = 3$ instead of $K = 4$ would lead to the loss of relevant patterns, such as "Cluster 1". Consequently, values of $K < 4$ result in missing essential patterns.

Furthermore, "cluster 4" from $k = 5$ and $k = 6$ could be interpreted as a variant of "cluster 3". Similarly, "cluster 5" from $k = 6$ might also be considered a variant from "cluster 1". Consequently, $k > 4$ values drives to found patterns variants instead of new relevant behaviors.

As result, $k = 4$ has been found as the best number of cluster based on both checking the Silhouette index and analyzing the k value sensibility.

6.1 Interpreting behaviors recognized by clustering

The training process results, along with the final clusters, are shown in Fig. 11. Additionally, each cluster has been

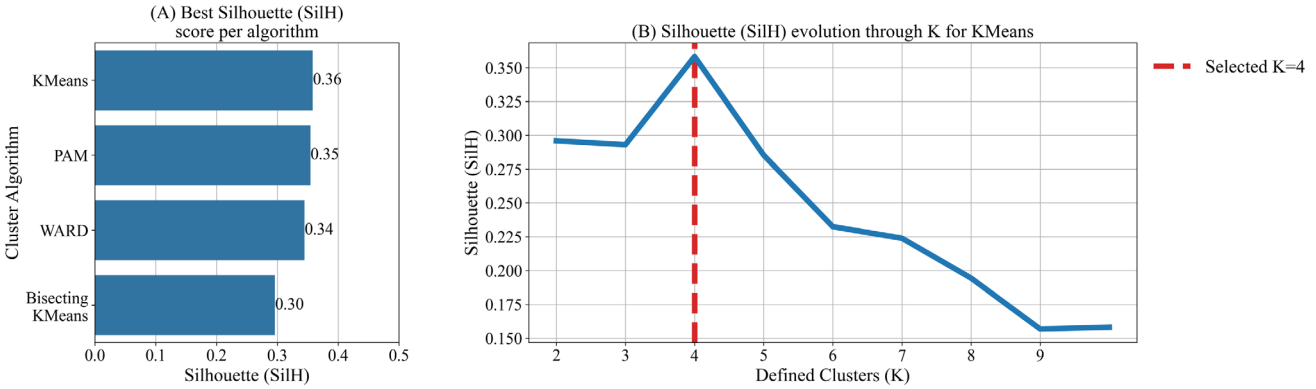


Fig. 9 Correlation Silhouette analysis of explored clustering algorithms. **A** Average score for selecting what method to be used. **B** *K* exploration score evolution for K-Means

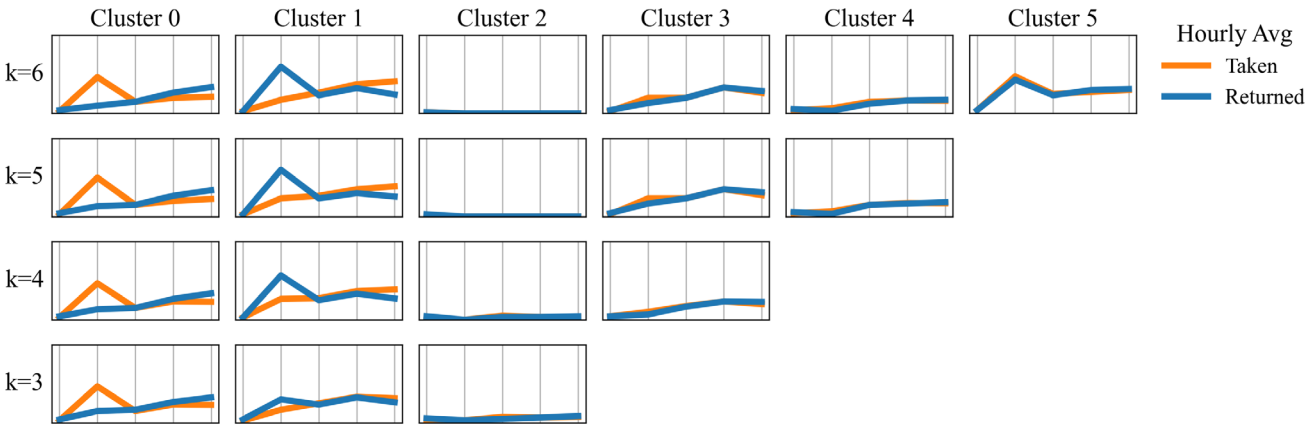
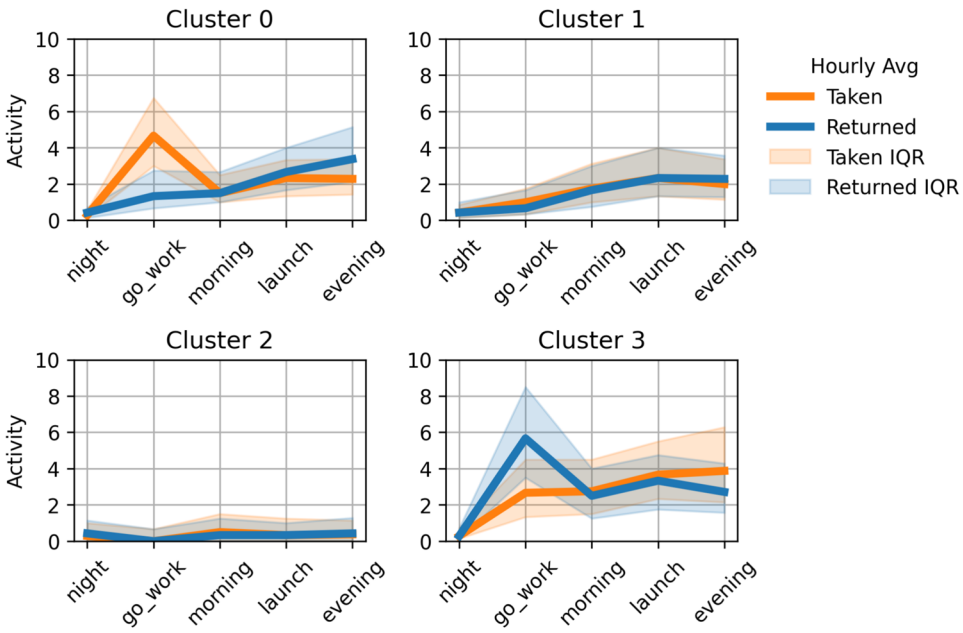


Fig. 10 Found clusters for several k values (between 3 and 6) near the best Silhouette score ($k = 4$). Notice that some interesting patterns are lost for $k < 4$ values. Moreover, variants from other patterns are found when $k > 4$

Fig. 11 The clusters obtained after the training process represent the evolution of the bicycles taken and returned over the course of a day



analyzed from a domain-knowledge perspective, considering both taken and returned bike behavior.

The obtained clusters also present other relevant individual dissimilarities:

- **Cluster 2** reflects regular and low usage. Its main difference from previously filtered inactive station days lies in the fact that clustered instances have at least one change in all the bins during a day.
- **Clusters 0 and 3** exhibit the same pattern but alternate between taken and returned bikes, respectively. Moreover, these trends usually appear during the week, reinforcing their relationship. It can be inferred that, normally, bikes are taken from stations in cluster 0 and returned to stations in cluster 3 during weekdays. Thus, they represent stations near residential and working areas.
- **Cluster 1** shows an activity peak that increases gradually throughout the morning and remains high during lunchtime and the evening. Despite this behavior potentially representing a wide range of user actions, for consistency, cluster 1 can be defined as "leisure" stations.

It is worth to mention that static days set, detected before applying clustering, will be added to the final solution classification as an extra cluster.

6.2 Using the model to classify new instances

Developing a clustering model based on correlation distance suffers a clear disadvantage. Other distance metrics such as Euclidean allow to directly assign new instance clusters by directly using taken and returned bikes vector. However, this inferring methodology could not be applied when the model is trained with correlation.

As a result, calculating new correlation vectors against training set ones is required for classifying new instances by using the trained model. (see Fig. 12b). That is imperative cause learnt centroids have been obtained from training correlation matrix (Fig. 12a). Therefore, using correlation vectors between instances not used in the training process would involve re-training the model every time upcoming records need to be classified. Consequently, training set instances have been saved to properly compute centroid distances at inference time.

Despite the clustering model has been trained only with 12.000 instances (3 systems, 40 stations per system and 100 days per station), extracted knowledge could be applied to new instances even from never seen systems. Hence, all available instances (239.440) from validation systems (Valence, Bruxelles and Amiens) have been classified by using this methodology. Computational efficiency of this predicting methodology is discussed in Sect. 7.2.

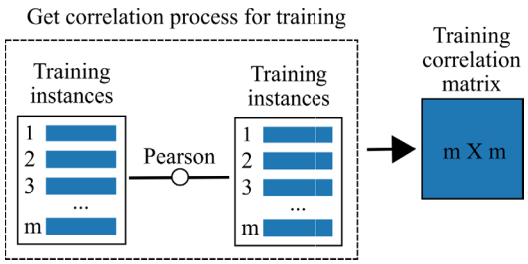
Finally, the whole model implementation methodology from original data preparation to inferring process is shown in Fig. 13.

6.3 Clustering validation

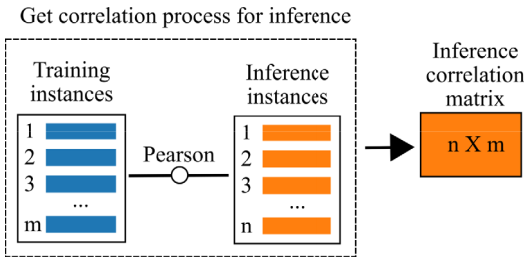
Evaluation metrics might help when verifying within and between cluster distances. Nevertheless, analyzing from a domain knowledge perspective usually provides a more coherent approach to ensuring that the extracted knowledge is correct. This fact becomes even more relevant in the multi-system approach presented in this work, as one of its main purposes is to apply knowledge from the trained system to previously unseen ones.

For that reason, to validate the developed clusters, a two-step methodology based on domain information has been designed.

The whole year data for all systems has been inferred, including both training and test ones. Notice that, as only less than 5% of the selected system data has been used for



(a) Training correlation process. Each training instance is correlated with themselves.



(b) Inference correlation process. Each inference instance is correlated just with all training instances but no with other inference instances.

Fig. 12 Correlation methodology for both training and inference processes

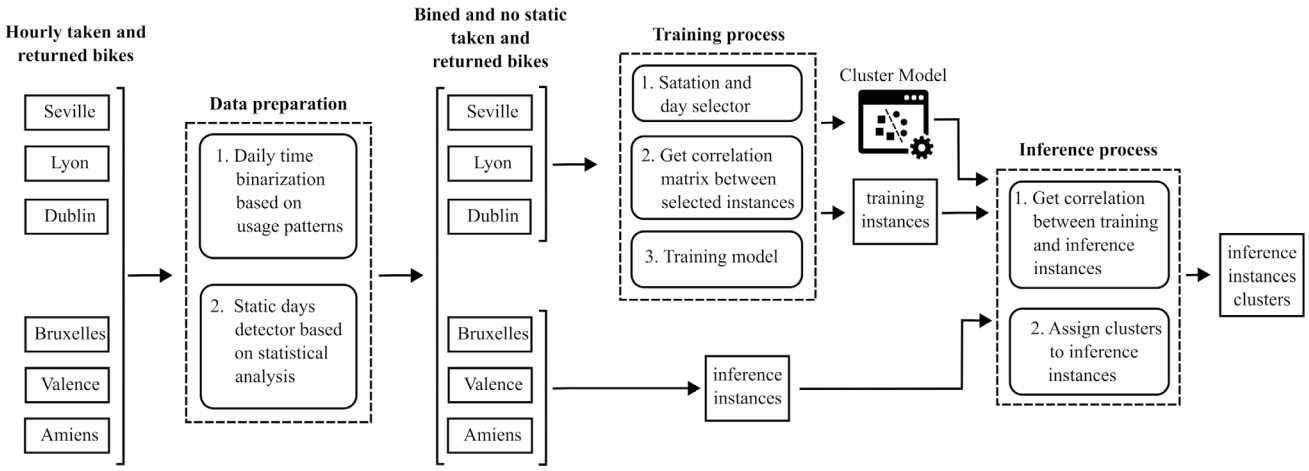


Fig. 13 Model development workflow. All implemented steps are summarized in this figure together with input data flow and resulting items during the process

training, the unused instances from training systems could also be considered for the evaluation process.

Thus, if the identified clusters truly represent a generalized pattern, the daily trends of taken and returned bikes for classified instances in each cluster should be quite similar across systems. This condition has been fulfilled by the developed model, as the discovered behaviors appear not only in systems seen during training but also in those solely used for testing (see Fig. 14).

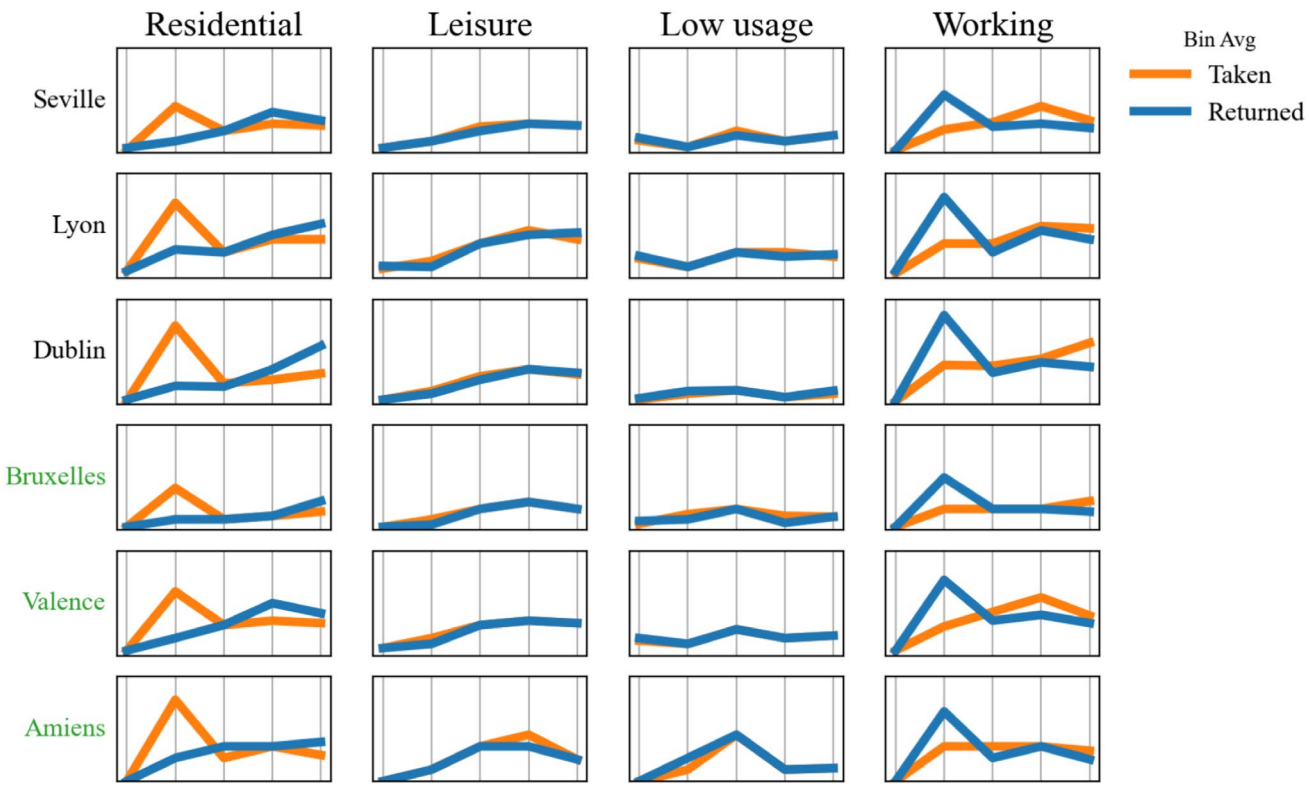


Fig. 14 Daily taken and returned bikes curve per cluster through systems. Full test system's labels are colored in green. Discovered behaviors keep similar despite the system has been used in training process or not

6.3.1 The geospatial dimension

Once extracted patterns have been guaranteed to make sense outside the training set, station location is another important domain feature to consider. Hence, verification methodology developed supports on GIS data to enable location validation understanding and automatization. For implementing that, The Urban Atlas European Project GIS Data (https://land.copernicus.eu/en/products/urban-atlas) has been used, as it provided open access to relevant interesting BSS domain GIS layers such as working areas or leisure facilities.

Several conditions the developed model needs to accomplish has been defined to ensure extracted knowledge to make sense. The conditions have been checked through a map visualization of most classified cluster to each station in a specific period. Accomplished conditions are the following:

- Stations primarily clustered as "Working" must be near working areas or in the city center. Additionally, they should generally not be clustered as "Working" on weekends (Fig. 15).
- Instances assigned to the "Leisure" cluster must be located near leisure areas or in the city center. Moreover, a higher number of instances in this cluster is expected to appear on weekends (Fig. 16).
- Residential areas are expected to be around the city center but not in immediate proximity (Fig. 17).
- Stations with low usage or considered static must increase as the distance from the city center grows (Fig. 17).

Therefore, it has been ensured that the trained clustering model not only is able to apply extracted patterns in new systems which have been never seen, but also these behaviors have been ensured to be right according to domain-specific factors.

7 Results

Results from BSS-based data analysis are usually interpreted in urban socio-geo-economic terms [31, 32]. The analysis can lead to useful conclusions for both institutional stakeholders and the company providing the service, particularly for those responsible for its maintenance and expansion.

In this work, clustering BSS has been approached from a different perspective compared to the studies carried out in the cited works. The results can be divided into two main points. First, the extracted knowledge and capabilities of the developed clustering solution are presented. Second, a comparison between previous clustering methodologies studied by academia becomes imperative to ensure that existing limitations have been overcome.

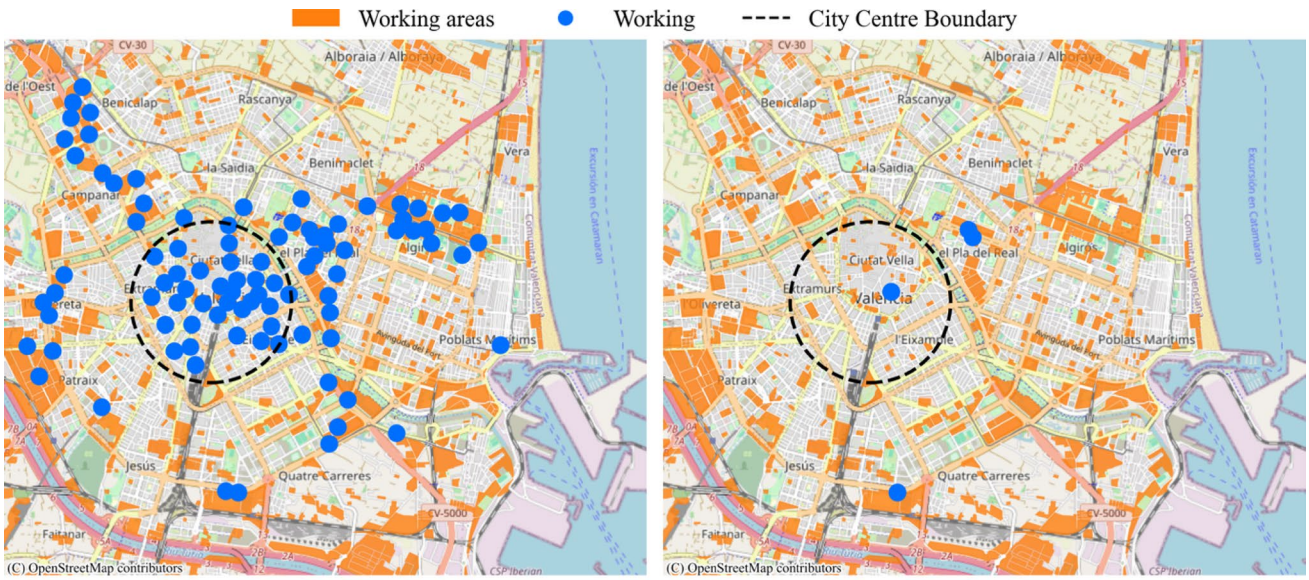


Fig. 15 Valence stations most classified as working split between weekdays or weekends. City centre is represented inside the slashed black ring for improving validation understanding

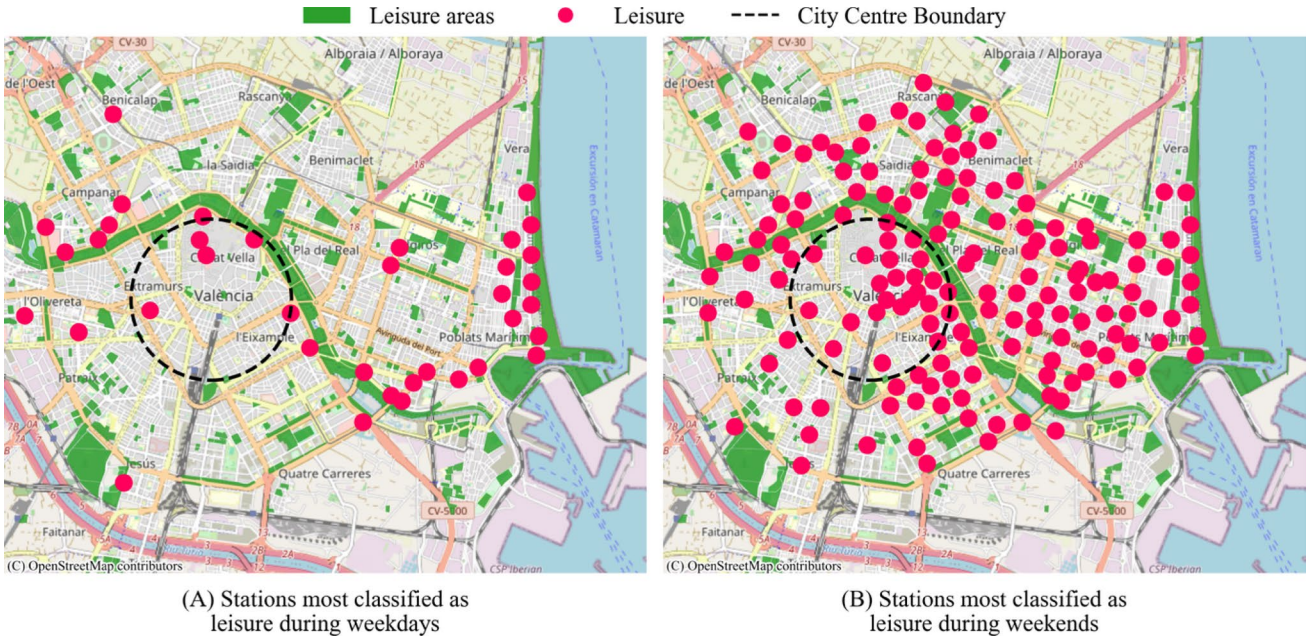


Fig. 16 Valence stations most classified as leisure split between weekdays or weekends. City centre is represented inside the slashed black ring for improving validation understanding

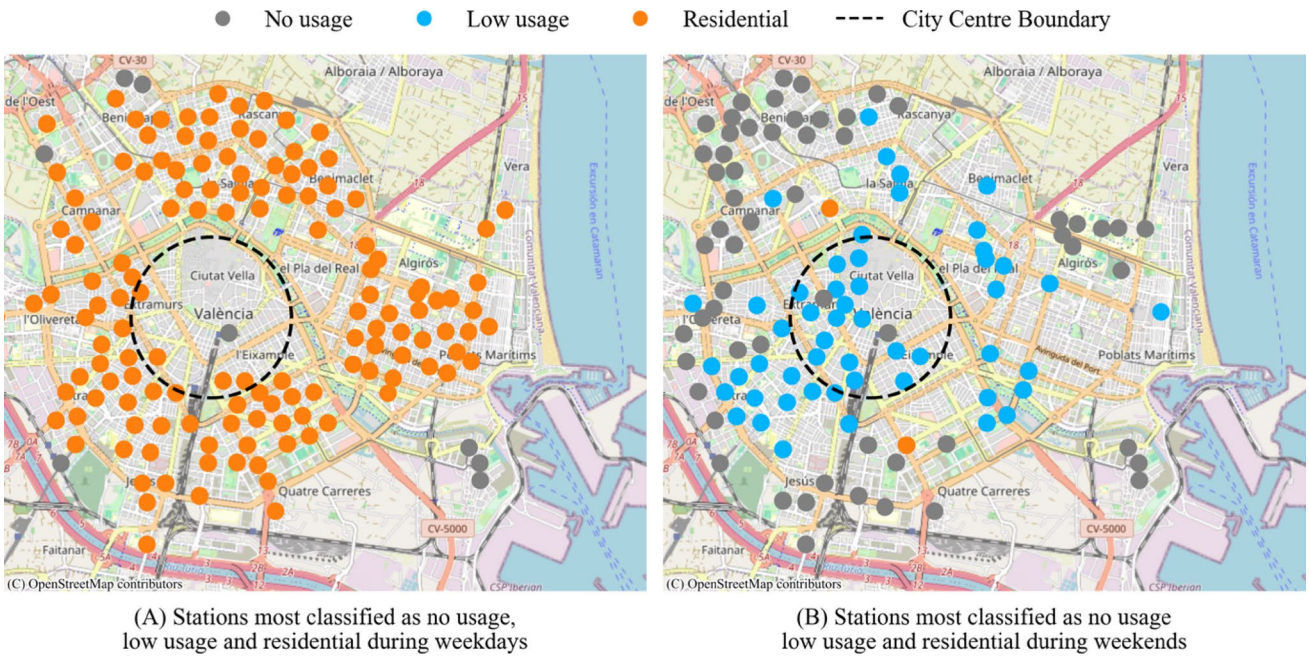


Fig. 17 Valence stations most classified as no usage, low usage and residential split between weekdays or weekends. City centre is represented inside the slashed black ring for improving validation understanding

7.1 Developed clustering model capabilities

Applying extracted knowledge from systems used in training to never-before-seen ones has been postulated as one of the main limitations in the literature. Consequently, this has also been considered one of the main objectives of this work.

In Fig. 18, the most frequent cluster for stations in three systems during the whole period is shown. The model has entailed coherent behaviors. Although no distinction was

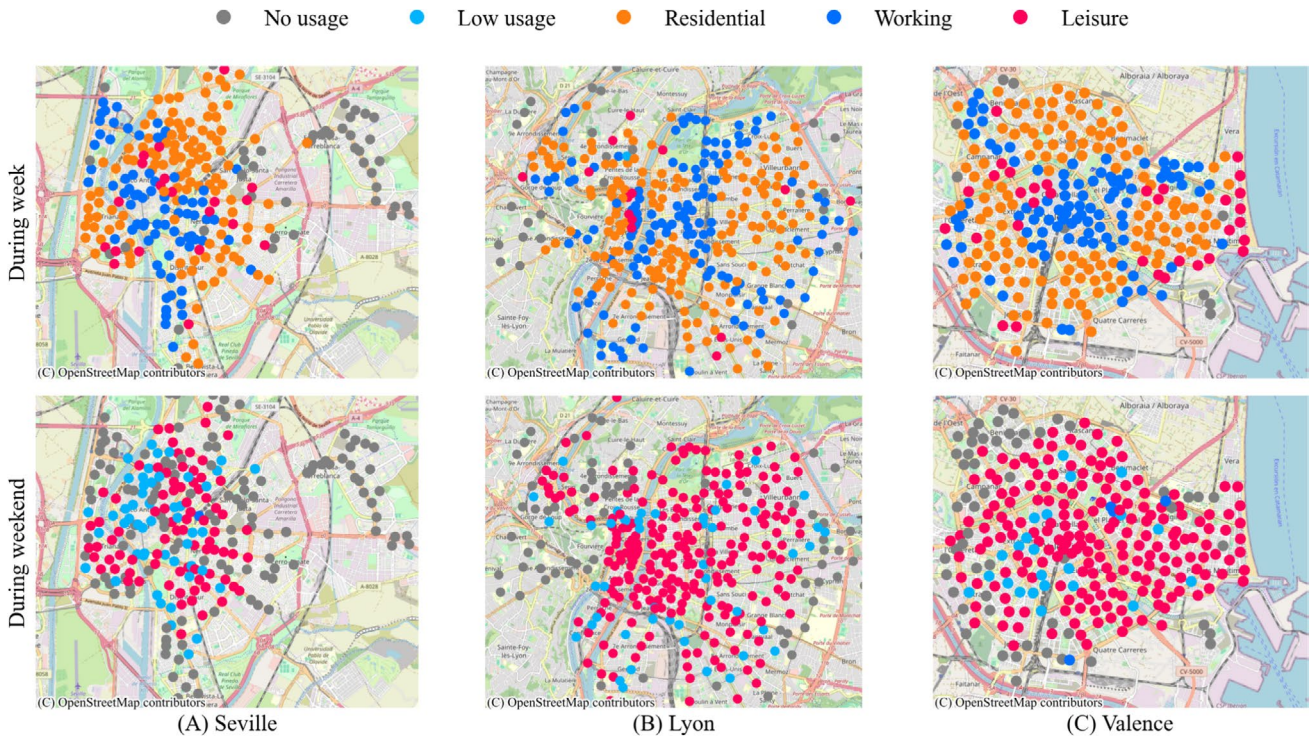


Fig. 18 Most classified cluster per each station on Seville, Lyon and Valence depending on working or weekend

implemented in the training process, different patterns have been found depending on work or weekend days. Additionally, stations near the periphery have been classified as "residential", whereas those located in the city center have been considered as "working" or "leisure".

It is worth noting that, despite finding "working" stations far from the city center, they are always located near offices or working areas, according to GIS data. "Leisure" instances have also been classified correctly, as their occurrence increases on weekends compared to working days. While "leisure" stations are typically found near the city center in Seville and Lyon, in Valence, they can be seen near the beach area even on working days.

We would like to emphasize that all of this has been classified using just a single trained model, without splitting between working and weekend days.

Concerning the day binarization methodology developed, despite the model having been trained with data from only 120 stations across 3 BSSs (Seville, Lyon, and Dublin), it can apply extracted knowledge even to new upcoming systems such as Valence. That is, the developed semantic unification of the day together with the use of Person Correlation as cluster input has enabled knowledge extraction between systems, leading to a multi-system analysis by training just a single clustering model. This fact has been presented in more detail in Sect. 6.3.

7.2 Comparison with the baseline clustering models

Comparison between the baseline clustering models and the one proposed in this work becomes necessary for several reasons. On the one side, developed model disadvantages against existing works need to be pointed. On the other side, demonstrating that some existing limitations in the literature have been overcome by the developed work turns imperative.

It is true that Correlation is presented as one of the keys to a multi-system clustering approach for Bike Sharing Systems by this work. Nevertheless, computational efficiency of this distance metric is supposed to be worse than other used in literature such as K-Shape or K-Means with DTW.

Consequently, it has been compared execution time of both implemented and baseline models at inference process, which may be considered a real usage case for any clustering developed model. First, Seville instances have been selected to be run as it is considered a middle-large size (260 stations). Next several time windows (1, 7, 30 and 90 days) have been inferred for all the stations in the system.

Run time for each methodology has been measured and presented in Table 3. As expected, K-Means with DTW reflects the best computational efficiency when production time. Using daily curves statistic together with a less time-consuming distance metrics drive to this result. In relation

Table 3 Computational required time for inferring developed and baseline models over several time windows

Method	1 day (260 instances)	7 days (1.820 instances)	30 days (7.800 instances)	90 days (23.400 instances)
K-Means correlation	4.62 s	6.29 s	15.16 s	47.57 s
K-Shpae (baseline)	0.16 s	1.01 s	4.32 s	12.77 s
K-Means DTW (baseline)	0.09 s	0.29 s	0.33 s	0.43 s

to correlation approaches, run time for the developed model in this work seems to be the highest.

Time consumption could be the main disadvantage between the proposed solution and baselines. Nevertheless, according to measured times in Table 3, a real production-time execution might not be compromised. Notice that full system data for a week or a month could be processed in few seconds. Nevertheless, knowledge extraction and its application is required to be compared between baseline models to grant this computational efficiency weakness to be compensated by better results.

The reuse of learned representations, classification, and BSS behavior understanding in new BSSs is a distinctive feature of the model. Cluster occurrence has been compared in Fig. 19 to measure knowledge application capabilities between the implemented and baseline models.

Cluster occurrence differs between systems for both baseline models. Notice that stations classified as "Working" by K-Means DTW seem to be quite low even in systems used for training. On the contrary, a more consistent occurrence pattern among systems has been found in the developed and K-Shape models' output, even in those systems not used in the training process. Notice that now a similar representation can be found for the "Working" cluster in almost all systems. Hence, K-Means DTW knowledge extraction capabilities could be considered worse than correlation approaches (notice that developed model solution has been validated against GIS data).

On the contrary, K-Shape approach seems similar to K-Means Correlation. The main difference could be the higher occurrence of "Working" instance in K-Shape which will be discussed next in this section.

A deeper analysis of the "Working" cluster can be interesting to ensure the improvements achieved by the developed approach. Stations considered as "Working" by both the baseline and the developed clustering model in this work are compared in Fig. 20.

Fig. 19 Cluster occurrence per system and stations comparison between the developed model and baselines. System not used in training process has been marked in green

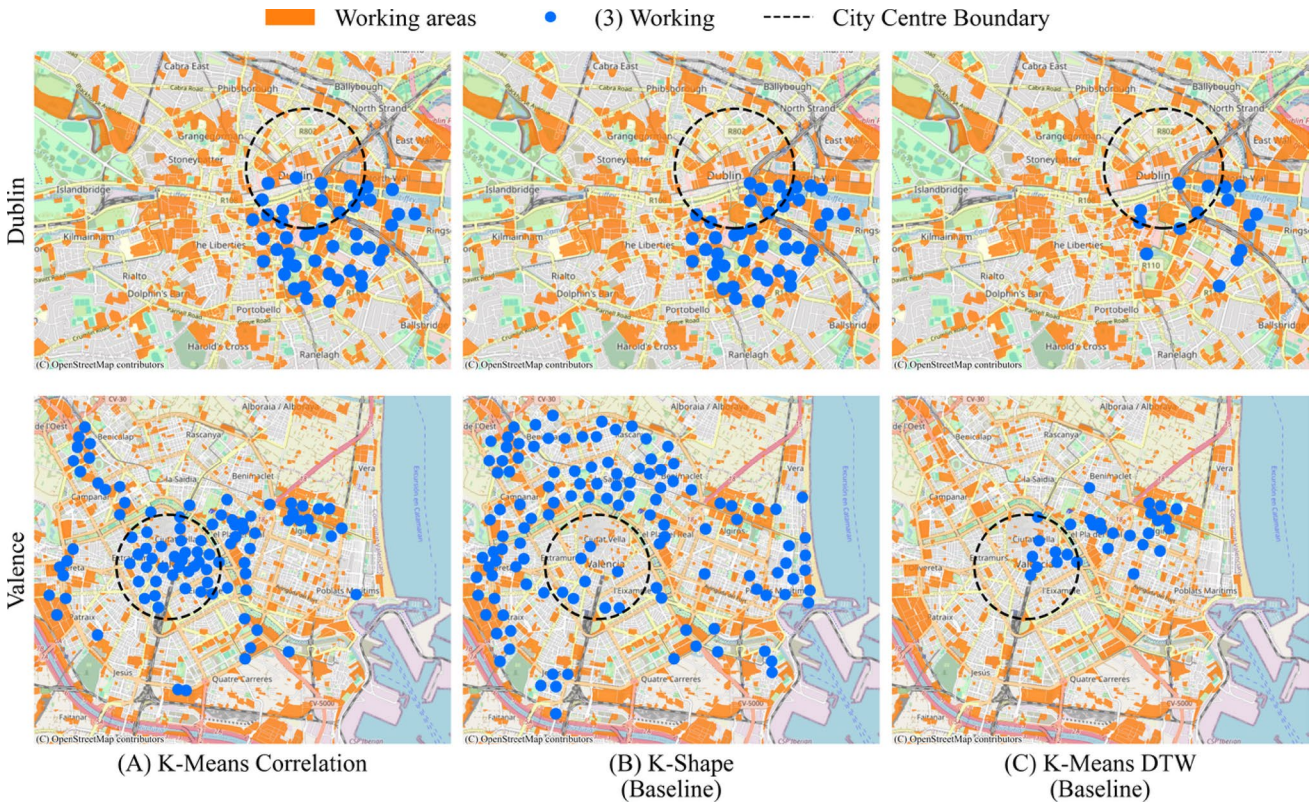


Fig. 20 "Working" stations found by the developed model (left column) and the baselines (middle and right columns). The Dublin system (used in the training process) is shown at the top. The Valence system (not used in the training) is shown at the bottom. City centre is represented inside the slashed black ring. Notice that prediction in a never-before-seen system has been enabled by using Correlation instead of Euclidean or DTW distances

The baseline models' weakness with respect to reusing learned knowledge in new BSSs has been highlighted. It is true that common patterns have been discovered in Dublin (used in training). Nevertheless, different results are found when applied developed techniques to Valence (a never seen system). On the one hand, despite few coherent working stations have been classified by K-Means DTW approach, found results are quite far from reality, where more "Working" stations are supposed to be found. On the other hand, an exceed of "Working" stations seems to be predicted by K-Shape. It is true that stations located near working GIS zones have been correctly classified. However, an important number of "Working" stations have been located in large residential areas or near the beach.

On the contrary, not only working stations in the city center of Valence or in the periphery near GIS working areas have been detected by the implemented clustering model in this work, but also well defined residential or leisure areas are not blended (see Fig. 20). Therefore, previous methodological limitations have been overcome, and a multi-system clustering approach has been better achieved by the developed study.

Dealing with seasonality has been identified as another relevant limitation of previous studies found in the literature. The inability of the baseline models to detect pattern changes during the analyzed period is shown in Fig. 21. Using only a single statistic for the analyzed period is considered the main cause of this limitation in the K-Means DTW technique, which also misclassified this station to "Residential" instead of "Working". Even K-Shape, which also targets to daily instances as the proposed work seems to not properly detect seasonality. However, these changes can be handled by the model implemented in this work (see also Fig. 21).

Finally, detecting special days such as festivities or events, which might change station behavior, seems to be crucial. The capabilities of the designed and baseline models for this task are presented in Fig. 22. On the one hand, the "Working" pattern has been discovered by both developed in this work and baseline models. On the other hand, the National Holiday on October 12th has only been detected by the K-Means Correlation clustering designed in this work.

Moreover, despite other distinct patterns on October 24th and 25th seem to be handled by both correlation approaches,

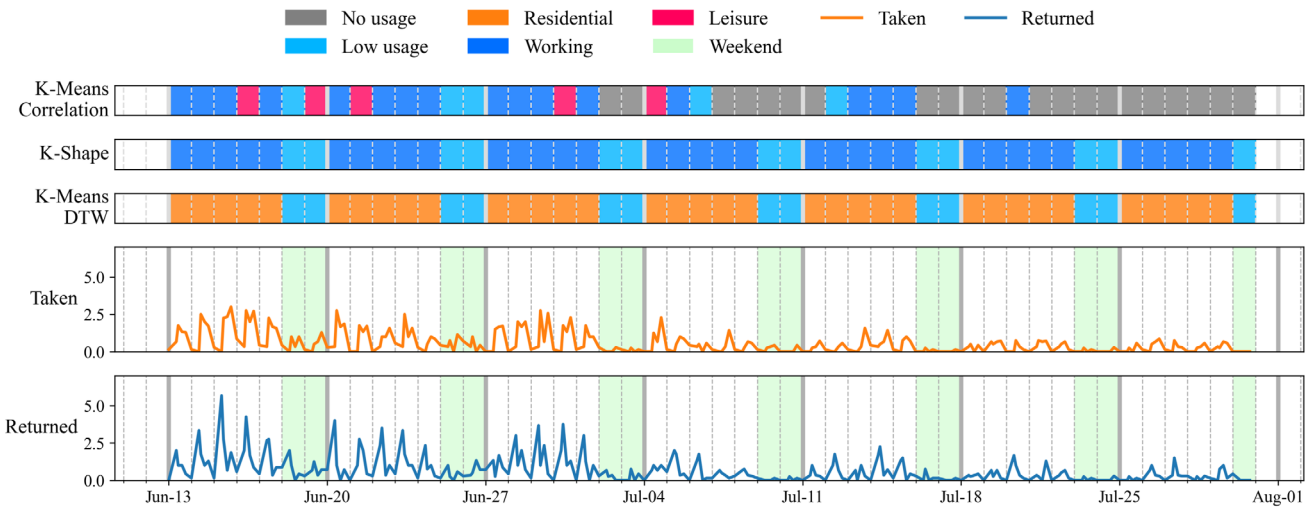


Fig. 21 Clustering results found by Correlation clustering (developed in this work) and Euclidean and DTW approaches (baselines) are presented for station near university in Seville. Last days of the academic year continued by first holiday period taken and returned bikes are shown. Despite activity considerably decreases by July 4th, baseline clustering models keeps similar than before. On the contrary, this pattern change has been detected by the developed Correlation approach in this work

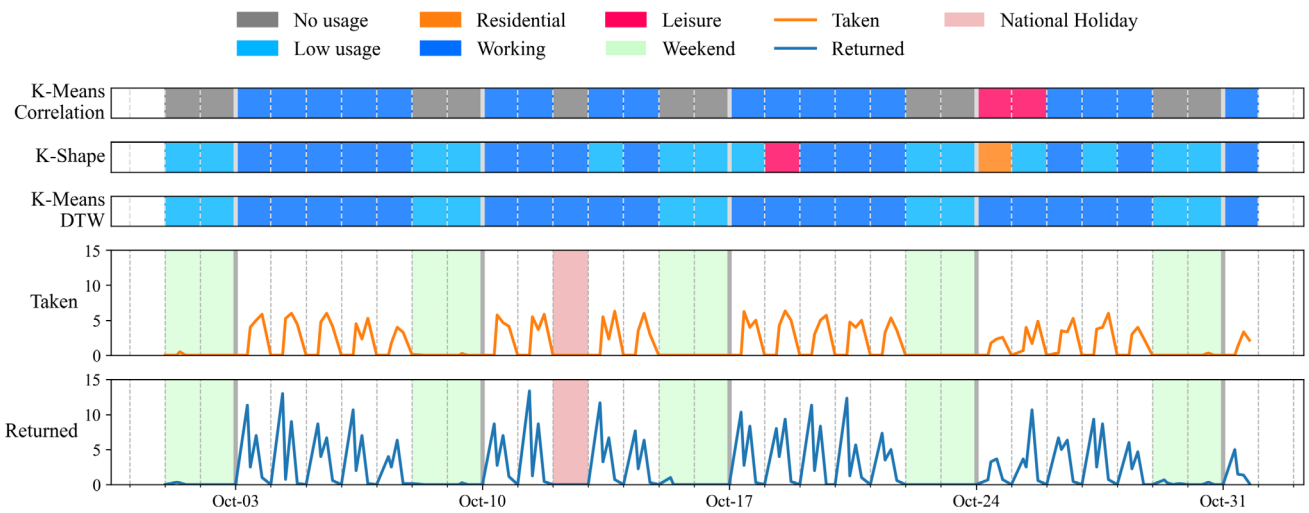


Fig. 22 Clustering results found by Correlation clustering (developed in this work) and Euclidean and DTW approaches (baselines) are presented for a period with special days (marked in red). October 12th National Holiday in Spain has only been discovered by Correlation approach

K-Shape seems to misunderstood those days behaviours (it not seems to be "Residential" or "Low usage") rather than presented solution, which has correctly classified it as "Leisure". Consequently, the capability of the implemented solution to deal with special situations as well as its improvement over baselines has been demonstrated.

8 Conclusions and future work

The proposed multi-system clustering approach, based on daily usage patterns, might offer an analytical tool for understanding and improving Bike Sharing Systems.

While there is room for further refinement and extension, the obtained results show the potential of this approach which might enhance the study and optimization of BSS as a sustainable model of urban transportation.

Training with several system data has been addressed by the daily binarization proposed in this paper, which translates a numerical concept (hourly) into a semantic one (such as "going to work"). This transformation could provide several advantages.

A better generalization could be reached as the model is trained on data from multiple cities. Additionally, as daily-time routines have been unified across cities, its application in never-seen systems might have been enabled.

Some limitations existing in previous clustering approaches might have been overcome by the implemented methodology. Developed model seems to effectively differentiate between working days and weekend behavior, capture seasonal trends, and even detect special events or holidays. This might represent an improvement over baseline methods. Focusing on classifying station daily usage patterns instead of station statistics during a given period and the use of Pearson Correlation as distance metric, seems to be one of the keys to this supposed upgrade. As demonstrated by the results obtained from the six European cities studied in this paper.

The use of Pearson Correlation as distance metric needs to be considered. On the one hand, by measuring similarity based on the shape of the usage curves rather than their absolute magnitude, common patterns across stations and BSSs of different sizes could be identified by the model. On the other hand, calculating correlation for all new instances is more time-consuming than other approaches, resulting into the main disadvantage of the proposed work.

Applying extracted knowledge to new instance without the necessity of retraining the model is allowed by the proposed inference process. It is true that this methodology reduces run time and enables using the implemented model in production time (running one week of a whole system only requires few seconds). Nevertheless, computational efficiency also remains worse than baseline methods.

All the implemented solution capabilities as well as its limitations and comparison between baseline models previously defined in the literature have been highlighted in Sect. 7.

8.1 Future work

Our model might serve as a confident base for several fields and applications related to BSS. First of all, integrating developed clustering results with additional data sources, such as weather conditions, public transport schedules, or socio-demographic information could enrich the analysis of BSS usage patterns.

With regards to Data Science applied to BSS data, an AI-driven Exploratory Data Analysis might result when combining the developed model with classical statistical methodologies. This could serve as a starting point for more practical methodologies.

First, outlier detection based on the daily clustering classification could be explored. Identifying the cause of misclassified stations in relation to their neighbours or even not expected daily behaviours could be considered interesting.

Next, merging cluster results with forecasting techniques related to BSS (for example, demand forecasting) might be useful. Combining station daily classification with other relevant features such as lagged demand or weather conditions might improve forecasting accuracy.

Reducing consumption time of the implemented methodology will be another important research line to explore. Despite the use of the model in BSS has been arrived, an efficiency improvement could enable the application of the proposed model in more demanding fields.

Finally, bringing the implemented model into production and running it daily for the nearly 30 systems recorded in the data acquisition process would be the next challenge for the authors of this work.

Acknowledgements We would like to thank the reviewers and the editor for their careful reading and insightful suggestions, which significantly improved the quality of this paper.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by F.M.S, G.A.A.C., and J.B.D.. The first draft of the manuscript was written by F.M.S., and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Funding for open access publishing: Universidad de Sevilla/CBUA. Francisco Márquez-Saldaña, Gonzalo A. Aranda-Corral and Joaquín Borrego-Díaz, received Grant PID2023-147198NB-I00 funded by MICI-U/AEI/10.13039/501100011033 (Agencia Estatal de Investigación) and by FEDER, UE.

Data availability Not applicable.

Code availability Not applicable.

Materials availability Not applicable.

Declarations

Conflict of interest Not applicable.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Xu X, Zuo W (2024) Does bike-sharing reduce traffic congestion? Evidence from three mega-cities in China. PLoS ONE 19(8):0306317. https://doi.org/10.1371/journal.pone.0306317
2. Geneletti D, Cortinovis C, Zardo L, Adem Esmail B (2019) Planning for ecosystem services in cities. Springer Briefs in Environmental Science. Springer, Cham. https://doi.org/10.1007/978-3-030-20024-4
3. O'Brien O, Cheshire J, Batty M (2014) Mining bicycle sharing data for generating insights into sustainable transport systems. J Transp Geogr 34:262–273. https://doi.org/10.1016/j.jtrangeo.2013.06.007
4. Reggiani G, Oijen T, Hamedmoghadam H, Daamen W, Vu HL, Hoogendoorn S (2022) Understanding bikeability: a methodology to assess urban networks. Transportation 49:897–925. https://doi.org/10.1007/s11116-021-10198-0
5. Froehlich J, Neumann J, Oliver N (2009) Sensing and predicting the pulse of the city through shared bicycling. In: Proceedings of the 21st international joint conference on artificial intelligence. IJCAI'09. Morgan Kaufmann Publishers Inc., San Francisco, pp 1420–1426
6. Canzler W, Knie A (2023) The future of mobility: Winners and losers and new options in the public space. WZB Discussion Paper SP III 2023-601, Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin
7. Ricci M (2015) Bike sharing: a review of evidence on impacts and processes of implementation and operation. Res Transp Bus Manag 15:28–38. https://doi.org/10.1016/j.rtbm.2015.03.003. (**Managing the Business of Cycling**)
8. Pearson L, Dipnall J, Gabbe B, Braaf S, White S, Backhouse M, Beck B (2022) The potential for bike riding across entire cities: quantifying spatial variation in interest in bike riding. J Transp Health 24:101290. https://doi.org/10.1016/j.jth.2021.101290
9. Natera Orozco LG, Battiston F, Iñiguez G, Szell M (2020) Data-driven strategies for optimal bicycle network growth. R Soc Open Sci 7:201130. https://doi.org/10.1098/rsos.201130
10. Szell M, Mimar S, Perlman T, Ghoshal G, Sinatra R (2022) Growing urban bicycle networks. Sci Rep 12(1):6765. https://doi.org/10.1038/s41598-022-10783-y
11. Liu S, Shen Z-JM, Ji X (2022) Urban bike lane planning with bike trajectories: models, algorithms, and a real-world case study. Manuf Serv Oper Manag 24(5):2500–2515. https://doi.org/10.1287/msom.2021.1023
12. Shimizu S, Akai K, Nishino N (2014) Modeling and multi-agent simulation of bicycle sharing. In: Mochimaru M, Ueda K, Takenaka T (eds) Serviceology for services. Springer, Tokyo, pp 39–46
13. Cipriano M, Colomba L, Garza P (2021) A data-driven based dynamic rebalancing methodology for bike sharing systems. Appl Sci. https://doi.org/10.3390/app11156967
14. Wang X, Sun H, Zhang S, Lv Y, Li T (2022) Bike sharing rebalancing problem with variable demand. Phys A 591:126766
15. Singla A, Santoni M, Bartók G, Mukerji P, Meenen M, Krause A (2015) Incentivizing users for balancing bike sharing systems. In: Proceedings of the AAAI conference on artificial intelligence, vol 29(1). https://doi.org/10.1609/aaai.v29i1.9251
16. National Academies of Sciences, Engineering, and Medicine: Micromobility Policies, Permits, and Practices (2022). The National Academies Press, Washington, DC. https://doi.org/10.17226/26815
17. Borrego-Díaz J, Chávez-González AM, Martín-Pérez MA, Zamora-Aguilera JA (2012) Semantic geodemography and urban interoperability. In: MTSR 2012: metadata and semantics research. communications in computer and information science, vol 343. Springer, pp 1–12
18. Rennie N, Cleophas C, Sykulski AM et al (2023) Analysing and visualising bike-sharing demand with outliers. Discov Data 1:1. https://doi.org/10.1007/s44248-023-00001-z
19. Feng Y, Affonso RC, Zolghadri M (2017) Analysis of bike sharing system by clustering: the vélib'case. IFac-Papersonline 50(1):12422–12427
20. Borgnat P, Abry P, Flandrin P, Robardet C, Rouquier J-B, Fleury E (2011) Shared bicycles in a city: a signal processing and data analysis perspective. Adv Complex Syst 14(03):415–438
21. Vogel P, Greiser T, Mattfeld DC (2011) Understanding bike-sharing systems using data mining: exploring activity patterns. Procedia Soc Behav Sci 20:514–523
22. Etienne C, Latifa O (2014) Model-based count series clustering for bike sharing system usage mining: a case study with the vélib'system of Paris. ACM Trans Intell Syst Technol (TIST) 5(3):1–21
23. Rennie N, Cleophas C, Sykulski AM, Dost F (2023) Analysing and visualising bike-sharing demand with outliers. Discover Data 1(1):1
24. Feng S, Chen H, Du C, Li J, Jing N (2018) A hierarchical demand prediction method with station clustering for bike sharing system. In: 2018 IEEE 3rd international conference on data science in cyberspace (DSC). IEEE, pp 829–836
25. Dai P, Song C, Lin H, Jia P, Xu Z (2018) Cluster-based destination prediction in bike sharing system. In: Proceedings of the 2018 artificial intelligence and cloud computing conference, pp 1–8
26. Chen L, Zhang D, Wang L, Yang D, Ma X, Li S, Wu Z, Pan G, Nguyen T-M-T, Jakubowicz J (2016) Dynamic cluster-based over-demand prediction in bike sharing systems. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing, pp 841–852
27. Zhu H, Shou T, Guo R, Jiang Z, Wang Z, Wang Z, Yu Z, Zhang W, Wang C, Chen L (2022) Redpacketbike: a graph-based demand modeling and crowd-driven station rebalancing framework for bike sharing systems. IEEE Trans Mob Comput 22(7):4236–4252
28. Li D, Zhao Y, Li Y (2019) Time-series representation and clustering approaches for sharing bike usage mining. IEEE Access 7:177856–177863. https://doi.org/10.1109/ACCESS.2019.2958378
29. Saldaña FJM, Aranda-Corral G, Borrego-Díaz J (2024) Bike sharing systems data interoperability by a unified station status concept and big data solutions. J Traffic Transp Eng (to appear)
30. Médard de Chardon C, Caruso G (2015) Estimating bike-share trips using station level data. Transp Res Part B Methodol 78:260–279. https://doi.org/10.1016/j.trb.2015.05.003
31. Jiménez P, Nogal M (2021) Analysis of real experiences using different sized bike sharing schemes in Irish cities. Transp Res Procedia 58:37–44. https://doi.org/10.1016/j.trpro.2021.11.006
32. Chalhoub Dourado G (2018) Bike-sharing system design - guidelines on conceiving and implementing a BSS as a public transport with a monocentric heterogeneous demand. Master thesis, Universitat Politècnica de Catalunya. https://upcommons.upc.edu/handle/2117/117643

33. Ashqar HI, Elhenawy M, Rakha HA, House L (2022) Quality of service measure for bike sharing systems. IEEE Trans Intell Transp Syst 23(9):15841–15849
34. Li Y, Zheng Y, Zhang H, Chen L (2015) Traffic prediction in a bike-sharing system. In: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, pp 1–10
35. Ashqar HI, Elhenawy M, Rakha HA (2019) Modeling bike counts in a bike-sharing system considering the effect of weather conditions. Case Studies Transp Policy 7(2):261–268
36. Marquez-Saldaña FJ, Aranda-Corral GA, Borrego-Díaz J (2022) Enabling knowledge extraction on bike sharing systems throughout open data. HCI in mobility. Transport, and automotive systems: 4th international conference, MobiTAS 2022, Held as Part of the 24th HCI international conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings. Springer, Berlin, pp 570–585
37. Foss AH, Markatou M, Ray B (2019) Distance metrics and clustering methods for mixed-type data. Int Stat Rev 87(1):80–109
38. Ghazal TM (2021) Performances of k-means clustering algorithm with different distance metrics. Intell Autom Soft Comput 30(2):735–742
39. Berthold MR, Höppner F (2016) On clustering time series using euclidean distance and Pearson correlation. arXiv preprint arXiv:1601.02213
40. Molchanov V, Linsen L (2018) Overcoming the curse of dimensionality when clustering multivariate volume data. In: VISIGRAPP (3: IVAPP), pp 29–39
41. Yang H, Zhang Y, Zhong L, Zhang X, Ling Z (2020) Exploring spatial variation of bike sharing trip production and attraction: a study based on Chicago's divvy system. Appl Geogr 115:102130
42. Hajela P, Lin C-Y (1992) Genetic search strategies in multicriterion optimal design. Struct Optim 4:99–107
43. Tipton E (2013) Stratified sampling using cluster analysis: a sample selection strategy for improved generalizations from experiments. Eval Rev 37(2):109–139
44. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
45. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 2:224–227
46. Tomašev N, Radovanović M (2016) Clustering evaluation in high-dimensional data. In: Unsupervised learning algorithms. Springer, pp 71–107
47. Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27
48. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process 26(1):43–49. https://doi.org/10.1109/TASSP.1978.1163055
49. Paparrizos J, Gravano L (2015) k-shape: efficient and accurate clustering of time series. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, pp 1855–1870
50. Crase S, Thennadil SN (2022) An analysis framework for clustering algorithm selection with applications to spectroscopy. PLoS ONE 17(3):0266369
51. Sun ED, Ma R, Zou J (2023) Dynamic visualization of high-dimensional data. Nat Comput Sci 3(1):86–100
52. Thrun MC, Ultsch A (2021) Using projection-based clustering to find distance-and density-based clusters in high-dimensional data. J Classif 38(2):280–312