Interaction Design and Architecture(s) Journal - IxD&A, N.65, 2025, pp. 90 - 114
DOI: 10.55612/s-5002-065-003

AI and the narrative of the everyday life: machine learning applied to the social mapping of rural music festivals.

Enrique Villamuelas García[1 2], Eva Hurtado Torán[2], Eduardo Roig Segovia [1],

[1] Escuela Técnica Superior de Arquitectura de Madrid, Universidad Politécnica de Madrid
[2] Universidad de Diseño, Innovación y Tecnología (UDIT)

Abstract. Given the growing ubiquity of AI and the consequent valorization of data in territorial representation, it becomes crucial to analyze how multimodal algorithms interpret spatial narratives and community dynamics. This research studies the effects of algorithmic automation on social mappings. Using "Balboa Observa" project, a collaborative web mapping initiative that documents Observatorio Music Festival in Spain, the study explores the interaction between collective cartography, ethnographic analysis, and AI-driven data processing. Unlike conventional AI practices, the prioritizes analyses that avoid imposing structured categories on everyday narratives, allowing inclusion of sensitive information for deeper festival impact understanding. Through multimodal algorithms like CLIP and ImageBind analyzing images, texts, and audio recordings, the research reveals how AI-generated spatial configurations differ from human interpretations and identifies biases from training data and algorithmic processes. The study highlights community participation and data ownership importance to mitigate biases and advocates for transparent, adaptable AI tools for social mapping.

Keywords: Social Mapping, Critical GIS, Socio-spatial Theory, Machine Learning, Multimodal Algorithms, Artificial Intelligence

1 Introduction

The research studies the effects of algorithmic automation on social mapping. Through the contrast of territorial analyses carried out by humans with artificial intelligence (AI) analyses, it aims to observe what distortions AI introduces in the representation of the territory. The proposal includes the use of analysis methods without fixed categories that reveal to local actors the orders and biases of digital representation. The concept of bias is approached not as an error to be corrected, but as an inherent and inevitable characteristic of AI systems, resulting from the specific perspectives and values contained in their databases. From this approach, the goal is to make the functioning of algorithms transparent and balance their role with that of humans, preventing these systems from obscuring the active participation of the local community in projecting their territorial reality.

In a first approach, a brief synthesis of the state of the art of the most prominent theories and references is carried out, to then apply the case study methodology. As a case study, we work in Balboa, a village in El Bierzo belonging to the popularly known as "España vaciada" [emptied Spain], an expression that designates rural areas of Spain that suffer depopulation [1]. Tourism has emerged as a transformative factor in rural environments that have transitioned from an agricultural-livestock model to one based on tourist activities. Balboa exemplifies this trend: although it has revitalized its livestock sector, tourism has become one of its main sources of income, especially in the form of music festivals such as Reggaeboa, Vibra Balboa and the Observatorio Festival.

The Observatorio Festival, beyond music, seeks to involve its attendees with the community through workshops and cultural projects that positively impact Balboa and its surroundings. One of these projects is "Balboa Observa", a web mapping initiative that annually documents the surroundings of Balboa during the four days of the Festival, exploring the relationship between collective cartography, ethnographic analysis, and AI databases.

In contrast to the extractivist character that conventionally characterizes data collection for AI, this research adopts an alternative approach inspired by ethnographic techniques that are more respectful of the reality to be represented [2]. The territory is collaboratively mapped from the point of view of its agents, giving them control so that it includes sensitive information such as images with recognizable faces, voices and personal stories, crucial information for a deeper understanding of the everyday narratives of the Festival.

The processing of data through AI algorithms presents considerable complexity, mainly due to its "black box" nature. This opacity makes it difficult for both the individuals who record the data and the communities subject to recording to develop an understanding of the process and question their results. The present research focuses on identifying methods that facilitate an early and accurate understanding of the effects of these algorithms to offer communities greater participation and control in this process.

2 Objectivity and Bias in Google Maps

Google Maps, along with services such as Apple Maps and Bing Maps, has achieved its popularity among digital mapping services thanks to massive data collection, which creates an apparently objective representation of the world, but designed to meet Google's commercial objectives. The company combines satellite imagery and topographic data for aerial views, while for street-level views it uses various vehicles and devices equipped with 360° cameras and lidar scanners. These include cars, tricycles, backpacks, and boats, which capture geolocated images and three-dimensional data [3, 4].

Government entities and organizations provide official information on buildings, roads, and public transport [5]. Simultaneously, real-time data on patterns of space usage and traffic are collected through the mobile devices of its users [6]. As a last resort to keep their information updated, they make use of AI-powered telephone bots with synthesized voices which, simulating human calls, obtain additional data from businesses and organizations [7].

Google Local Guides, initially known as Panoramio, represents the only space apparently ceded to local actors in the Google Maps ecosystem. Acquired in 2007, Panoramio was originally a social network for sharing geolocated photographs [8]. After its acquisition and subsequent rebranding, the service has evolved towards a more map monetization-oriented approach. Although the platform promotes ideals of community participation, local empowerment, and generation of social impact, in practice, Google Local Guides focuses on collecting reviews and photographs of commercial establishments and tourist attractions [9].

Based on this information, Google Maps performs a more exhaustive processing of geospatial information and user-generated content known as knowledge extraction or "knowledge mining". This process includes the use of computer vision algorithms that, for example, analyze Street View images to extract data such as business names and hours [10], or study social patterns of entire neighborhoods by observing parked vehicles [11] or satellite images [12]. These algorithms, which employ neural networks, have the ability to work with enormous amounts of unstructured data, including images, text, or videos. The learning of these algorithms materializes in an abstract and numerical representation of the data, which crossed with geographic data allows for the automatic execution of complex tasks, such as, for example, analyzing visuals to detect unregistered routes [13] or guiding users through a territory by seeking routes that do not correspond to the most optimal path [14].

The apparent objectivity attributed to satellite imagery and AI processing devalues human contributions and contrasts with the interpretative nature of traditional cartography [15, 16]. This machine-based approach that collects data in an opaque manner, often without the knowledge or consent of the affected individuals, not only raises ethical questions but also excludes any alternative reality to that perceived by the machine [17]. As a result, the representation of geographic space in digital cartography is constrained by commercial interests and algorithmic biases, creating a limited image that systematically over-represents and invisibilizes minority or local perspectives [17].

3 The Possibility of an Alternative Model

The democratization and improvement of web technologies, the creation of open databases, and research into new algorithmic analyses have driven the development of digital cartography outside the business sphere [18]. Works such as those by the urban planning studio 300,000 Km/s show how these new tools allow for the digital visualization of large territorial databases, revealing dimensions previously imperceptible to the human eye, such as the spatial distribution of pollution [19], the state of buildings [20], or urban uses and events [21]. The interactivity and near real-time updating of these visualizations facilitate the study of events that unfold in both space and time, such as music festivals [22] or demonstrations in public spaces [23]. As a consequence of the acceleration of information, early detection of events requiring immediate intervention becomes possible, as in the monitoring of forest fires [24]. However, these techniques not only offer new visualization methods but also present new challenges derived from their technical complexity and the use of large databases.

In the public sector, the use of this information for automated decision-making presents both challenges and opportunities. AI governance poses a dilemma for governments: on the one hand, they must protect citizens from potential algorithmic harm; on the other, they are tempted to increase their efficiency through the use of algorithms [25].

A correct digital representation of the territory depends directly on the quality of the data, which has led companies and institutions to become interested in digitizing their pre-existing information and designing new methods for data collection [26]. To avoid public rejection, companies like Google choose to hide these processes totally or partially behind apparently benign projects, such as their maps service, which is presented as a public service to democratize terrestrial cartography [3]. Data collection has evolved from manual and evident methods towards increasingly invisible and efficient techniques, increasing the ability to obtain information without the user being aware of it [27].

The introduction of AI models in geospatial data analysis has brought innovation along with a greater degree of complexity. In contrast to traditional spatial analysis approaches, which are based on principles such as Tobler's First Law of Geography [28] where geographic proximity determines relationships between elements, machine learning algorithms are capable of extracting complex patterns from large amounts of data, revealing connections that go beyond simple spatial proximity. Within machine learning models, neural networks or deep learning allow these algorithms to work with unstructured data, such as images, audio, or texts.

Machine learning data classification can use two main categories depending on whether or not a predefined order is sought: supervised and unsupervised. Supervised classification uses an algorithm with a predefined structure to apply it to new data. In the CLIP and the City research, the CLIP algorithm is used to classify urban panoramic images according to predefined cultural and architectural labels, spatially mapping the perceived characteristics of the city of Rome [29]. On the other hand, unsupervised classification infers a structure by observing patterns in the data, which implies a transition from deductive to inductive reasoning, starting from raw data to discover patterns. In the Urban Grammar AI research project, unsupervised learning is used to identify typological families of urban plots. This approach allows for the discovery of new categories of urban plots based on data such as shape, use, or construction period, which facilitates comparisons of urban fabric in different regions without relying on predefined classifications [30].

However, the statistical nature of these algorithms can blur nuances essential for a deep understanding of the phenomena they analyze [31]. These methods are based on mathematical functions that statistically calculate output values from input values. As a result, the generated models are statistical representations that approximate reality, but may not capture all of its complexity. This limitation becomes evident when attempting to quantify qualitative data. Algorithms, working exclusively with numbers, require the numerical translation of concepts previously considered impossible to measure. Various investigations seek to quantify qualitative concepts such as habitability, human perception of the city, or the subjective emotions produced by urban streets, through visual scores given by citizens to Street View images [14, 32, 33]. These practices promote the idea that these concepts can be measured and predicted, rather than considering them as consequences of factors impossible to anticipate. However, the translation of urban complexity through static images and user scores eliminates other sensitive and temporal dimensions of these environments. This simplification biases the representation, offering an idealized version of urban reality [27].

The handling of personal and sensitive information represents one of the main challenges in data analysis. In late 2023, OpenAI, the company that owns ChatGPT, created the Data Partnerships program with the aim of collaborating with global organizations in creating extensive databases that "reflect human society", explicitly avoiding the collection of "sensitive or personal information" [34]. The complexity and variety of this type of information, added to privacy considerations, frequently leads to the anonymization of data. Although this process does not affect some analyses, in other cases, such as in the study of specific communities, it can result in the loss of significant details [27].

Despite the apparent difficulties in working with non-anonymous and sensitive data, there are fields such as medicine where AI analyses must always be accompanied by rigorous guarantees of privacy, transparency, and respect for individual data ownership. In the context of territorial analysis, territorial intelligence, a concept developed by Jean-Jacques Girardot, facilitates researchers, actors, and communities in acquiring a deeper knowledge of their environment, allowing them to manage their development more effectively. This approach recognizes the inherent complexity of territories and proposes the appropriation of information and communication technologies as a crucial step for local actors to initiate a learning process, enabling them to act in a pertinent and effective manner [35].

As Latour points out, the ubiquity of digital data and the immediacy of "real time" incapacitate our ability to process and understand information, resulting in a superficial and inauthentic experience that threatens to homogenize it [36]. Collaborative maps such as "Queering the map", which records personal experiences about the LGBTQ2IA+ collective around the world [37], "Aporee", which focuses on geolocated sound [38], and Native Land Digital, which maps the past and present territories of indigenous nations [39], propose methodologies that, although slower and more complex, are necessary to preserve minority and alternative narratives that deviate from the normative. Co-creation between different actors and perspectives offers visions of space that reflect a diversity of experiences and situated knowledge.

The research seeks to study how automation affects social mapping. To do this, it replicates the standard methodology used in data science and machine learning, applying it to the collaborative mapping of a music festival's environment. The process includes: (1) data collection, (2) data processing, and (3) AI analysis. Manual analysis of unstructured data formats such as images, texts, and audio is usually very slow and costly due to the inherent complexity of each format. For this reason, we propose using multimodal algorithms capable of extracting patterns from large amounts of data in different formats and cross-referencing them in very little time. This classification will attempt to detect interesting patterns for agents in their mapping process.

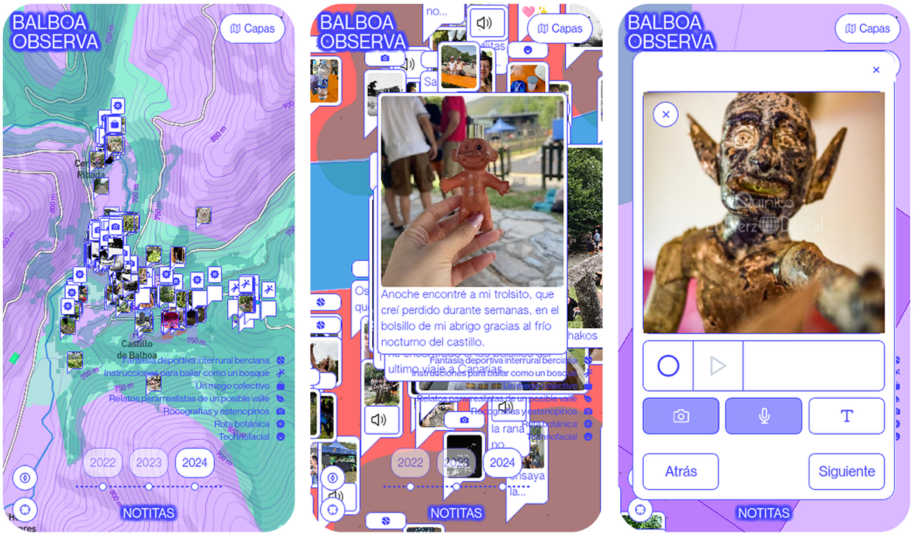
To compare human and computational perception, three analysis methods are proposed. First, a spatiotemporal analysis is proposed, an approach similar to that performed during social mapping where only physical position and recording time are taken into account to observe their distribution. The second involves supervised classification using multimodal algorithms, the conventional method of using this type of algorithm. In it, two data of different formats are compared to find connections between patterns, as can be seen in the CLIP and the City project [29]. The third analysis proposes a less conventional approach by using unsupervised classification on the activations of multimodal algorithms. Instead of studying the final response of the algorithm, its internal activations or reactions are studied, a process similar to an electroencephalogram. Through explanatory techniques normally used in algorithm training, it is possible to partially reveal the real internal connections and biases of representation, allowing mapping participants to automatically discover, among enormous amounts of data, patterns invisible to the human eye. Inspired by the unsupervised methodology of the Urban Grammar AI project [30], this analysis does not impose patterns on the algorithm but lets the algorithm freely organize the information. It is important to note that these algorithmic analyses do not intend to replace but complement human perception, providing different ways of understanding information.

4 Data Collection

Data collection was carried out during the years 2022, 2023 and 2024, focusing on the village environment during the four days of the Festival. With the aim of creating an interactive and real-time record, a specific web application was developed for the event, accessible at www.balboaobserva.es (Figure 1). This tool, optimized for mobile devices, allowed attendees to upload data (photos, texts or audios) on a map of Balboa and visualize the contributions of other participants. The recording methodology was inspired by the "Photovoice" technique, defined by Caroline Wang and Mary Ann Burris as "a process by which individuals can identify, represent and enhance their community through the use of a specific photographic technique, entrusting cameras to individuals to act as recorders and potential catalysts for change in their own communities" [2].

To promote its use, social networks, posters and postcards with QR codes were used, in addition to conducting explanatory workshops. Participants in the recording were guided to record the *everyday life* of the Festival, including not only the main scheduled events, but also more routine situations outside the program. Before uploading each piece of data, participants had to explicitly accept that the information would be public and would be used for this study, thus obtaining their consent. Although the application does not record the identity of who uploads the information, the data itself is not anonymized, and may contain personal and sensitive information.

Finally, it should be noted that biases in participation were identified, with a predominance of young users and a notably lower participation of the village's regular residents compared to Festival attendees. Although the objective of the research is not so much to obtain a faithful representation of reality as to understand the distortion introduced by machine learning algorithms, these biases should be considered for the interpretation of the results and to avoid erroneous conclusions.


BALBOA OBSERVA
Capas
Castillo de Balboa
Fantasía deportiva interrural berciana
Instrucciones para bailar como un bosque
Un riego colectivo
Relatos pararrealistas de un posible valle
Rocografías y estenopinos
Ruta botánica
Tecnofacial
2022
2023
2024
NOTITAS
ultimo viaje a Canarias
Atrás
Siguiente

Anoche encontré a mi trolsito, que creí perdido durante semanas, en el bolsillo de mi abrigo gracias al frío nocturno del castillo.

Fig. 1. Balboa Observa map interface where participants can easily upload images, audios and texts during the festival.

5 Data Processing

After registration, the data was manually validated. Duplicates, those located outside the village limits or unrelated to the Festival were excluded, and incorrect locations were corrected. The content of images, texts and audios was not modified. Over the three years, 1107 images, 233 texts, and 76 audios were collected. The data collected during this time was aggregated into a single representative day, allowing for a more robust and generalizable analysis. This methodological decision is justified by the notable consistency observed in the spatio-temporal distribution patterns, attributable to the permanence of schedules and spaces of Festival activities during the three years analyzed.

To discover deeper connections between the data, beyond spatial or temporal correlations, it is proposed to process the information using multimodal algorithms. These algorithms are capable of cross-referencing concepts between different formats, such as the images, texts, and audios from the map. Two types of algorithms are proposed:

- CLIP (Contrastive Language-Image Pre-training), which connects text and images without the need for specific labels [40]
- ImageBind, which extends these capabilities to other modalities, including audio [41].

It is important to note that the identified relationships are influenced by the original training data of the models, so their use introduces biases in the interpretation of results. Although it would be ideal to train the multimodal algorithms with data relevant to the study to obtain more precise connections, this process is complex and computationally expensive. Therefore, the option is to use pre-trained models, recognizing that the established connections may not perfectly fit the specific context, but offer a solid basis for analysis.

The multimodal algorithms process images, audios and texts through one of their parts called encoder that encodes them into numerical vectors called *embeddings* and that act as abstract representations of the information with which the algorithm can work. These vectors are composed of a set of dimensions or parameters, whose quantity varies according to the algorithm employed. In this case, the largest and most advanced models are chosen such as CLIP-ViT-L/14@336px which uses vectors of 765 parameters and ImageBind_huge which employs 1060 to encode the information [40, 41]. It is essential to highlight that these numerical values are not directly interpretable by human beings; their understanding is reserved exclusively for the algorithm.

6 Data Analysis

Following the methodology employed in social mapping, an analysis oriented to generate new visualizations of the territory is carried out. In this case, three differentiated analyses are implemented with the objective of comparing both the benefits and the limitations inherent to each methodological approach.

6.1 Spatiotemporal Analysis

The collected data were manually classified through visual analysis to identify spatial and temporal clustering patterns. The results show that the highest density of points is concentrated in specific areas of the village, particularly around the central spaces and during the main moments of the Festival. In spatial terms, six main areas were identified from north to south where activity was most intense: Chis's Palloza; the Camping; the central space of the village known as the Pradera; the residential area of the town; the Amphitheater and the Forest. The temporal analysis reveals distinctive usage patterns: scarce activity in the morning, concentration in the Pradera and Chis's Palloza at midday, shift to the Camping during dinner time, and focus on the Amphitheater and Chis's Palloza at night (Figure 2).

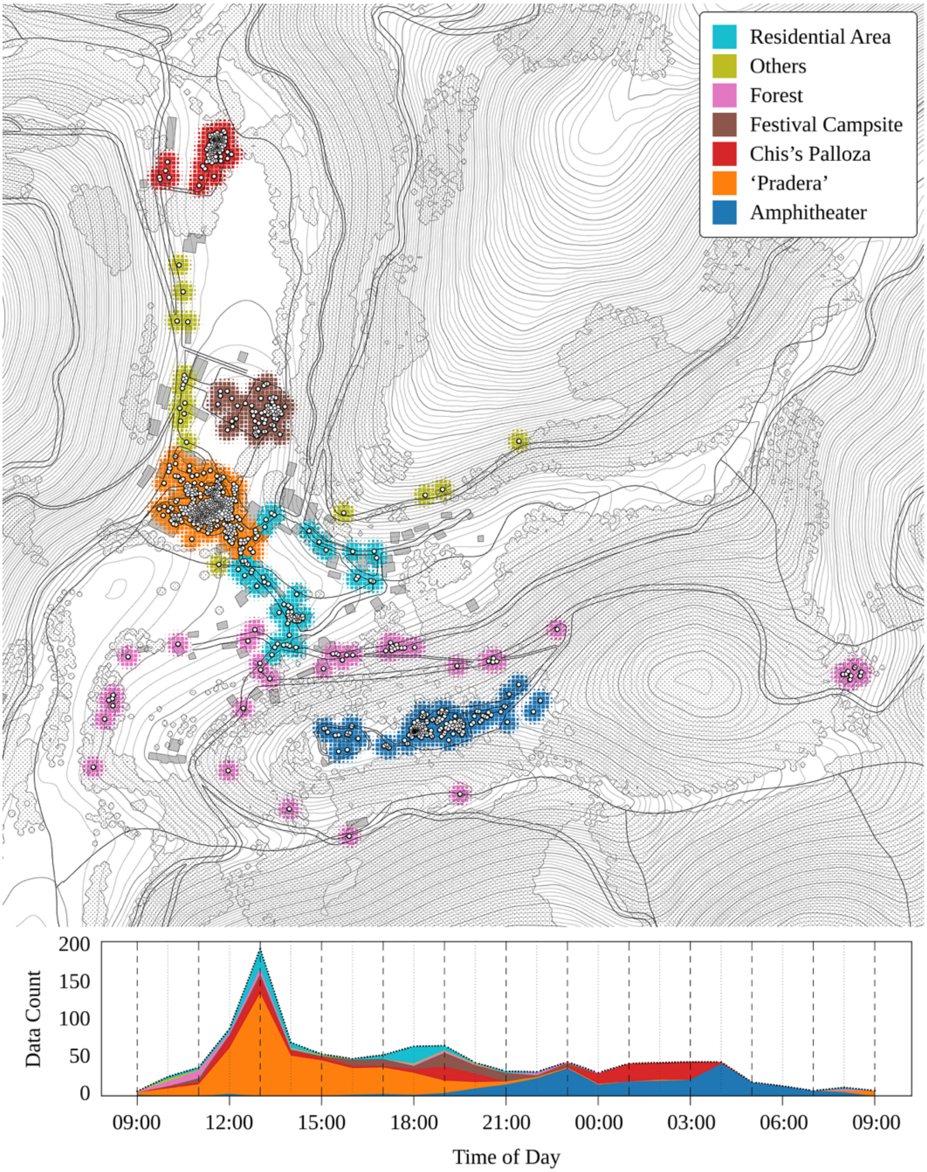


Fig. 2. Spatiotemporal distribution of aggregated data across Balboa's main spaces.

The combination of spatial and temporal analyses, as shown in Figure 3, allows for the disaggregation of data that share space but not time, revealing specific usage patterns. This is evidenced in the dual use of spaces such as Chis's Palloza, used for workshops during the day and concerts at night, or a plot in the Forest that hosts daytime workshops and in the early morning becomes an extension of the nighttime party.

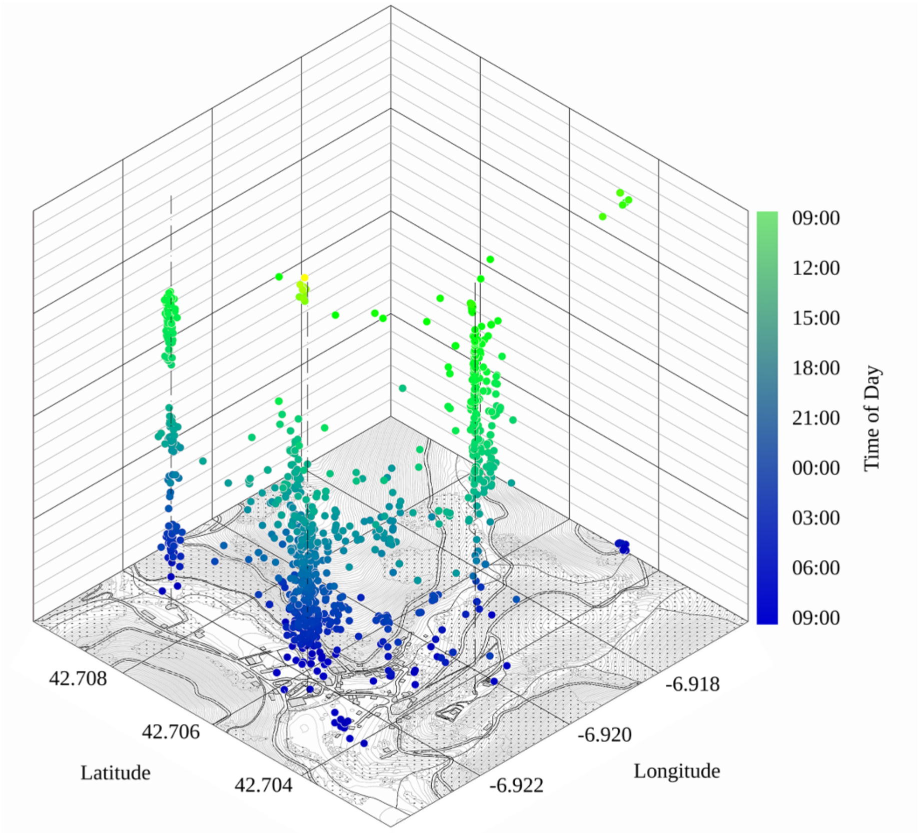


Fig. 3. 24-hour Spatio-temporal Distribution of Aggregated Data.

6.2 Supervised Analysis

The classification of information through multimodal models leverages the zero-shot capabilities of these algorithms, that is, their ability to classify data that was not included during their training phase. This approach constitutes the methodological basis proposed in CLIP and ImageBind research [40, 41] and represents the standard method in studies employing these algorithms [29].

The classification process performs the comparison of embeddings from data with different formats through cosine similarity calculation. This study specifically implements the comparison between images and textual categories, given that images present greater territorial coverage. Multiple festival activities were defined and grouped into five categories: Community and Leisure, Music and Party, Workshops and Talks, Daily Needs, and Organization and Logistics. To optimize model performance, considering its original training base, texts were written in English preceded by the expression "A picture of a," thus imitating the characteristic descriptions of its training database. Additionally, to address possible misalignment between representations, multiple variations of each textual description were generated, including terms specific to the Spanish context such as "siesta."

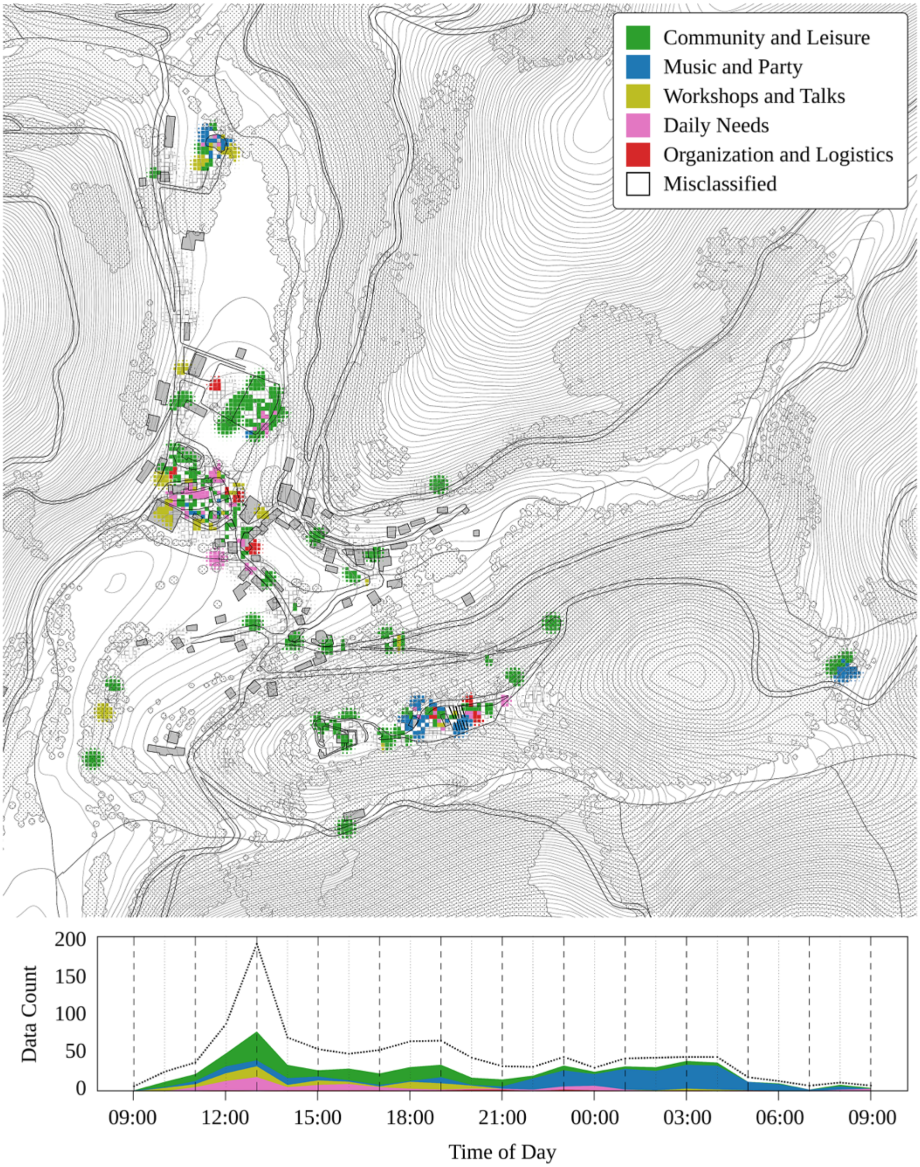


Fig. 4. Spatiotemporal distribution of categories correctly detected by zero-shot classification and, in white, the misclassified data.

The system individually calculates the similarity between each image and the activities. They are processed through a softmax function that transforms them into normalized probabilities that sum to 100%, identifying the category with the highest connection according to the model. No probability threshold was established to determine the validity of classifications, both due to the absence of a clear criterion and to evaluate the general certainty of the system. Instead, a manual filtering process was implemented image by image to verify the correspondence between automatic classification and the correct category. Only those results that presented alignment with the researchers' evaluation were counted in the final analysis.

After comparative evaluation of both algorithms, exclusively the results obtained through CLIP were selected, given that ImageBind showed consistently lower probabilities and a greater number of errors detected during manual review. Of the total 1107 images, only 56% of the images were correctly categorized. The algorithm had problems mainly categorizing images where no clear activities appeared or with clear misalignments between the algorithm and Balboa when classifying local village people as people in costumes. Of the total well-categorized data, 58% were below a probability of 0.5, indicating that although it categorized correctly, in most cases it did not have very high certainty. The categories with the most errors were "people relaxing in shade," probably due to the use of the word "shade" which had more weight than the rest of the phrase, and "people taking a siesta," which generated very disparate results due to the Spanish term. The results of the algorithm can be seen in Figure 4 compared with the total data that was analyzed.

6.2 Unsupervised Analysis

The third analysis seeks to categorize the data without using predefined categories. This approach avoids confronting the algorithm with representations structured differently from its training database, allowing it to organize the information according to its intrinsic structure. This analysis, used in research such as Urban Grammar, studies the different characteristics of each data point to identify groups that share common features.

In this case, each data point includes as characteristics its spatial and temporal variables, incorporating as an innovative element the numerical values of the embeddings. These values contain the characteristics identified by the algorithm. Although they are encoded and it is not possible to extract the numerical values corresponding to individual characteristics, it is feasible to use these values together with spatial and temporal data to detect groups with similar values and, consequently, comparable characteristics.

Each embedding acts as a coordinate in a multidimensional space, where its location reflects the relationship with other elements in the distribution learned by the system. This multidimensional space is known as "latent space" and constitutes a representation of the model learned from the training database. In this space, points that present similar characteristics are grouped together, while those less related distance themselves from each other. For example, if we draw a vector between the points representing the concepts of "day" and "night", and move its origin to the point representing "Concert in the Pradera", the model could point to the concept "Concert in the Amphitheater". The organization of information in the latent space facilitates the identification of patterns and similarities that would not be evident in the original data, allowing the discovery of connections between different aspects of urban space planning and use.

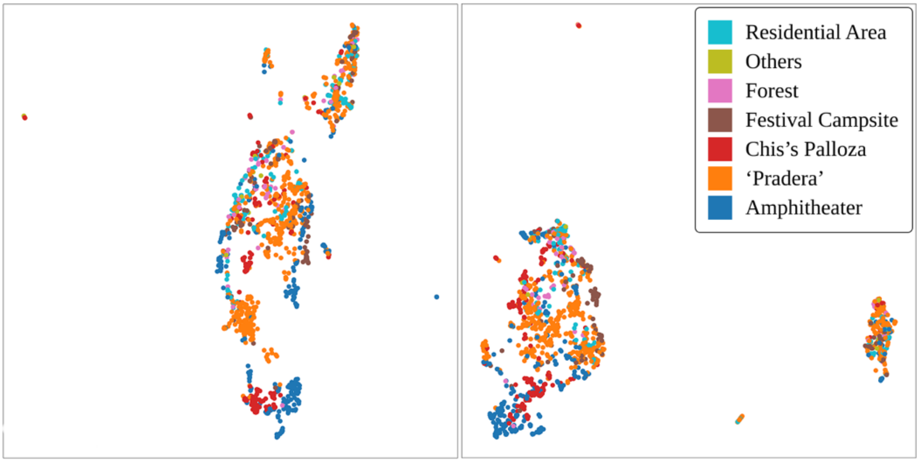


Fig. 5. Balboa's main spaces distribution on the latent spaces of ImageBind (left) and CLIP (right).

The multidimensional latent space requires dimensional reduction techniques for its visualization. Among these techniques, such as PCA, t-SNE or UMAP, UMAP is selected for its superior ability to preserve data structure in high-dimensionality embeddings [42]. These techniques transform complex data into two-dimensional or three-dimensional visual representations, maintaining the original relationships [42]. This visualization allows the projection of information associated with the main spaces of the town, previously ordered by its geospatial position, into the latent space generated by the algorithms. As shown in figure 5, the algorithmic connections distort and mix the spaces of the town, transcending the traditional spatial and temporal correlations of urbanism [28].

To identify areas of the town where common concepts exist, the processed data was analyzed using clustering techniques. Clustering, an unsupervised learning method, was employed to group similar data into sets called clusters, based on the similarity of their characteristics. Two clustering algorithms were selected: K-means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). K-means, chosen for its effectiveness in analyzing multidimensional data, groups the data into a specific number of clusters, minimizing intra-cluster variance [43]. DBSCAN, selected for its better performance with irregularly shaped clusters and its robustness against noise, groups points in high-density regions, separating them from low-density regions [44].

The analysis was conducted by separately examining the spatiotemporal characteristics, CLIP embeddings, and ImageBind embeddings, as well as all their possible combinations. The concatenation of data and embeddings is a common practice that allows for increasing dimensionality and offering a richer representation to the algorithm. The clustering methods were applied to both the original datasets and those subjected to dimensional reduction using UMAP. This reduction facilitates the algorithm's task by operating with a fewer number of variables, although it entails a loss of details in the representation.

Additional datasets were created using only images due to the detection of the "Modality Gap" phenomenon, which caused the clustering to return groups exclusively of texts and others of audios, justifying the separate analysis of images to obtain more coherent results. To determine the optimal number of groups that best represented the dataset, various validation techniques were used:

- Elbow Method: Used exclusively for K-means analysis, it provides an intuitive visualization of where improvement stabilizes as the number of clusters increases.
- Silhouette Index: Measures the similarity of a data point to its own cluster compared to other clusters, providing a measure of the quality and consistency of the groups [45].
- Davies-Bouldin Index (DBI): Offers a more quantitative measure of the separation between clusters and takes into account the dispersion within them.
- Noise Point Analysis: Applied specifically in DBSCAN, it evaluates the algorithm's effectiveness in identifying outliers.

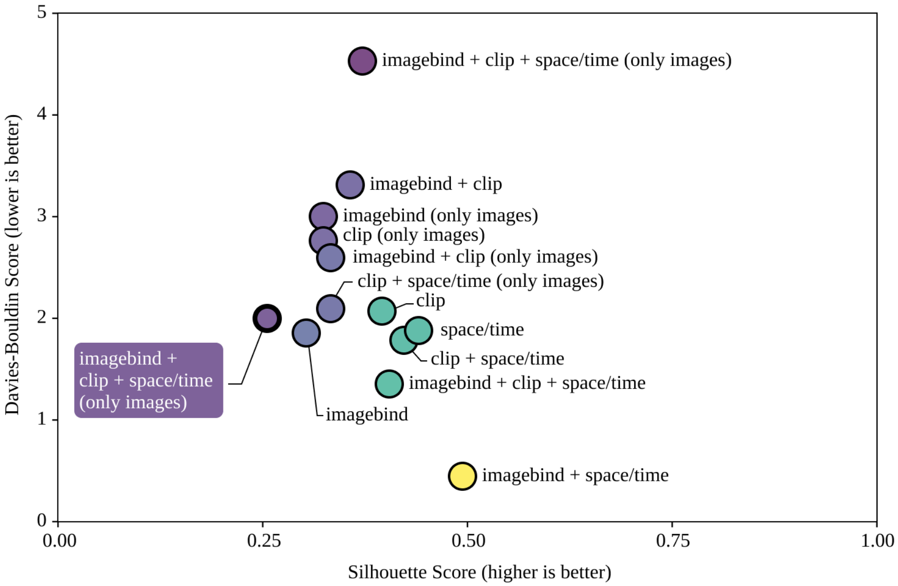


Fig. 6. Comparison of DBSCAN clustering performance across all analyses, where the effect of the "Modality Gap" produces generally lower scores in clusterings based solely on images.

These techniques were selected for their complementarity, offering a comprehensive evaluation of the quality and robustness of the clustering. The Silhouette Index was prioritized due to its ability to simultaneously evaluate internal cohesion and separation between clusters [45]. In a complementary manner, a visual exploration of the content was carried out to understand the main characteristics of the clusters. The centers of each group were calculated and the closest data points were obtained for visualization, considering them as the most representative of each cluster.

One of the resulting models is selected for more in-depth study. Analyses that did not include spatio-temporal parameters and at least one of the algorithms were discarded, as data combination is one of the research objectives. The analyses that included texts and audios caused the clustering algorithm to generate specific clusters for these media, evidencing the effect of the "Modality Gap" and the persistent difficulty in achieving real multimodal understanding [46]. Consequently, these analyses were discarded. Between K-means and DBSCAN, K-means had very poor performance in detecting clusters, likely due to the complex and non-convex shapes that can be observed in the UMAPs. Therefore, it was decided to choose DBSCAN.

For the selection of the final clustering analysis, three evaluation criteria were established: (1) optimal values in the Silhouette and Davies-Bouldin indices that confirm the quality of the grouping, (2) visual coherence of the clusters through graphical inspection, and (3) theoretical relevance for the analysis of the everyday dimension of the Festival. The dataset that integrates spatio-temporal information with CLIP and ImageBind embeddings met these requirements most satisfactorily. Although, in figure 6, it is observed that this dataset presented the lowest Silhouette index among the analyses based solely on images, it recorded the most favorable Davies-Bouldin index and showed clusters with highly significant distinctive characteristics for the proposed social mapping.

The final analysis presents 10 clusters that capture different aspects of the Festival, as can be seen in figures 7 and 8. These can be visually grouped into broader categories for a better understanding of the representation offered by the algorithm. Firstly, areas related to nighttime festivities are identified. The "Performance Area" represents the core of the parties, the stages, sound equipment, and artists. In contrast, the "Festive Periphery" shows more everyday situations on the margins of nighttime concerts such as eating at food stalls or informal post-concert parties.

Daytime leisure activities are reflected in several data families. The "Daytime Leisure Area" points to spaces that host events organized by the Festival such as concerts, workshops, and games. It also indicates other uses such as rest areas where attendees sunbathe or enjoy siestas. The areas called "Human portraits" focus on humans as protagonists, highlighting their interactions, outfits, and selfie culture, rather than the space where it occurs. The "River Beach" is distinguished as a differentiated meeting point, dedicated to bathing activity in the dammed area of the river.

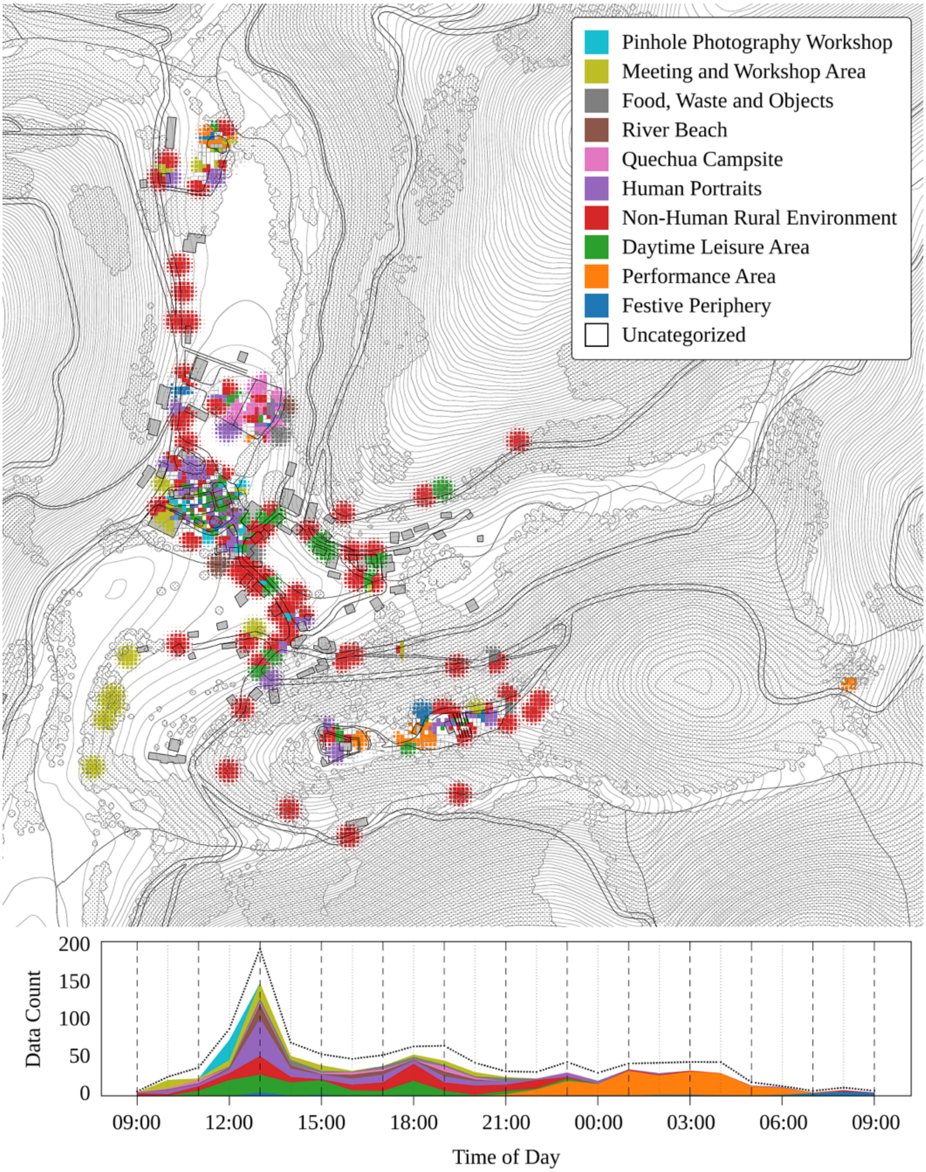


Fig. 7. Spatiotemporal distribution of algorithm-detected clustersand and, in white, the uncategorized data.

The habitable condition of the Festival is evidenced in groupings such as the "Quechua Campsite", characterized by a high density of Quechua brand tents from Decathlon. "Food, Waste and Objects" shows the concentration of objects related to ways of inhabiting these informal settlements. Likewise, cultural and collaborative activities are represented by the "Meeting and Workshop Area", which exhibits human grouping patterns oriented towards workshops or group sessions. A specific example is the "Pinhole Photography Workshop", characterized by the use of homemade cameras to take black and white photographs. Finally, the "Non-Human Rural Environment" group offers an interesting contrast, showing a notable absence of direct human presence and focusing on non-human elements, crops, and rural structures of the surroundings.

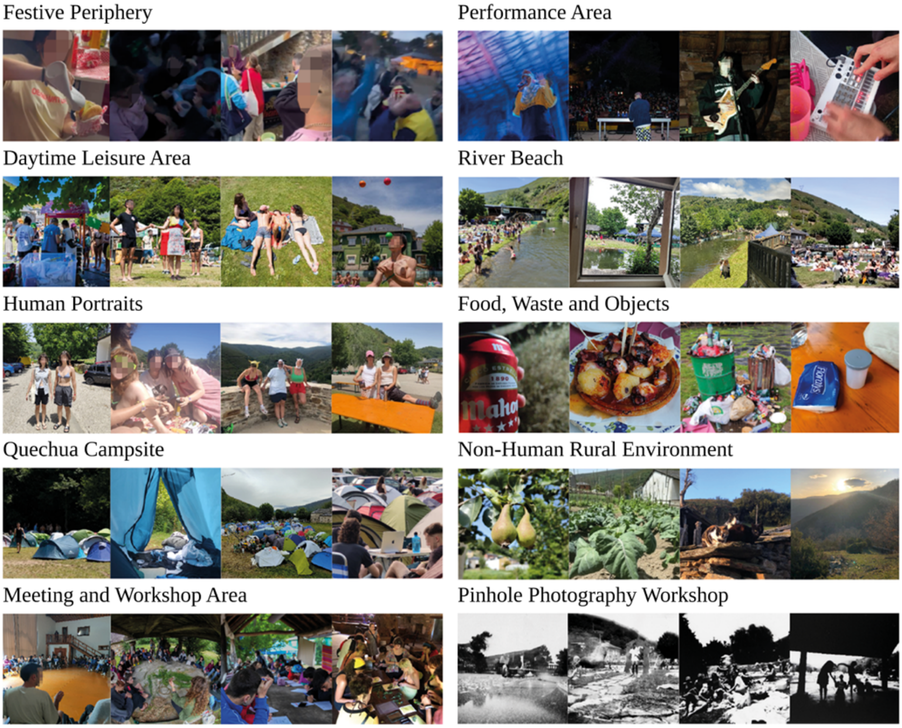


Fig. 8. Images sampled from cluster centroids (the mean point of all data points in each cluster)

7 Discussion

Artificial intelligence (AI) based analyses offer a new perspective on the territorial planning of Balboa and the daily life of the Festival Observatorio, generating structures different from traditional human planning. Compared to the conventional use given to multimodal algorithms in data analysis, the unsupervised method proposed in the research offers more interesting arrangements. Although this approach brings new views to the territory, it also presents significant challenges when applied to social mapping processes, including algorithmic biases, the need for transparency in decision-making, and inherent technical complexity.

The geographic visualization of the groupings generated by the algorithm presents significant differences compared to those made by humans, particularly in their spatio-temporal distribution. However, the representation of these groups in the latent space reflects an ordering closer to what a person could perform, as can be verified in figure 9. This similarity is attributed to the fact that the training database uses information generated and labeled by people [40].

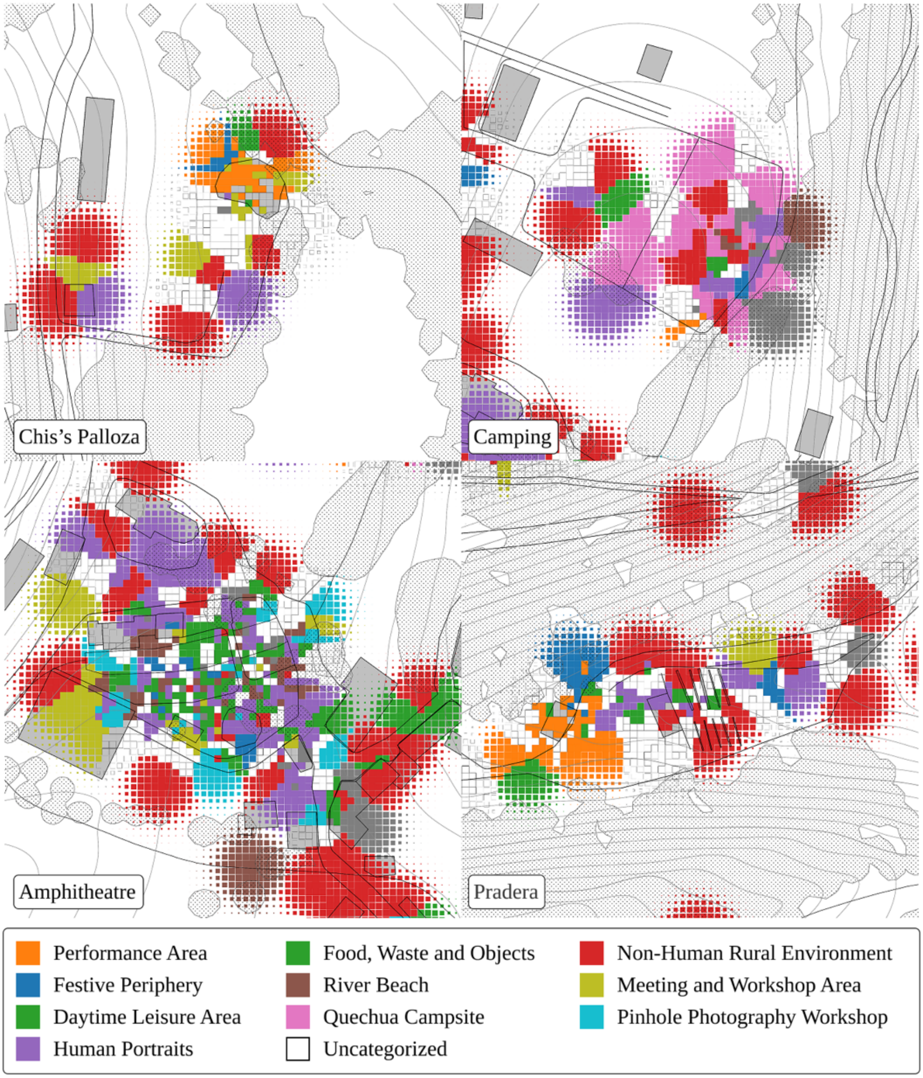


Fig 9. Cluster spatial distribution in data-dense regions.

The difference lies in that artificial intelligence can process thousands of different data simultaneously, finding measurable relationships between them. This leads it to propose more heterogeneous and specific zones, combining areas, activities, and times, in contrast to the human tendency to create more homogeneous structures to handle large amounts of information. The unsupervised analysis through multimodal algorithms made it possible to disaggregate and understand from new perspectives the data grouped in the same cluster within high-density areas, such as the Pradera, Chis's Palloza, the Camping, and the Amphitheater, as evidenced in figure 9. The superposition of multiple areas in these spaces generates spatial configurations that are not intuitive for a human, but that can be understood as the superposition of multiple realities in the same space and time. This approach overcame the limitations of simple spatiotemporal analysis, offering a deeper and more nuanced understanding of the festival dynamics. The results of the supervised analysis also fell behind due to the need to introduce categories and the difficulty of these coinciding with those of the algorithm. The freedom that the unsupervised methodology allows the algorithm made it possible to identify patterns and relationships that were not evident in the direct visualization of the data, providing new perspectives on the distribution of the data.

These analyses stand out for their ability to generate categories from the particular to the general, allowing the emergence of unexpected classifications. For example, it was possible to differentiate between core activities of the festival (such as concerts and workshops) and more peripheral and everyday situations. Additionally, these methods allow for identifying previously undetectable biases, such as the formation of groups based on distinctive visual characteristics. Examples of this are the "Human Portraits" area, due to the selfie format, or the images from the "Pinhole Photography Workshop", which formed an outlier due to their black and white style. These results underscore the importance of the chosen approach in data collection, as they reflect the intention to represent the daily life of the Festival.

Unsupervised analysis presents advantages over supervised analysis in handling data that does not fit into any predefined category. While in supervised analysis, misclassified data detected through visual inspection is eliminated from the representation, uncategorized data in unsupervised analysis simply indicates that the algorithm does not find significant relationships between these elements and the rest of the dataset. This does not imply erroneous categorization, as their position in the latent space and their proximity to other groups can also provide valuable information about their characteristics. Furthermore, as shown in figure 10, the visualization of the latent space is considerably more interpretable than the individual numerical probability results produced by zero-shot classification for each data point. Although this representation presents some bias due to dimensional reduction, which inevitably involves information loss, the ability to observe relationships between all data simultaneously facilitates a more intuitive and comprehensive interpretation of the information contained in the dataset.

The use of these tools to analyze the daily realities of the Festival through algorithms also faces significant challenges due to the multiplicity of final representations and the dependence on human interpretation. Although there are mathematical methods to evaluate the reliability of interpretations, these do not completely guarantee the objectivity of the analysis. For example, the results could suggest that the concatenation of spatiotemporal connections with multimodal embeddings facilitated the formation of clusters, but this hypothesis requires confirmation through additional analyses. Currently, there is no objective method to determine which ordering generated by the algorithm is closer to reality, beyond visual observation. This implies that all orderings produced by the various combinations could be relevant and it would be erroneous to interpret the data groupings as real spatial configurations without due verification and interpretation.

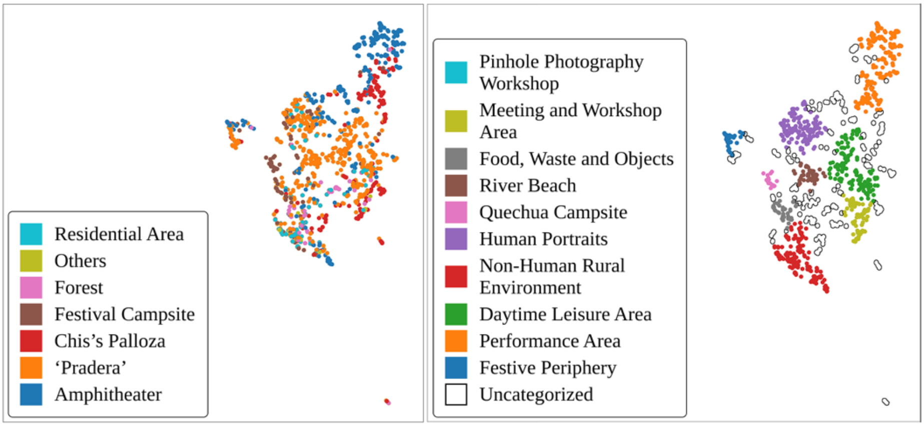


Fig 10. Comparison of latent space distributions: primary Balboa areas (left) vs. algorithm-detected clusters (right).

An additional problem identified in this research is that the multimodal algorithms used have predefined relationships that may not correspond to the studied reality, given that they have been trained with internet data not specialized in territorial analysis [47]. Although clustering is performed through unsupervised learning without a predefined structure, the training data of multimodal algorithms generate a structure that introduces biases. This affects their ability to analyze everyday and specific realities that deviate from the distribution of these algorithms, potentially concealing alternative realities.

A possible solution would be the alignment of the algorithm through the creation of specific databases on territory, urbanism, or citizen participation. The representations proposed in this research help detect the biases introduced by these algorithms. These databases should be updated with contemporary and relevant information for each application context, with the citizen participation of the communities. In the specific case of music festivals in rural environments, it would be necessary to expand the study to more events, collecting detailed and specific information from each context to obtain a more complete picture of the phenomenon.

However, this approach also presents problems by requiring the transfer of sensitive community data to third parties. The creation of these algorithms not only implies technical complexity, but also involves social and political decisions comparable to the definition of urban laws. A more respectful alternative with community data would be to create methods capable of guiding the algorithm, modifying its responses to align them in other directions while the data never leaves the community's possession. An example of this is LoRA (Low-Rank Adaptation) technology, an algorithm that couples to machine learning systems and allows selectively adjusting certain parameters of the models to reduce specific biases without the need to retrain the entire system. [48]

The processing speed of these algorithms observed in the research would allow real-time analysis both during data collection and in the final mapping phase, giving participants greater control over the final result. The implementation of these algorithms in social mapping processes and citizen participation presents an interesting potential as tools to efficiently handle situations with abundant information. In the final phases of mapping, when participants analyze the results to extract conclusions, these algorithmic analyses would complement the human perspective by providing a computational vision that can reveal non-evident patterns and relationships, thus enriching the collective interpretation process.

However, it is crucial to implement adequate measures in data processing to avoid erroneous conclusions derived from the overwhelming amount of information handled. Given the emerging nature of this technology, it is necessary to develop protocols that guarantee the reliability of results and the transparency of algorithmic processes. The appropriate use of these systems involves establishing clear criteria for result validation, interpretation protocols that combine automated analysis and human supervision, and feedback mechanisms that allow user communities to understand and evaluate the algorithmic processes employed. Although the transparency of these complex systems is not yet resolved and their technical complexity is high, the main concepts and intuitions are easy to understand.

The results indicate that the improvement of the tool should focus on making the flow of information transparent, quickly revealing the relationships that the algorithm articulates. It is crucial to show which realities are overrepresented in the final configuration to allow timely corrections. To improve the direct applicability of these algorithmic methods, it is crucial to develop techniques that allow the algorithm to reveal real connections and main characteristics of the groupings in a more objective manner. To mitigate biases, the tool should facilitate their detection in situ during data recording, allowing a faster and more complete understanding of the situation. The tool should offer flexibility to modify mapping parameters during recording, adapting to unforeseen situations. For example, it can promote the inclusion of the perspective of local residents, whose participation in data recording was limited. It is important to consider that the use of mobile phones can pose a barrier for certain users with difficulties accessing these media. The accessibility of the tool is fundamental so that people without technical knowledge can modify it and obtain an accurate representation of the data.

Data ownership should remain in the hands of its users and legitimate owners. To achieve this, it is essential to provide local actors with appropriate computational tools, whose technical elements are generally affordable in terms of cost and maintenance. The use of local and private servers offers an additional layer of security for personal and sensitive data, especially valuable when there is a preference not to expose information to external systems. It is crucial that the service be subject to review by any community member, ensuring that the collection and use of information are strictly limited to the agreed purposes. The implementation of more transparent participatory processes in AI systems should include detailed forms of consent, granting individuals greater control over the final representations of their data. Likewise, it is fundamental to develop mechanisms that allow participants to understand how their information will be used and have the ability to modify or withdraw their consent at any time.

8 Conclusions

Research on AI applications in social mapping reveals both advantages and challenges. Its main strength lies in the ability to quickly process large volumes of data, establishing specific connections between large amounts of information. However, its use must be careful and should not replace other established practices. Although the analytical complexity of this technology may appear more reliable than human analysis, the lack of real verifications makes human supervision by the involved communities indispensable.

The research reveals that the collaborative development of multimodal algorithms with local actors is fundamental to adequately reflect local realities in social mapping. This participatory approach would allow capturing the relationships of interest and specific characteristics of each study environment. The results of the case study on music festivals in rural settings underscore the need to expand research to a wider variety of events and contexts to obtain a more complete and nuanced understanding of these social phenomena. Future studies should consider how to adapt and refine these algorithms to improve their accuracy and relevance in different social mapping scenarios.

The importance of improving transparency and flexibility in AI-based social mapping tools is highlighted. Future developments should focus on more clearly revealing the relationships articulated by the algorithms and allowing real-time adjustments. It is essential to develop methods that facilitate the objective identification of main connections and characteristics in the data, as well as the detection of possible biases in the results. The potential application of these analyses in real-time opens up new possibilities for active citizen participation in social mapping processes.

The ethical application of AI in social mapping requires maintaining ownership and control of data in the hands of users and local communities. It is imperative to develop more transparent and participatory processes in the collection and use of personal data, implementing more detailed and flexible consent mechanisms. The feasibility of using local infrastructures for these tools offers a promising solution to ensure the security of sensitive data. The future of this field must prioritize empowering communities in managing their own data and in supervising the AI processes used in social mapping, thus ensuring a more informed and equitable citizen participation.

In this pioneering stage of AI, the desirable transparency and democratization should extend to the algorithms themselves. Revealing to individuals and communities the interpretations that these produce, explaining their generation and origin, would allow greater control over these processes and facilitate collaboration in the development of multimodal algorithms more representative of everyday realities. This openness is fundamental if one intends to think the world in collaboration with other intelligences and that they entrust their quality and trust in citizenship.

Acknowledgments. To Santiago Gómez Plata for helping to start this project and to the Observatorio Festival and the town of Balboa for their support.

References

1. Sanz Simón C.: Del Molino Molina, Sergio (Eds.): La España vacía. Viaje por un país que nunca fue. Madrid, Turner Noema, 2016. 291 pp Cuad Hist Contemp, 40, (2018) https://doi.org/10.5209/CHCO.60347
2. Wang C., Burris M.A.: Photovoice: Concept, Methodology, and Use for Participatory Needs Assessment Health Education and Behavior, 24, (1997) https://doi.org/10.1177/109019819702400309
3. Anguelov D., Dulong C., Filip D., Frueh C., Lafon S., Lyon R., Ogale A., Vincent L., Weaver J.: Google street view: Capturing the world at street level Computer (Long Beach Calif), 43, (2010) https://doi.org/10.1109/MC.2010.170
4. How Street View works and where we will collect images next, https://www.google.com/intl/en_uk/streetview/how-it-works/
5. Create and publish your own Street View imagery https://www.google.com/intl/en_uk/streetview/contribute/
6. Google Maps 101: How AI helps predict traffic and determine routes, Google Blog, https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/
7. About phone calls from Google Assistant https://support.google.com/business/answer/7690269?hl=en
8. Panoramio, https://www.panoramio.com/
9. Local Guides, https://maps.google.com/localguides/
10. Kopf J., Chen B., Szeliski R., Cohen M.: Street slide: Browsing Street level imagery ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010 (2010) https://doi.org/10.5209/CHCO.60347
11. Gebru T., Krause J., Wang Y., Chen D., Deng J., Aiden E.L., Fei-Fei L.: Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States Proc Natl Acad Sci U S A, 114, (2017) https://doi.org/10.1073/pnas.1700035114
12. Bieri V., Zamboni M., Blumer N.S., Chen Q., Engelmann F.: OpenCity3D: What do Vision-Language Models know about Urban Environments? (2025) https://doi.org/10.1109/WACV61041.2025.00503
13. World scale inverse reinforcement learning in Google Maps https://research.google/blog/world-scale-inverse-reinforcement-learning-in-google-maps/
14. Quercia D., Schifanella R., Aiello L.M.: The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city HT 2014 - Proceedings of the 25th ACM Conference on Hypertext and Social Media (2014) https://doi.org/10.1145/2631775.2631799
15. Lyons S.: Satellite surveillance and the orbital unconscious New Media Soc, (2023) https://doi.org/10.1177/14614448231187352
16. Wood D.: The power of maps Sci Am, 268, (1993) https://doi.org/10.1038/scientificamerican0593-88
17. Graham M., De Sabbata S., Zook M.A.: Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information Geo, 2, (2015) https://doi.org/10.1002/geo2.8
18. OpenStreetMap, https://www.openstreetmap.org/
19. Air/ Aria/ Aire A + U-Architecture and Urbanism, (2021)

20. Ciutat Vella’s Land-Use Plan, https://urbannext.net/ciutat-vellas-land-use-plan/
21. Santamaria-Varas M., Martinez-Diez P.: Cartografías de la ciudad nocturna a través del Big Data Obra digital, (2014) https://doi.org/10.25029/od.2014.41.6
22. Larsen J.E., Sapiezynski P., Stopczynski A., Mørup M., Theodorsen R.: Crowds, bluetooth, and rock’n’roll: Understanding music festival participant behavior PDM 2013 - Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia, Co-located with ACM Multimedia 2013 (2013) https://doi.org/10.1145/2509352.2509399
23. Argota Sánchez-Vaquerizo J., Delso Gutiérrez R., Gómez Saiz A.: Watching Puerta del Sol. On Protest Space and its Temporal Conflicts Open!, (2017)
24. Lever J., Arcucci R.: Sentimental wildfire: a social-physics machine learning model for wildfire nowcasting J Comput Soc Sci, 5, (2022) https://doi.org/10.1007/s42001-022-00174-8
25. Kuziemski M., Misuraca G.: AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings Telecomm Policy, 44, (2020) https://doi.org/10.1016/j.telpol.2020.101976
26. Arribas-Bel D.: Accidental, open and everywhere: Emerging data sources for the understanding of cities Applied Geography, 49, (2014) https://doi.org/10.1016/j.apgeog.2013.09.012
27. Thatcher J., O’Sullivan D., Mahmoudi D.: Data colonialism through accumulation by dispossession: New metaphors for daily data Environ Plan D, 34, (2016) https://doi.org/10.1073/pnas.1700035114
28. Tobler W.R.: A Computer Movie Simulating Urban Growth in the Detroit Region Econ Geogr, 46, (1970) https://doi.org/10.2307/143141
29. del Castillo N.: CLIP and the City: Addressing the Artificial Encoding of Cities in Multimodal Foundation Deep Learning Models. In: On Architecture Challenges in Design, pp. 100–109 (2023) https://doi.org/10.60152/eun81fru
30. Arribas-Bel D., Fleischmann M.: Understanding (urban) spaces through form and function Habitat Int, 128, (2022) https://doi.org/10.1016/j.habitatint.2022.102641
31. Can AI Solve Science?, http://writings.stephenwolfram.com/2024/03/can-ai-solve-science
32. Naik N., Philipoom J., Raskar R., Hidalgo C.: Streetscore-predicting the perceived safety of one million streetscapes IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2014) https://doi.org/10.1109/CVPRW.2014.121
33. Arturo, http://arturo.300000kms.net/
34. OpenAI Data Partnerships, https://openai.com/index/data-partnerships/
35. Girardot J.-J.: Inteligencia Territorial y Transición Socio-Ecológica Trabajo, 23, (2011) https://doi.org/10.33776/trabajo.v23i0.956
36. May, J: Signal. Image. Architecture (2019)
37. Queering the map, https://www.queeringthemap.com/
38. Aporee, https://aporee.org/maps/
39. Native Land Digital, https://native-land.ca/
40. Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I.: Learning Transferable Visual Models From Natural Language Supervision Proceedings of Machine Learning Research. Vol. 139 (2021) https://doi.org/10.48550/arXiv.2103.00020
41. Girdhar R., El-Nouby A., Liu Z., Singh M., Alwala K.V., Joulin A., Misra I.: ImageBind One Embedding Space to Bind Them All Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2023- June (2023) https://doi.org/10.1109/CVPR52729.2023.01457
42. McInnes L., Healy J., Saul N., Großberger L.: UMAP: Uniform Manifold Approximation and Projection J Open Source Softw, 3, (2018) https://doi.org/10.21105/joss.00861

43. Lloyd S.P.: Least Squares Quantization in PCM IEEE Trans Inf Theory, 28, (1982) https://doi.org/10.1109/TIT.1982.1056489
44 Ester M., Kriegel H.P., Sander J., Xu X.: A Density-Based Algorithm for Discovering Clusters A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Proceedings - 2nd International Conference on Knowledge Discovery and Data Mining, KDD 1996 (1996) https://dl.acm.org/doi/10.5555/3001460.3001507
45. Rousseeuw P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis J Comput Appl Math, 20, (1987) https://doi.org/10.1016/0377-0427(87)90125-7
46. Liang W., Zhang Y., Kwon Y., Yeung S., Zou J.: Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning Advances in Neural Information Processing Systems. Vol. 35 (2022) https://doi.org/10.48550/arXiv.2203.02053
47. Schuhmann C., Beaumont R., Vencu R., Gordon C., Wightman R., Cherti M., Coombes T., Katta A., Mullis C., Wortsman M., Schramowski P., Kundurthy S., Crowson K., Schmidt L., Kaczmarczyk R., Jitsev J.: LAION-5B: An open large-scale dataset for training next generation image-text models Advances in Neural Information Processing Systems. Vol. 35 (2022) https://doi.org/10.48550/arXiv.2210.08402
48. Hu E., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W.: LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS ICLR 2022 - 10th International Conference on Learning Representations (2022) https://doi.org/10.48550/arXiv.2106.09685