IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 February 2025, accepted 8 April 2025, date of publication 17 April 2025, date of current version 28 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3562095

RESEARCH ARTICLE

Spatiotemporal Epidemiological Similarity Based on Patient Trajectories

LORENA PUJANTE-OTALORA[1], MANUEL CAMPOS[1,2], AND JOSE M. JUAREZ[1]
[1]Medical Artificial Intelligence Laboratory (MedAILab), University of Murcia, 30100 Murcia, Spain
[2]Murcian Bio-Health Institute, IMIB-Arrixaca, 30120 Murcia, Spain

Corresponding author: Lorena Pujante-Otalora (lorena.pujante@um.es)

This work was supported in part by CONFAINCE project under Grant Ref: PID2021-122194OB-I00 funded by Ministerio de Ciencia, Innovación y Universidades (MICIU)/Agencia Estatal de Investigación (AEI)/10.13039/501100011033 and by European Regional Development Fund (ERDF)/European Union (EU); and in part by Fundación Séneca, Región de Murcia (Spain) under Grant Ref: 21460/FPI/20.

ABSTRACT Healthcare-associated infections are a significant public health concern. This study presents the Spatiotemporal Epidemiological Similarity based on Patient Trajectories (StESPT) method to provide clinicians with support in investigating the epidemiological relationships between infected patients, making detecting outbreaks and identifying possible transmission routes. To this end, the StESPT retrieves the movements of infected patients through the hospital and transforms them into trajectories. Then, a Trajectory Distance Measure Algorithm (TDMA) was used to quantify the spatiotemporal similarity between the trajectories of each pair of patients. Finally, based on the results of each TDMA, the k-means clustering algorithm was used to group patients, which can help understand the spread of the infection. We compared the suitability of three commonly used TDMAs (Dynamic Time Warping (DTW), Spatiotemporal Linear Combine similarity (STLC), and Spatiotemporal Longest Common Subsequence (ST-LCSS)) that were modified to measure similarity instance of distance. For each TDMA, we also proposed a version that adapts better to the semantics of our problem. The StESPT method was tested with a synthetic simulation of *Clostridium difficile* infection in a hospital. The results of the modified version of the ST-LCSS-WTW best reflected the spatiotemporal relationships between patients and led to the k-means clustering revealing two potential outbreaks.

INDEX TERMS Spatiotemporal analysis, healthcare-associated infections, outbreak detection, trajectory similarity.

I. INTRODUCTION

In recent years, healthcare-associated infections (HAI), also known as *nosocomial infections*, have become a major public health concern. It is associated with the global emergence of multidrug-resistant (MDR) bacteria and other pathogens. The difficulty in treating MDR infections leads to an increase in their severity and spread. As a result, HAIs are a major cause of death and increased morbidity among hospitalized patients. Among healthcare facilities, hospitals are the most affected by HAIs [1], [2], [3], [4].

Inside a hospital, infections can be transmitted through both direct and indirect contact. *Direct contact* is defined as physical contact or contact within the threshold distance for transmission by droplets or fomites, for example, through respiratory activity. *Indirect contact* can occur through a chain of people (e.g., healthcare workers (HCW)) or a contaminated environment (droplets and fomites can remain suspended or on surfaces such as clothing, furniture, or medical instruments) [5]. Contact tracing plays a significant role in controlling infections. However, even in hospitals, it is conducted manually, which may delay the time required for halting transmission [6]. In addition, contact tracing establishes transmission routes based on direct contact, so that indirect contact may be missed.

In general, population density and mobility patterns have a major influence on the spread of an outbreak [7]. In the case of hospital environments, the spatial structure of the

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

hospital, physical distributions of patients, and their trajectories (sequence of movements within the hospital) over time are determining factors in the transmission of HAI.

Hospital epidemiologists must study whether the current situation can be characterized as an outbreak when there are a high number of infections in a hospital during a period. First, they must **investigate the epidemiological relationships between infected patients**. Generally, an alert for possible outbreaks is issued when the cumulative number of cases of a bacterium (or other pathogen) reaches a threshold in a spatial location (service, hospitalization unit, or floor) during a period (usually one week) [8], [9]. The outbreak is confirmed when it is proven that there is some epidemiological spatiotemporal relationship between the infected patients. Second, they **detect possible transmission routes** by identifying groups of patients with similar trajectories to control the infection.

Regarding the representation of spatiotemporal information, in [10], we presented a model defined as a knowledge graph (KG) that links the temporal dimension that describes everything that happens to each patient in the form of events (e.g., movements or positive microbiological tests) with the spatial dimension. This last dimension describes the hospital layout with several levels of detail in the form of a hierarchy ranging from bed to building (physical part), and a simple representation of the healthcare workers' organization (logical part). This model allowed us to trace all patient movements during hospitalization. Its data come from the Hospital Information System (HIS), which usually records patients' beds (or rooms) and medical units over time.

In this work, we present the **Spatiotemporal Epidemiological Similarity based on Patient Trajectories (StESPT)** method, which provides clinicians with support in these two tasks. For the first task, we proposed **quantifying how spatiotemporally connected infected patients have been based on their trajectories.** The base assumption of the StESPT is that patients with closer or more prolonged contact (direct or indirect) with an infected patient should, in theory, be at a higher risk of infection [7]. The closeness between the two patients can be interpreted as having a similar trajectory, which can be computed using a Trajectory Distance Measure Algorithm (TDMA). Based on the model defined in [10], we transformed the events registered on the HIS of the stays of the infected patients into trajectories. For the second task, the similarity between the trajectories of each pair of patients can be graphicly represented and used as input for a clustering algorithm to identify the closest patients and, therefore, plausible transmission routes.

Within StESPT, we evaluated three state-of-the-art TDMAs and proposed three extensions based on the formers to determine which best reflects patients' spatial and temporal relationships.

The proposed method is demonstrated by a simulation study in which we use its steps to analyze the spatiotemporal connections between patients infected with *Clostridium difficile*(*CD*) in a hospital over a short period of time. We proved the existence of the outbreak and identified the groups of patients among whom transmission was most likely.

The remainder of this paper is organized as follows: Section II provides background information related to the state-of-the-art (Section II-A), the desirable characteristics for TDMAs to use (Section II-B), and some definitions of preliminary concepts about trajectories (Section II-C). The steps of the StESPT are described in detail in Section III. Sections III-C and III-D are of special interest because they define an equation to evaluate the spatiotemporal epidemiological similarity between points and trajectories, respectively. Section IV presents an experimental evaluation of the StESPT, the results of which are discussed in Section V. Conclusions are summarized in Section VI.

II. BACKGROUND

A. LITERATURE REVIEW

Major advances in sensor technology and devices with GPS or other technologies to track their position have led to large amounts of spatiotemporal data describing the movement of objects and people, known as *trajectories*. Consequently, interest in the analysis of trajectory data has increased in the last few years in several domains that include transportation, behavioral studies, traffic prediction, animal migration, and finding points of interest [11]. Trajectory analysis covers several tasks such as indexing, classification, clustering, data querying, and mining. However, for all of them, computation of the distance or similarity between trajectories is a fundamental component. Thus, it must be carefully defined to adequately capture the data semantics and underlying relations [12]. For this purpose, TDMAs have been widely studied for decades.

In the last few years, with the emergence of coronavirus disease 2019, interest in trajectory analysis focused on epidemiology has increased. Positional data from GPS [7], [13], mobile phone operators [14], and Bluetooth have been used for several tasks, such as contact tracing and the detection of mobility patterns of people in cities or regions. For indoor cases, specifically for hospital environments, patients' movements are also gathered from Electronic Health Records (EHR) and HIS [15], [16], [17], [18], [19] or generated with simulation models [20]. Some of these studies are characterized by creating contact networks between patients, where contact occurs when two patients have been in the same *ward* (understood as hospital pavilions or a set of rooms from the same medical service) [16], [18], [19], [21] or room [20], [22] at some point during the day. Network and statistical analyses were performed on these networks to detect patient clusters. A 24-hour period may be too long for a detailed analysis of a patient's movements because they may be in several rooms throughout the day. In addition, for fine-grained analysis, the *ward* concept may cover a large and highly compartmentalized area, considering that all patients in it are equally close. When the spatial unit is a *room*, including information about the hospital layout or the physical distance between rooms might be helpful in performing a more complete analysis.

Some studies in which patient trajectories were analyzed at a finer grain have been conducted [17], [23]. In [17], a spatiotemporal proximity equation was used to quantify the likelihood of disease transmission between infected inpatients. The proximity is expressed as the inverse of the distance. A trajectory was created for each patient to record the ward in which they had been each day. However, no TDMA is used to guide the computation of the total proximity between two trajectories: they add the proximity from each point on one trajectory to each point on the other trajectory. The spatial distance between wards was based on the proportion of patients who moved from one ward to another over a year [24]. Finally, they used *continuous K-nearest neighbors* to filter the closest patients. In [23], the hourly positions of some patients were tracked using smartphone GPS sensors to create their trajectories. Subsequently, two TDMAs (*Dynamic Time Warping* and *Longest Common Subsequence*) are used to obtain a *similarity matrix* between the patients, which was used as the input of a hierarchical clustering method to detect routine patterns.

B. CRITERIA FOR TDMAS

We propose a detailed analysis based on the similarity between patient trajectories, computed using a TDMA. The studies [11], [12], [25] are three surveys in which some of the most used TDMA are characterized, classified, and compared. We selected the algorithms that met the following criteria:

- The distance measure considers (or can do) spatial and temporal information because both dimensions influence transmission by air and contaminated surfaces.
- They are defined *discretely*. They only consider the sampling points that define the trajectory, not the movement in between, because patients are not in constant motion.
- They are *elastic* measures. They allow the comparison of one-to-many/none points, and not only the i-th point of one trajectory to the i-th point of another (*lock-step* measures). We aim to determine which patients are more likely to transmit the infection, so a restrictive algorithm would not count situations such as indirect contact or when two patients have been close but not at the same time. In addition, the patients may not have shared locations with the rest of them.
- They are *heuristic-based measures*, also called *non-learning*. Their strategy is to identify optimal point matches by comparing pairs of points from two trajectories and computing the total distance by aggregating the average values between all these points [26]. They can easily be adapted to different domains. In contrast, *learning measures* rely on machine-learning techniques to reconstruct high-dimensional input data into a new low-dimensional representation. They are typically developed for a specific domain, the most common of which is the road network.

From the TDMAs that met these features, we chose the following three for comparison: ***Dynamic Time Warping*** (**DTW**), ***Spatiotemporal Longest Common Subsequence*** (**ST-LCSS**), and ***Spatiotemporal Linear Combine***(**STLC**). Some other commonly used TDMA, such as *Edit Distance on Real Sequence, Hausdorf Distance* or *Frechet Distance*, do not meet all these requirements, so they were not selected for this work.

C. PRELIMINARY CONCEPTS

This section introduces the preliminary concepts about trajectories.

Definition 1 (Trajectory): Trajectory T is an ordered sequence of sampling points that represents the path followed by a moving object. In this study, the trajectory T_n describes the movements of patient n through the hospital.

$$T_n = \{p_1, p_2, \ldots, p_m\}, where\ p_i\ (1 \leq i \leq m) \tag{1}$$

$m = |T_n|$ denotes the length of the trajectory, that is, the number of points. We use T_n^i to denote the i-th point in T_n.

Definition 2 (Sampling point): The sampling point p_i (for simplicity, *point*) represents a patient's spatial location at a specific timestamp. In this work, we take as a basis the spatiotemporal model presented in [10] (see Fig. 1). The spatial location is described as a pair with the IDs of two of its concepts: the *Seat* where patients were located and the *HospitalizationUnit* (*HU*) that cared for them.

$$p_i = \left(\left(Bed_{ID}, HospitalizationUnit_{ID}\right), timestamp\right) \tag{2}$$

Definition 3 (Head(T)): It is a function that returns the first sampling point of the trajectory T.

$$Head\ (T) = p_1 \tag{3}$$

Definition 4 (Rest(T)): It is a function that returns the tail sampling points of trajectory T except for the first one.

$$Rest\ (T) = [p_2, p_3, \ldots, p_m] \tag{4}$$

Definition 5 (Trajectory distance measure algorithm (TDMA):) A TDMA is a function $d\ (T_1, T_2) \in \mathbb{R}_0^+$ that evaluates the distance between two trajectories T_1 and T_2 based on the distance between their sampling points [11].

III. THE STESPT METHOD

This section presents the steps that form the StESPT method for epidemiological research. First, we examine patients infected with a specific bacterial species at a particular time. Then, their stays in the hospital are transformed into trajectories. The following steps consist of defining the spatiotemporal epidemiological similarity between two sampling points and using a TDMA to calculate the similarity between the trajectories of each pair of patients. Using this result, we create a *similarity matrix* that is represented as a heatmap. Finally, we apply k-means clustering to the similarities between patients to detect groups of spatiotemporally close patients.

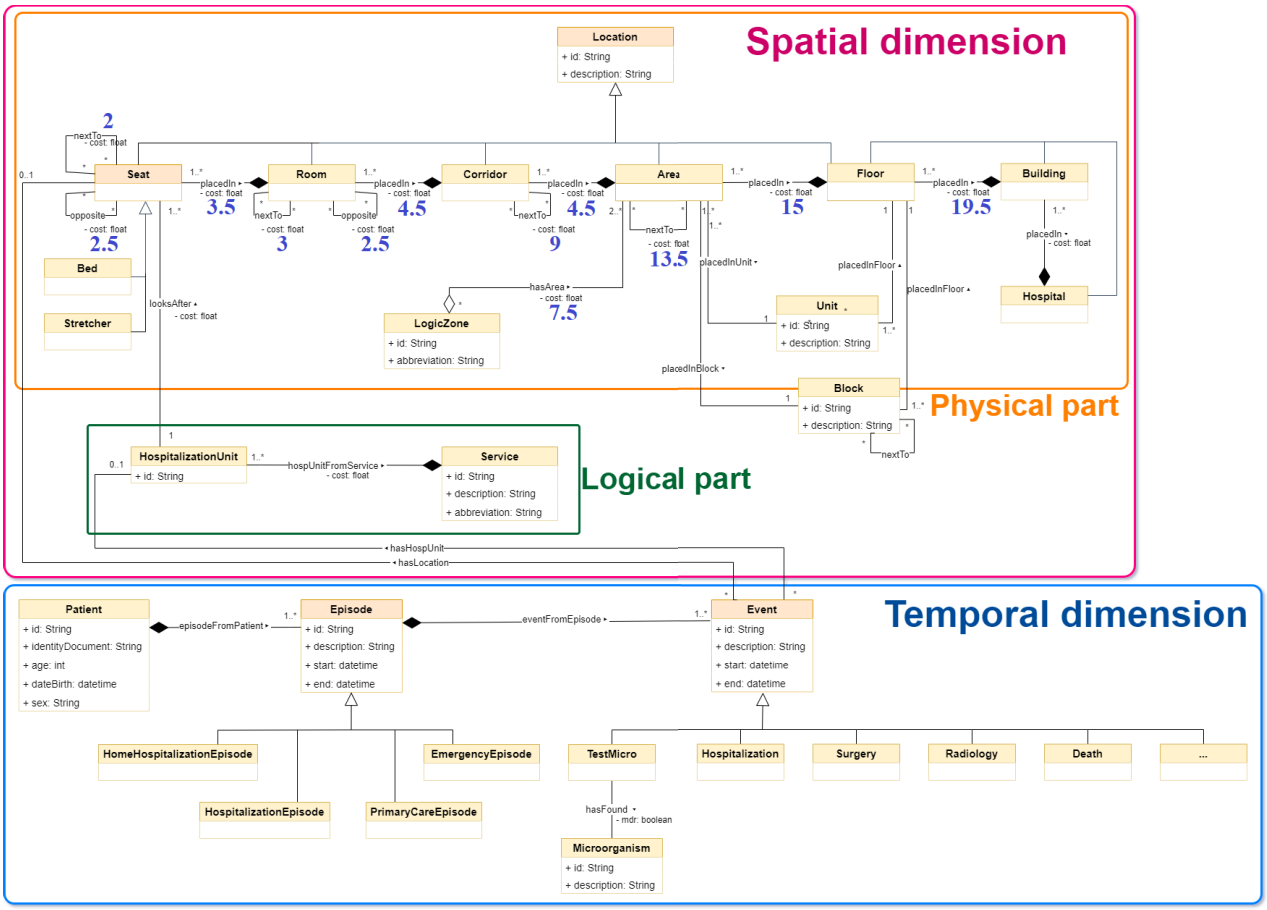


FIGURE 1. UML Class diagram of the spatiotemporal epidemiological model. In blue, the values for the property cost of each edge that we used in the experiment to calculate the spatial distance and similarity between two beds. These values have been calculated based on the dimensions of a hospital bed and a generic room with two beds.

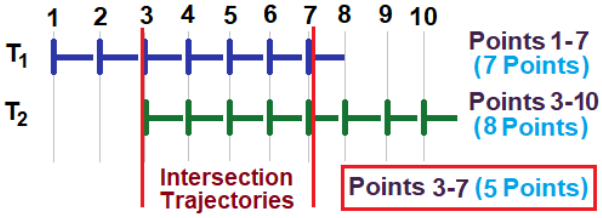


FIGURE 2. Temporal representation of the intersection trajectories between two trajectories.

A. STEP 1: GET INFECTED PATIENTS

The first step of StESPT is to retrieve patients who can be part of a possible outbreak. For this task, we used one of the epidemiological queries defined in [10], specifically *Q2—Detection of an outbreak in a Location*. This query retrieved those patients who were on a determined floor and had a positive microbiological test of a specific bacterium for a certain period, regardless of whether this positive test occurred while the patient was on the floor. Note that a positive test indicates that we know that the patient is infected from the moment we obtain the result; however, this time does not have to coincide with when they become infected. There are infections with an incubation time of several days, leading to patients being infected in one location of the hospital and having a positive result in another.

B. STEP 2: TURN PATIENTS' STAYS INTO TRAJECTORIES

The second step was to retrieve the trajectories of patients from the first step. We propose defining a *Search Time for the Patients' Trajectories* (*STPT*), which should not be the same as the search time for infected patients. The end time should be the date of the last positive test because, after this, all patients are already infected, so contact between them will not affect the infection. The start time of STPT should be before the start time to search for infected patients. Considering that a positive test does not represent the exact time of the infection, for the start time, we added a margin of several days back to the date of the first positive test. For example, we can add the median incubation time of the bacterium according to scientific literature.

We look for all *Events* with a connected *Location* that occur during the *STPT*. These events are converted into a trajectory of sample points using a sampling rate (SR) such that the points on all trajectories are time-aligned (the first and last points on each trajectory may not coincide). Each point is defined as in Definition 2.

When comparing pairwise trajectories to calculate their similarity, we select the intersection points between the two trajectories. That is, only the time that both patients have been in the hospital will be evaluated, not the time that only one patient has been hospitalized. To simplify, we refer to these temporarily intersecting paths as *intersection paths (IT)*. Fig. 2 shows two trajectories (ignoring spatial information) that temporally coincide from points 3 to 7. When calculating the similarity between the patients, only the sub-trajectories between these points (marked in red) are compared.

C. STEP 3: SPATIOTEMPORAL EPIDEMIOLOGICAL SIMILARITY BETWEEN POINTS

We define the spatiotemporal distance between two sampling points as a weighted and linear combination of the spatial distance, d_{sp}, and the temporal distance, d_{temp} [17], [26], [27].

$$d_{ST}(p_1, p_2) = \alpha \cdot d_{sp}(p_1, p_2) + (1 - \alpha) \cdot d_{temp}(p_1, p_2) \tag{5}$$

$\alpha \in [0,1]$ can be modified depending on the disease and the spatial and temporal characteristics of the outbreak; therefore, we can adjust the relative importance of space and time in the spatiotemporal distance.

However, we aim to measure how close the trajectories are, not their distance: the shorter the distance between them, the more similar they are. Therefore, similarity can be interpreted as the opposite of distance. Thus, we needed to define the spatiotemporal similarity between sampling points.

1) SPATIAL DISTANCE

The spatial dimension of the sampling points is generally defined using two-dimensional or three-dimensional coordinates. Their distance is usually measured using the Euclidean distance. We use a ''qualitative distance'' based on the spatial hierarchy of *Locations* from the KG in [10]. Given two *Seats*, their spatial distance is defined as the sum of the ''*cost*'' property of the edges of the *physical spatial dimension* that form the shortest path between them (see Fig. 1).

A matrix stores the distance between each pair of hospital *Beds*. The matrix values will be normalized to the interval [0,1] to make the spatial and temporal distances comparable.

HCW may also represent a means of transmitting the infection: patients cared for by the same HCW are more likely to be infected when there is an infection in the *HU* or *Service*. To represent this situation, the spatial distance between two sampling points is reduced by multiplying it by a value between (0,1). In our experiments, this value will be 0.,5 when patients are cared for by the same *HU,* and 0.7, when by the same *Service*.

2) TEMPORAL DISTANCE

The temporal distance between two sampling points was defined based on two aspects. The first component is the absolute value of the difference between the timestamps ($diff_{tmp}$). Similar to the spatial distance, the difference is normalized to [0,1]. The maximum value is the difference between the timestamps that define STPT.

The second component is the transmissibility (or transmission speed) of the infectious disease. Direct and indirect contact is more likely to result in a new infected patient when the disease transmissibility is higher. Hence, it can be used to determine whether two points are temporally close. To represent this relationship between the temporal difference and transmissibility, we multiplied the temporal difference by the *transmission rate* (β) [28] as the transmissibility parameter.

$$diff_{tmp}(p_1, p_2) = \left| p1_{time} - p2_{time} \right| \tag{6}$$

$$d_{tmp}(p_1, p_2) = \beta \times diff_{tmp}(p_1, p_2) \tag{7}$$

3) SPATIOTEMPORAL EPIDEMIOLOGICAL SIMILARITY

Up to this point, we have defined the spatial and temporal distances separately; therefore, it is possible to calculate the spatiotemporal distance between two points using (5). Similarity can be interpreted as the inverse of distance, so we inverted the equations to calculate the spatial and temporal distances as linearly as possible. When the points have the same *Location*, the spatial similarity must be zero; and when they have the same timestamp, the temporal similarity must be zero.

In the case of temporal distance, β must also be inverted: high values of β would return a larger temporal distance and, thus, a lower probability of infection.

The definitions of spatial and temporal similarities between sampling points are the following:

$$sim_{sp}(p_1, p_2) = 1 - d_{sp}(p_1, p_2) \tag{8}$$

$$sim_{tmp}(p_1, p_2) = e^{\ln(\beta) \times \left| p1_{time} - p2_{time} \right|} \tag{9}$$

Based on (5), spatiotemporal epidemiological similarity is defined as a weighted and linear combination of (8) and (9) with $\alpha \in [0,1]$.

$$sim_{ST}(p_1, p_2) = \alpha \cdot sim_{sp}(p_1, p_2) + (1 - \alpha) \cdot sim_{tmp}(p_1, p_2) \tag{10}$$

D. STEP 4: TDMA

To evaluate the closeness between two patients, we studied three TDMAs. We modified them slightly so that they calculate similarity instance of distance. In addition, we designed an extension of each algorithm to adapt better to the semantic requirements of our problem. In total, six TDMA were compared.

1) DTW

Dynamic Time Warping(*DTW*) [29] is an algorithm designed for time series that can also be applied to trajectories. DTW is defined recursively to search through all possible point alignments between two trajectories, T_1 and T_2, where the sum of the spatial distances between points is minimal and meets the following requirements:

- T_1^1 at least matches T_2^1.

- $sim_{ST\_DTW}(T_1, T_2) = ST\_DTW(T_1, T_2)^1$
- $d_{ST\_DTW}(T_1, T_2) = ST\_DTW(T_1, T_2)^2$

where

- $ST\_DTW(T_1, T_2) = \begin{cases} (0,0), & \text{if } |T_1| = 0 \text{ and } |T_2| = 0 \\ (0,\infty) & \text{if } |T_1| = 0 \text{ or } |T_2| = 0 \\ \begin{pmatrix} sim_{ST}(Head(T_1), Head(T_2)) + A^1 \\ d_{ST}(Head(T_1), Head(T_2)) + A^2 \end{pmatrix} & otherwise \end{cases}$

where

- $A = select\ couple\ from\ B\ where\ B^2\ is\ minimum,$
- $B = \{ST\_DTW(T_1, Rest(T_2)) \cup ST\_DTW(Rest(T_1), T_2) \cup ST\_DTW(Rest(T_1), Rest(T_2))\}$

FIGURE 3. Definition of ST-DTW. It returns a couple, where the first element is the spatiotemporal similarity between T_1 and T_2, and the second is the distance. Superscripts represent the position of an element within a tuple. In red, the functions to calculate the spatiotemporal similarity and distance between two points. In DTW, they are changed for sim_{sp} and d_{sp}.

- T_1^m at least matches T_2^m.
- Each intermediate point of each trajectory matches one or more points of the other, without returning.

Thus, DTW returns the minimum difference between the two trajectories. Our version of DTW calculates both the distance and the similarity between points, where distance is used, as in its original definition, to determine the optimal path. Another change is that the similarity will be zero when the length of one or both trajectories is zero: if any alignment can be made, no similarity should be added to the total.

It is noteworthy that DTW only evaluates the spatial dimension; therefore, it uses the d_{sp} and sim_{sp}.

2) ST-DTW

We propose a new *Spatiotemporal DTW* (*ST-DTW*) that extends DTW to evaluate the time and infection transmissibility. The change in *ST-DTW* is that it uses the definitions of spatiotemporal distance and similarity defined in (5) and (10), respectively. Fig. 3 shows the definition of the ST-DTW.

3) STLC

Spatiotemporal Linear Combine similarity(*STLC*) [11], [30] proposes a linear combination of spatial and temporal similarities between two trajectories to obtain their spatiotemporal similarity. Spatial and temporal similarities are computed separately. To explain STLC, we focus on the spatial dimension. It calculates the *point-to-trajectory* spatial distance for each point of each trajectory, that is, the minimum spatial or temporal distance between that point and a point of the other trajectory. These distances are transformed into similarities and the average similarity is calculated for each trajectory. Then, they are added. Temporal similarity is calculated using the same process. Finally, the spatial and temporal similarities are combined linearly using the parameter $\alpha \in [0,1]$. Fig. 4(a) shows the original definition of the STLC.

We adapted STLC to use the spatial and temporal similarity definitions (8) and (9). Fig. 4(b) shows our adaptation of the STLC. Note that selecting the point with the minimum distance and transforming it into similarity is equivalent to choosing the point with the maximal similarity.

$$sim_{STLC}(T_1, T_2) = \alpha \cdot sim_{sp}(T_1, T_2) + (1-\alpha) \cdot sim_{tmp}(T_1, T_2)$$

where:

- $sim_{sp}(T_1, T_2) = \frac{\sum_{T_1^i \in T_1} e^{-d_{sp}(T_1^i, T_2)}}{|T_1|} + \frac{\sum_{T_2^j \in T_2} e^{-d_{sp}(T_2^j, T_1)}}{|T_2|}$
  - $d_{sp}(p, T) = min_{q \in T}\{d_{sp}(p, q)\}$
- $sim_{tmp}(T_1, T_2) = \frac{\sum_{T_1^i \in T_1} e^{-d_{tmp}(T_1^i, T_2)}}{|T_1|} + \frac{\sum_{T_2^j \in T_2} e^{-d_{tmp}(T_2^j, T_1)}}{|T_2|}$
  - $d_{tmp}(p, T) = min_{q \in T}\{d_{tmp}(p, q)\}$

(a)

$$sim_{STLC}(T_1, T_2) = \alpha \cdot sim_{STLC_{sp}}(T_1, T_2) + (1-\alpha) \cdot sim_{STLC_{tmp}}(T_1, T_2)$$

where:

- $sim_{STLC_{sp}}(T_1, T_2) = \frac{\sum_{T_1^i \in T_1} sim_{sp}(T_1^i, T_2)}{\lambda} + \frac{\sum_{T_2^j \in T_2} sim_{sp}(T_2^j, T_1)}{\lambda}$
  - $sim_{sp}(p, T) = max_{q \in T}\{sim_{sp}(p, q)\}$
- $sim_{STLC_{tmp}}(T_1, T_2) = \frac{\sum_{T_1^i \in T_1} sim_{tmp}(T_1^i, T_2)}{\lambda} + \frac{\sum_{T_2^j \in T_2} sim_{tmp}(T_2^j, T_1)}{\lambda}$
  - $sim_{tmp}(p, T) = max_{q \in T}\{sim_{tmp}(p, q)\}$

(b)

FIGURE 4. Definition of STLC. (a) Original STLC. The equations to transform distances into similarities are marked in green (spatial) and blue (temporal). (b) Our adaptation of STLC. Changes are in orange and purple.

$$sim_{JSTLC}(T_1, T_2) = \frac{\sum_{T_1^i \in T_1} sim_{ST}(T_1^i, T_2)}{\lambda} + \frac{\sum_{T_2^j \in T_2} sim_{ST}(T_2^j, T_1)}{\lambda}$$

where:

- $sim_{ST}(p, T) = max_{q \in T}\{sim_{ST}(p, q)\}$

FIGURE 5. Definition of JSTLC.

The original STLC algorithm calculates the average *point-to-trajectory* similarity for each trajectory and dimension. Thus, it does not consider the proportion of time that patients have been in the hospital compared with the entire STPT. In the most extreme case, if two patients have been in the closest possible *Bed* for their entire *intersection trajectories*, their similarity will be the maximum regardless of the length of the ITs. We changed the denominator to obtain the average similarities for parameter λ, which is the difference between the points defining STPT.

4) JSTLC

We propose an extension of STLC called *Joint Spatiotemporal Linear Combine similarity* (*JSTLC*). It directly calculates the spatiotemporal point-to-trajectory similarity instead of separately calculating the spatial and temporal similarities. With this modification, we want to overcome the fact that the spatially closest point may not be the temporally closest point. Fig. 5 shows the definition of the JSTLC.

5) ST-LCSS

The *Spatio-Temporal Longest Common Subsequence*(*ST-LCSS*) [11] returns the size of the longest common subsequence between two trajectories, T_1 and T_2. This subsequence includes all matches between two points of each trajectory, but with some conditions: a sampling point cannot match with more than one point, the pairs of matched points do not have to be consecutives, but they must follow the order of the trajectories (if T_1^i and T_2^j match, T_1^{i+k} cannot match with T_2^{j-l}), and there can be points without a match.

A strict definition of a match is that two sampling points share the same values for spatial and temporal dimensions. However, it can be difficult to find any two points with the exact time and location. Thus, the ST-LCSS adds two parameters, ε and δ, to define the maximum distance in space and time to limit the search for matching two points.

The original ST-LCSS already measures the similarity between trajectories (the size of the LCSS), so it has barely been modified. In our version of the algorithm, ε and δ represent the minimum spatial and temporal similarities between two sampling points to match, and not the maximum distance.

The most significant change is that the result of our ST-LCSS is both the length of the LCSS and the sum of the spatiotemporal similarities between each pair of matching points. The solution of the ST-LCSS will be the combination with the highest number of matches, and, in the case of a tie, the solution with the highest spatiotemporal similarity will be selected. Fig. 7 shows the definition of ST-LCSS.

6) ST-LCSS-WTW

We designed a variation of ST-LCSS, called *ST-LCSS With Time Window*(*ST-LCSS-WTW*), to improve its performance in the measurement of similarities between the trajectories of infected patients. The ST-LCSS follows a greedy scheme and does not explore the entire space of solutions: when it finds a match, it does not reconsider its choices in future steps. Thus, an optimal solution may not be found. For example, two points that do not have the highest similarity may match leading to a result with a lower number of matches or similarity. In addition, the similarity between T_1^i and T_2^{j+1} may be higher than that of the chosen match with T_2^j. We aim to identify as many connections as possible between each pair of patients that could have contributed to the transmission of the infection. We propose the addition of a time window ω and explore several possible solutions for any given step. Given two sampling points T_1^i and T_2^j, we evaluate the matches between each point and any other point in the other trajectory within the interval $[T_2^{j-\omega}, T_2^{j+\omega}]$ or $[T_1^{i-\omega}, T_1^{i+\omega}]$, respectively. For each match, ST-LCSS-WTW is called recursively. From all the available solutions, the algorithm returns the one with the largest number of matches.

ω is defined in terms of the sampling points. We propose that the time it represents be smaller than the value of $diff_{temp}$ used to calculate δ. Otherwise, there would always be points to evaluate, whose temporal similarity would never be within the defined range.

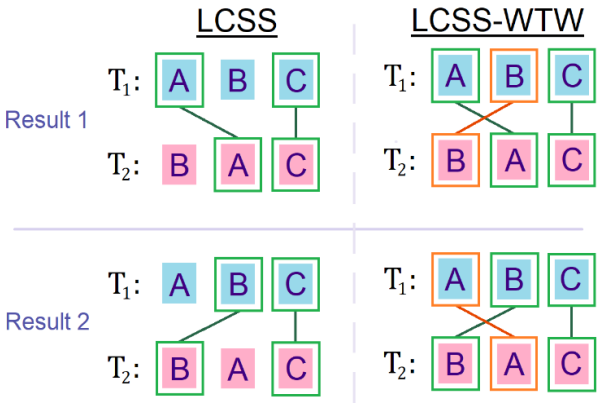


FIGURE 6. Example of executing ST-LCSS and ST-LCSS-WTW over a same pair of trajectories.

Fig. 8 shows the definition of the ST-LCSS-WTW. A representative example of the difference between the ST-LCSS and ST-LCSS-WTW is shown in Fig. 6, where T_1 and T_2 are represented as strings. A match occurs when two points have the same letter, and there is a maximum difference in one position between the two points. In the ST-LCSS-WTW, ω is 1. T_1 and T_2 represent two strings with the same letters in different orders. When executing the ST-LCSS, it returns two matches: when it matches *A*s (or *B*s), the first *B*(or *A*) can no longer be matched, and thus the only possible match is between *C*s. However, in ST-LCSS-WTW, both pairs of *A*s and *B*s can match such that the result is three matches.

7) RESULTS OF THE TDMA AS A SIMILARITY MATRIX

As a result of the execution of each TDMA for each pair of patients, we obtained a squared matrix $M_{p\times p}$ called *similarity matrix*(*SM*), where p is the number of infected patients and $M_{i,j}$ represents the similarity between the intersection trajectories of patients i and j. Thus, it summarizes the spatiotemporal similarity among patient trajectories, and each row (or column) is a *similarity vector*(*SV*) of size p with the similarities between a patient and the others (including themselves).

Results from DTW, ST-DTW, ST-LCSS and ST-LCSS-WTW range $[0, |STPT| \times \Upsilon]$, where $|STPT|$ is the number of points of STPT when sampled and Υ is the maximum possible similarity between two sampling points (two neighboring beds at the same timestamp, or the same bed at two consecutive timestamps). However, results from the STLC and JSTLC range [0,2] when spatial and temporal distances are in the range of [0,1], as in this work. To compare the SMs from the six TDMA, we normalized the results to [0,1]. These normalized values can be represented as a heatmap, supporting the visual detection of clusters. For example, a row (or column) with several warm cells represents a highly interconnected patient, and several warm cells that share the same set of patients can represent a cluster of close patients. Note

- $sim_{ST\_LCSS}(T_1, T_2) = ST\_LCSS(T_1, T_2)^1$
- $|LCSS|(T_1, T_2) = ST\_LCSS(T_1, T_2)^2$

where

$$\circ\ ST\_LCSS(T_1, T_2) = \begin{cases} (0,0), & \text{if } |T_1| = 0 \text{ or } |T_2| = 0 \\ \begin{pmatrix} ST\_LCSS\big(Rest(T_1), Rest(T_2)\big)^1 + sim_{ST}\big(Head(T_1), Head(T_2)\big), \\ ST\_LCSS\big(Rest(T_1), Rest(T_2)\big)^2 + 1 \end{pmatrix}, & \text{if } sim_{sp}\big(Head(T_1), Head(T_2)\big) \geq \varepsilon \\ & \text{and } sim_{temp}\big(Head(T_1), Head(T_2)\big) \geq \delta \\ A & otherwise \end{cases}$$

where

- $A = select\ couple\ from\ B\ where\ B^2 is\ maximum,$
- $B = \{ST\_LCSS\big(T_1, Rest(T_2)\big) \cup T\_LCSS(Rest(T_1), T_2)\}$

FIGURE 7. Definition of our adaption ST-LCSS. It returns a couple, where the first element is the spatiotemporal similarity between T_1 and T_2, and the second is the length of the LCSS. Superscripts represent the position of an element within a tuple.

- $sim_{ST\_LCSS\_WTW}(T_1, T_2, i, j, U) = ST\_LCSS\_WTW(T_1, T_2, i, j, U)^1$
- $|LCSS|(T_1, T_2, i, j, U) = ST\_LCSS\_WTW(T_1, T_2, i, j, U)^2$

where

$$\circ\ ST\_LCSS\_WTW(T_1, T_2, i, j, U) = \begin{cases} (0,0), & \text{if } i \geq |T_1| \text{ or } j \geq |T_2| \\ \begin{pmatrix} sim_{ST}\big(T_1^i, T_2^j\big) + A^1, \\ 1 + A^2 \end{pmatrix}, & \text{if } |B| > 0 \\ C & otherwise \end{cases}$$

where

- $A = select\ couple\ from\ B\ where\ B^2\ is\ maximum$
- $B = \{ST\_LCSS\_WTW\big(T_1, T_2, i+1, j+1, U \cup \{T_1^i, T_2^j\}\big) \mid sim_{sp}\big(T_1^i, T_2^j\big) \geq \varepsilon, sim_{temp}\big(T_1^i, T_2^j\big) \geq \delta, T_1^i \notin U, T_2^j \notin U\}$
  $\cup\ \{ST\_LCSS\_WTW\big(T_1, T_2, i, j+1, U \cup \{T_1^{i+x}, T_2^j\}\big) \mid -\omega \leq x \leq \omega, i + x > 0, sim_{sp}\big(T_1^{i+x}, T_2^j\big) \geq \varepsilon, sim_{temp}\big(T_1^{i+x}, T_2^j\big) \geq \delta, T_1^{i+x} \notin U, T_2^j \notin U\}$
  $\cup\ \{ST\_LCSS\_WTW\big(T_1, T_2, i+1, j, U \cup \{T_1^i, T_2^{j+x}\}\big) \mid -\omega \leq x \leq \omega, j + x > 0, sim_{sp}\big(T_1^i, T_2^{j+x}\big) \geq \varepsilon, sim_{temp}\big(T_1^i, T_2^{j+x}\big) \geq \delta, T_1^i \notin U, T_2^{j+x} \notin U\}$
- $C = select\ couple\ from\ D\ where\ D^2\ is\ maximum,$
- $D = \{ST\_LCSS\_WTW(T_1, T_2, i, j+1, U\}) \cup ST\_LCSS\_WTW(T_1, T_2, i+1, j, U)\}$

FIGURE 8. Definition of ST-LCSS-WTW. It returns a couple, where the first element is the spatiotemporal similarity between T_1 and T_2, and the second is the length of the LCSS. Superscripts represent the position of an element within a tuple. ω is the time window expressed as sampling points, i and j are the indices to traverse T_1 and T_2, and U is the set of sampling points that already have a match. For a better reading, some parts have been marked.

that the similarity between two patients is not considered a transmission probability. It is a numerical representation of the proximity in space and time of two patients when compared to others.

E. STEP 5: CLUSTERING METHOD

We used a clustering algorithm to group the patients based on their similarity to each other, so that close patients were in the same cluster. These clusters could represent different outbreaks or patients whose shared *Locations* or *HU* must be investigated as transmission foci. We clustered the results for each TDMA. For this task, we used the k-means algorithm, whose input was the similarity vectors (SV) for each patient. K-means clustering is a widely used partitional clustering algorithm. Using an iterative process, a dataset is decomposed into K disjoint partitions (*clusters*), represented by a *centroid*(the mean vector of all points assigned to a given cluster). It aims to minimize each point's distance (Euclidean distance) to its cluster centroid [31], [32]. As a result, SVs (and their patients) are grouped into several clusters.

K-means requires K to be specified, so we searched for the optimal value for each TDMA output. For this purpose, we executed the *Silhouette method*: we executed k-means for each TDMA output with different values of K ranging *[2, p]* and selected the one whose *silhouette coefficient*(SC) was the highest. SC is a commonly used cluster validation measure that combines *cohesion* (how close the points of the same cluster are to each other) and *separation* (how far apart the points of different clusters are). In particular, the *separation* is calculated based on the average distance between a cluster's points and that of the closest cluster. SC is defined in the interval $[-1,1]$, where positive values indicate a high separation between clusters (points are '*well-clustered*'), negative values imply overlapping clusters and values close to 0 indicate that data are uniformly distributed, so it is difficult to determine to which cluster each point belongs [33], [34].

IV. EXPERIMENT

The suitability of our method was proven through an experiment to analyze a *Clostridium difficile* (CD) outbreak in a hospital. We compared the six proposed TDMAs to determine which could best capture the semantics and relationships among infected patients.

CD is currently one of the most significant HAI and the main cause of infectious diarrhea in hospitalized patients.

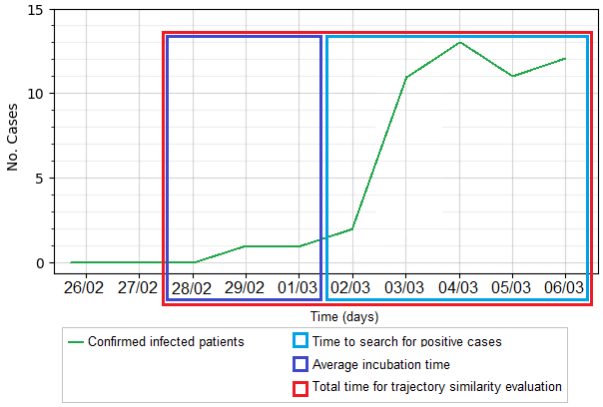


FIGURE 9. Number of infected cases of CD per day. When the number decreases it might be because the recovery or the death of the patients. In blue, the time for which we looked for positive cases. In purple, the extra time we added to search for the patients' events. STPT is in red.

In recent years, its incidence and severity have increased in North America and Europe [35], [36]. CD is a spore-forming bacterium that can be transmitted from person to person or through contact with contaminated surfaces. Room decontamination and hand hygiene of HCW are some of the most followed strategies to reduce its transmission, although it has proven to be a challenge to achieve [35]. Therefore, direct and indirect contact between patients may play a crucial role in the propagation of CD.

A. DATASET

We used a realistic simulation model of HAI [37] to generate patient flow through hospital beds and services. The simulation model had a discrete time with steps of 8 h, representing three work shifts: morning (8:00-15:59), afternoon (16:00-23:59) and night (00:00-7:59). Several patients were infected with CD during the simulation. Our dataset was generated using the same values as in the experiment in [37] for the parameters that describe the epidemiological behavior of the infection.

We generated a simulation for the first three months of the year 2024. From February 28 to March 6, the evolution of CD cases might represent an outbreak. Fig. 9 shows the number of patients confirmed to be infected daily during this period.

We designed a hospital layout based on a large hospital according to the Spanish standards for an influence area of 250,000 people. Our hospital had 729 beds distributed between four floors. *Floor 0* had four Accident and Emergency (A&E) rooms with five beds each, four Intensive Care (IC) rooms with 4 four beds each, 24 radiology rooms, and 27 operating theatres. The other three floors were used for hospitalization. *Floors 1* and *2* had 13 hospitalization units (HU) of seven services each, and *Floor 3* had 14 HUs of three services each. Each HU had seven to ten rooms, and each room had two beds. In total, there were 17 services for hospitalization, one for A&E, one for IC, and one for radiology. All services except *Radiology* had one HU for surgeries.

B. RESULTS

In the following sections, we present the results for each step of the StESPT method.

1) STEP 1: GET INFECTED PATIENTS

As seen in Fig. 9, from March 2 to March 4, there was a significant increase in up to 13 CD cases, which remained stable for the next two days. Most of these patients were on *Floor 2* at some time during these days, so they might be spatiotemporally related. We looked for patients who had positive tests from **March 2 to March 6** and spent some time on ***Floor 2***, obtaining **17 patients**.

2) STEP 2: GET PATIENTS' TRAJECTORIES

The first and last positive tests of the 17 patients from *step 1* were on March 2 and March 6, respectively. The CD incubation period is typically less than one week, and some studies have determined that infection can occur within 48-72 hours after exposure [38], [39]. We will add three days before March 2 to obtain the events of the 17 patients. Therefore, the **STPT** will be conducted from **February 28 to March 6**.

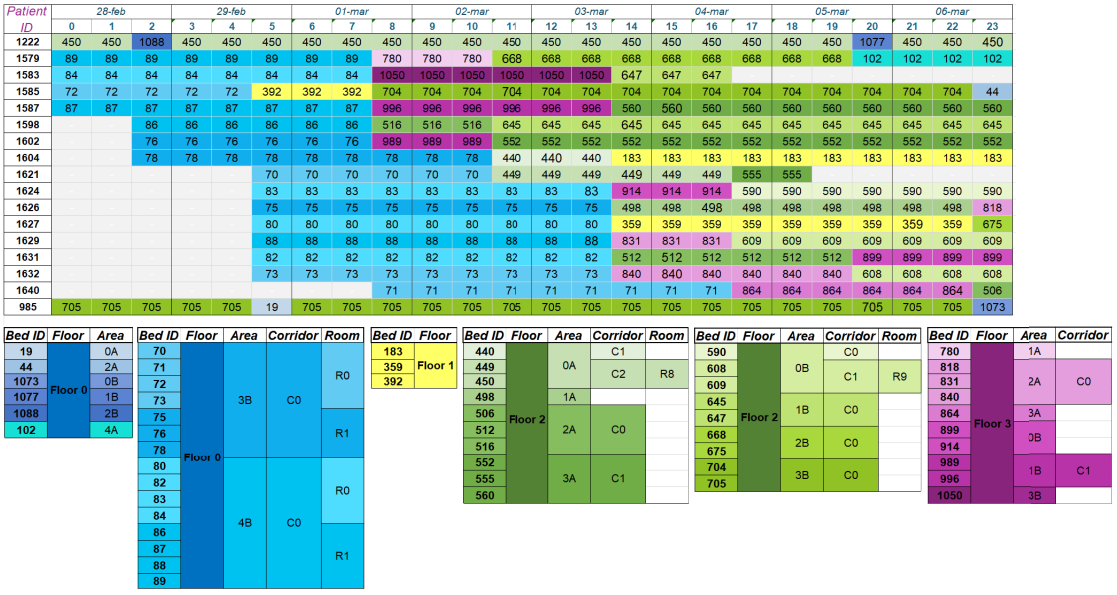
The simulation model has a precision of 8 h; therefore, it was the sampling rate when transforming events into trajectories. We evaluated the intersection trajectories for a maximum of 8 days (**24 sampling points**). Fig. 10 shows the trajectories of each patient, distinguishing between the beds where they were hospitalized (Fig. 10(a)) and the HU that cared for them (Fig. 10(b)).

We can see that patients have different start and end dates of their trajectories, and that some patients are not in nearby locations at close times. This scene leads to the evaluation of common trajectories of different lengths (the shortest has nine steps, and the longest has 24) and diverse spatiotemporal circumstances. However, almost all the patients had two characteristics. First, their trajectories begin on *Floor 0*, specifically in the two adjacent areas (*3B* and *4B*) dedicated to the A&E service. Second, they had at least one point on *Floor 2*. Several patients coincided in the same or adjacent *Areas*. Additionally, some Services and HUs are shared by several patients, such as *Service 7* and *HU 8B*.

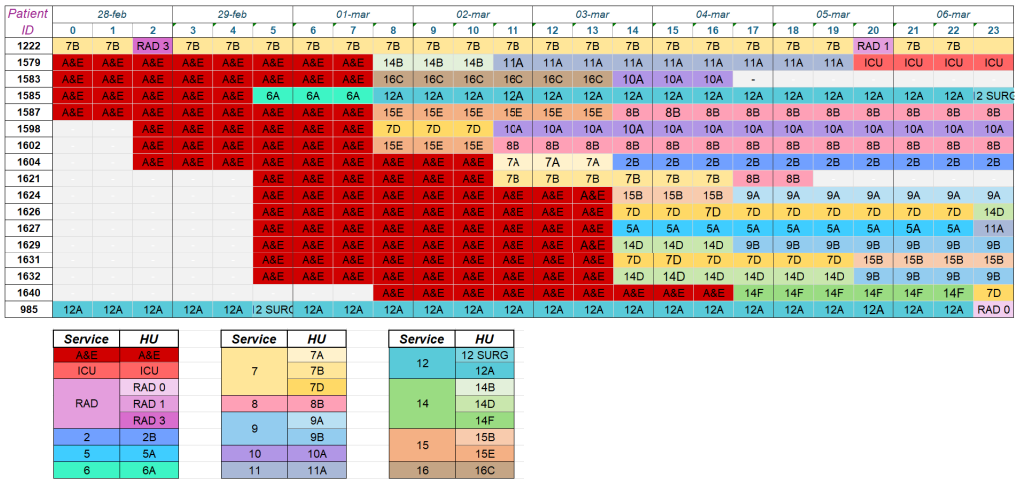
3) STEPS 3 AND 4: SPATIOTEMPORAL SIMILARITY BETWEEN TRAJECTORIES

In this section, the similarity matrix (SM) from executing each TDMA is presented as grid-normalized heatmaps (see Fig. 11), where low similarity is represented in dark blue and higher similarities are in orange-red tones.

When comparing the results from DTW (Fig. 11(a)) and ST-DTW (Fig. 11(b)), they initially appear to be completely different: in the DTW heatmap, most values are below 0.25, with some groups of patients having values around [0.45, 0.65], and in the ST-DTW heatmap, there is an overall increase in similarity of approximately 0.15-0.3, with most values above 0.4. However, both heatmaps exhibited similar patterns. *Patients 1632* and *1579* had medium-high values



(a)



(b)

FIGURE 10. Representation of the trajectories of the 17 CD-infected patients. (a) shows the ID of the Bed where the patient has been at each point. Cells of the similar colours represent Beds that are in the same Room, Corridor, Area or Floor. For example, *Beds 71* and *72* are in the *Area 3B* from Corridor C0 in *Area 3B* from *Floor 0*. (b) shows the name of HU that cared for the patient. Cells of similar colours represent HU from the same Service.

(greater than 0.45), with more than half of the patients in both heatmaps, adding *patients 1222*, *1626,* and *1631* to this list of highly interconnected patients in the heatmap of ST-DTW. These patients presented some of the highest similarity values in the heatmap.

For the ST-LCSS and ST-LCSS-WTW, we evaluated both algorithms using several input parameters values. Specifically, we tested that the maximum temporal difference between two points was 2, 5, and 8 points (they were in the same 24, 48, and 72 h, respectively), and that the maximum spatial distance was between beds in the same *Area* or *Floor.* For the ST-LCSS-WTW, we considered time windows of 2, 5, and 8 points. The results of all the tested combinations were very similar. The most significant changes (on the order of hundredths) occurred when allowing matches on the **same*Floor*within 48 h** and, for ST-LCSS-WTW, a **time window of 5 points**. This section presents the results for these values (Figs. 11(e) and 11(f) also show the heatmaps using these values). Using higher values requires a considerably higher level of computation whereas providing practically identical results.

ST-LCSS and ST-LCSS-WTW presented similar results: most cells had values in the range [0.2, 0.4], with two groups whose values were in the range [0.5, 0.65]. The difference is that the highest similarities have an increase of approximately 0.1 points in ST-LCSS-WTW. It is worth mentioning that most of the intersection trajectories have a length between 12 and 15 points, and that STPT has

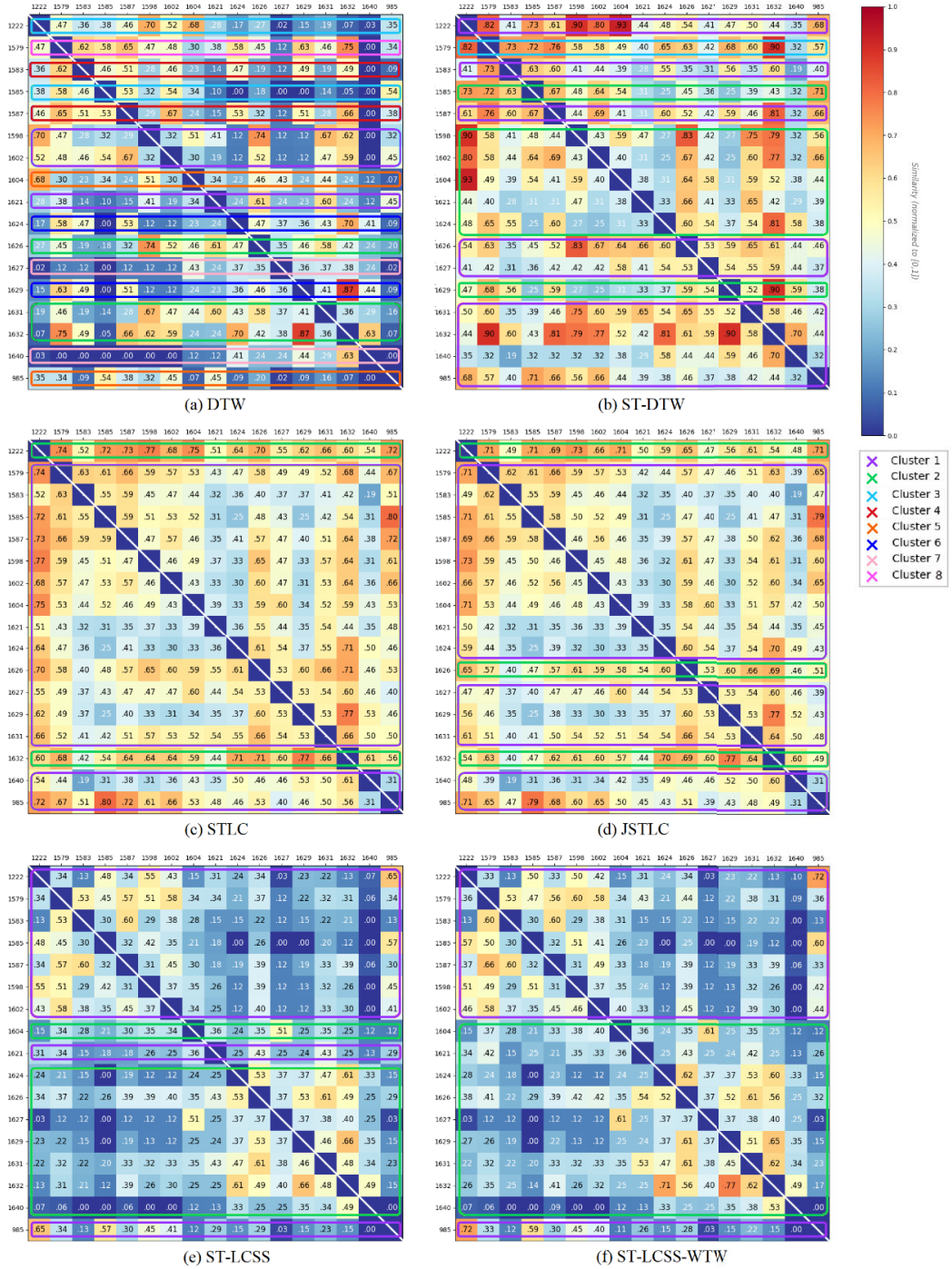


FIGURE 11. Heatmaps with spatiotemporal trajectory similarity between the 17 patients infected with CD. Each map shows the results of one of the TDMAs presented in this work. Patients' IDs are shown as the indexes of the heatmaps and arranged alphabetically. There is an overlapped layer to show the clusters identified with the k-means algorithm, so that the rows of the patients that belong to the same cluster are highlighted in the same color.

24 points. If we observe the heatmap by rows (or columns), we can see that the range of values of the similarities between each patient and the rest is narrow and constant. However, when patients show medium-high similarities, we can identify the group of patients to whom they are close and those to whom they are not. It is different in DTW and ST-DTW heatmaps, where each patient's rows (or columns) have different ranges and variances, making visual analysis difficult.

The results from STLC and JSTLC were similar, with slightly lower values for JSTLC than for STLC. In both heatmaps, most values were medium-high, approximately [0.4, 0.6], and did not present remarkable changes in either the heatmap as a whole or in rows. However, we can see

some patients with overall similarities that are higher (such as patients *985*, *1222,* and *1632*) or lower (such as patients *1621*, *1624*, *1629*, and *1640*) than the others. Some of these patients had the longest and shortest trajectories.

Heatmaps can be helpful tools for performing a preliminary analysis of how patients are spatiotemporally related. This visual interpretation can be affected by how the rows and columns of the heat map are arranged, and a representation of patient trajectories, such as the timelines shown in Fig. 10, should be used as a support. Below, we present an overall interpretation based on the patients with the highest similarities. All patients started their trajectories in beds from the A&E Service on *Floor 0*, which could be the origin of the infection. Subsequently, two outbreaks appeared: a group with *patients* from *1222* to *1602*and *patient 985*, and another group with *patients* from *1624* to *1640*. The first group included patients who first entered the hospital, with their trajectories largely overlapping when they were on *Floor 0* and some adjacent *Areas* on *Floor 2*. In the second group, we suspect that *HU 7D* and *Service 14* are related to the spread of infection. These groups can be observed in the six heatmaps in Fig. 11.

4) STEP 5: CLUSTERING METHOD

Patients were grouped by applying the k-means clustering algorithm to the SM for each TDMA. For each SM, we tested the values for K in the range $[2, p]$ and ran k-means 10 times with different initializations for each value of K. Finally, for each TDMA, we selected the clustering result with the highest silhouette score (SC). Table 1 shows the average SC of the clusters of the SM for each TDMA. For each cluster, a set of quantitative measures is also shown, such as the number of patients, the average and standard deviation of the similarities of the SV in the cluster, the average distance of a patient in the cluster to any other patient in the cluster (inner distance), the average distance of a patient in the cluster to any other patient in the closest and farthest clusters (closest and furthest outer distances), and the average SC. The distance was calculated as the Euclidean distance between the SVs of two patients. Fig. 12 shows the clustering of the patients for each TDMA. Patients are represented by their SVs, which were reduced to a two-dimensional space. In general, the figures represent a cloud of points where it is difficult to distinguish any defined group of patients. Hence, the average SC of clustering on each TDMA (in this section, we refer to the SM returned by each TDMA as the name of the algorithm) is not very high, between 0.13 and 0.25. An interpretation [40] of SC is that a value of over 0.7 is ''strong,'' over 0.5 ''reasonable,'' and over 0.25 ''weak,'' representing points that are uniformly distributed.

Clustering on DTW divided the patients into eight groups. These clusters were small (each cluster had one to three patients) and had a high standard deviation (0.2, and the average similarity of their SVs was 0.3). There were clusters with a low inner distance, such as *clusters 6* and *7*, and others with a high inner distance, such as *clusters 1* and *2*. The average closest outer distance of each cluster also varied, which led to clusters with a SC ''stronger'' than others. Consequently, the average SC of the clustering for DTW was 0.25.

Similar to DTW, in ST-DTW, there are some patients whose SVs are similar to each other but different from the rest. However, unlike DTW, the clustering solution for ST-DTW with the highest SC grouped the patients into two large clusters and one with only one patient. The diversity among the patients in each cluster is reflected in their high standard deviations regarding the average similarity of their SVs (0.2 and 0.5, respectively), and that their average inner and outer distances were similar. Consequently, the SC of each cluster was low, with an average value of 0.116, which was lower than that of DTW. Note that the SC of *cluster 3* was zero because it only had one patient.

In contrast to DTW and ST-DTW there were no SVs with noticeably low or high distances to the others in STLC and JSTLC. Consequently, patients were grouped into two clusters of different sizes that only differed in one patient: in STLC, clusters had 2 and 15 patients, and in JSTLC, clusters had 3 and 14 patients. It must be noted that the clusters with the fewest patients are formed by those with the highest overall similarity: *1222* and *1632* in STLC, and also *1626* in JSTLC. Notably, the highest similarities of these patients do not necessarily coincide in the same SV positions, although the inner distances of the small clusters are slightly lower than those of the large clusters. However, it should also be noted that the two clusters of each TDMA have similar standard deviations (they can be considered low), and inner and outer distances. As previously mentioned, the STLC and JSTLC are formed by uniformly distributed points. Only slight differences between the inner and outer distances in STLC and JSTLC led to an average SC of 0.21 and 0.13, respectively.

In the case of clusters of ST-LCSS and ST-LCSS-WTW, we found SVs with overall low similarities but with some patients being spatiotemporally close. These patients were divided into two clusters that coincided with the two areas shown in Figs. 11(e) and 11(f), which grouped the highest similarities. The fact that there are many SVs whose cells can be divided into two differentiated ranges of values indicates that the standard deviation of the similarities of the clusters is high with respect to the average similarity: in both TDMAs, they are around 0.17. The clusters of both TDMAs were similar and presented similar average SCs: 0.195 for ST-LCSS and 0.188 for ST-LCSS-WTW. These average SCs were among the highest of all the clustering solutions. The average SC of ST-LCSS-WTW was slightly lower because of the higher inner distances of its clusters, which may be explained by the similarities between the close patients being higher in ST-LCSS-WTW than in ST-LCSS, whereas those between distant patients remained the same.

V. DISCUSSION

In this section, we discuss several topics related to the results of the TDMAs and the k-means clustering.

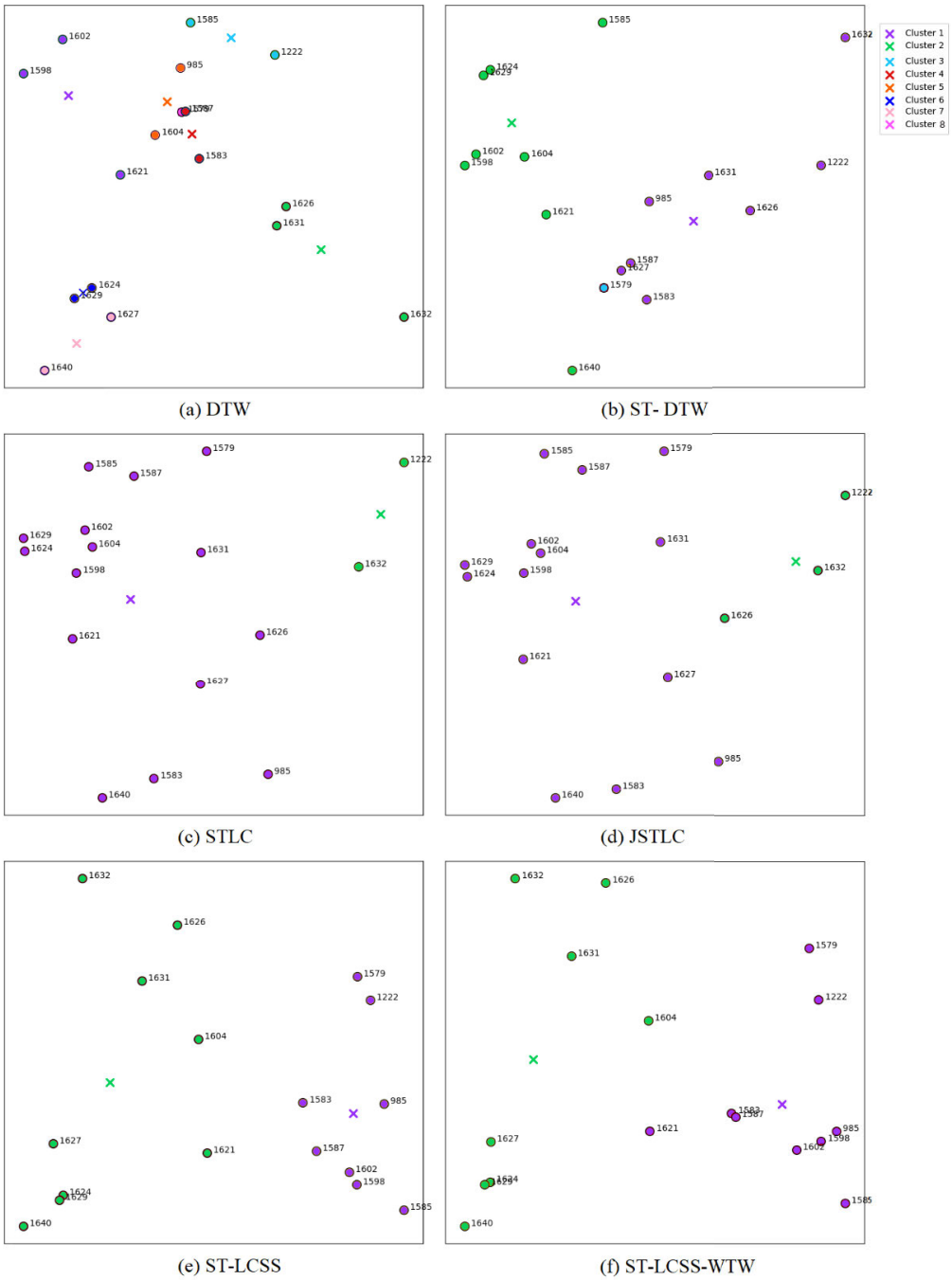


FIGURE 12. Representation of clusters in a two-dimensional space. The SVs of each patient are linearly reduced to two dimensions using a Singular Value Decomposition of the data. Centroids of the clusters are represented as a X.

A. DTW AND ST-DTW

DTW is an algorithm where the first and last points of the two trajectories must match and each of their intermediate points must match another at least once. These restrictions lead to the length of intersection trajectories (IT) was an important factor when calculating their similarity. Although DTW only evaluates the spatial dimension, the definition of the algorithm leads to the fact that the similarity is not only influenced by the trajectories having spatially close beds, but also by when these beds are used. Therefore, high similarity is not guaranteed only by points that are spatially and temporally aligned.

Fig. 13 illustrates three different examples in which DTW was used. In Fig. 13(a),the beds of T_1 and T_2 are close such that almost all the points match one-to-one. Fig. 13(b) shows an edge case where T_1^1 and T_2^1 are very close, all points up to the penultimate of T_1 occur in the same bed, and the beds from the rest of the points of T_2 are close to the bed of the last point of T_1. Consequently, there are matches between T_2^1 and all points in T_1 except its last, and between the last point in

TABLE 1. Clustering of the results of each TDMA using the K that provides the highest SC for each.

TDMA	SC	Clusters							
		ID	No. Pats	AVG Similarity	Standard Deviation of Similarity	AVG Inner Distance	AVG Closest Outer Distance	AVG Furthest Outer Distance	AVG SC
DTW	0.209	1	3	0.326	0.208	0.900	1.058	1.378	0.150
		2	3	0.385	0.215	0.852	1.172	1.436	0.273
		3	2	0.272	0.22	0.835	1.096	1.417	0.238
		4	2	0.335	0.208	0.852	1.043	1.225	0.183
		5	2	0.258	0.177	0.825	1.043	1.267	0.209
		6	2	0.316	0.220	0.539	1.044	1.362	0.484
		7	2	0.173	0.175	0.623	1.115	1.417	0.441
		8	1	0.428	0.220	-	1.070	1.417	0
		AVG		0.295	0.203	.0775	1.082	1.357	0.247
ST-DTW	0.116	1	9	0.490	0.194	1.062	1.172	1.179	0.094
		2	7	0.454	0.199	0.882	1.181	1.227	0.253
		3	1	0.578	0.213	-	1.227	1.243	0
		AVG		0.507	0.202	0.972	1.176	1.203	0.116
STLC	0.210	1	15	0.465	0.166	0.947	1.116	Same	0.151
		2	2	0.596	0.144	0.870	1.190	Same	0.268
		AVG		0.531	0.163	0.914	1.049	Same	1.130
JSTLC	0.130	1	14	0.448	0.160	0.986	1.082	Same	0.089
		2	3	0.553	0.165	1.842	1.016	Same	0.171
		AVG		0.501	0.163	0.914	1.049	Same	1.130
ST-LCSS	0.195	1	9	0.271	0.166	0.804	1.030	Same	0.220
		2	8	0.268	0.168	0.856	1.030	Same	0.170
		AVG		0.269	0.167	0.830	1.030	Same	0.195
ST-LCSS-WTW	0.188	1	8	0.273	0.172	0.872	1.084	Same	0.196
		2	9	0.268	0.168	0.897	1.096	Same	0.181
		AVG		0.270	0.170	0.885	1.090	Same	0.188

T_1 and all points in T_2, except for its first. In both examples, the spatiotemporal similarity would be high, but there are more matches in Fig. 13(b). Thus, its similarity would be higher. In the experiment, a situation similar to that shown in Fig. 13(b) could occur between patients *1579* and *1632*. Finally, Fig. 13(c) is a version of Fig. 13(b), where T_1^2 and T_2^2 are in close beds far from those of T_1^1 and T_2^1. In addition, the rest of the central beds in T_2 were at an intermediate distance from those of T_1. With these modifications, the number of matches and their overall similarity were lower than in Fig 13(b).

ST-DTW follows the same scheme as DTW but also evaluates the temporal dimension. If we continue with the examples in Fig. 13, this change will not significantly affect that shown in Fig. 13(a). However, the results in Figs. 13(b) and 13(c) were oppositely affected. In Fig. 13(b), some of the intermediate points of the trajectories would match each other, reducing the number of matches. In Fig. 13(c), the number of matches would not change, but the final similarity would increase. In DTW, points that are distant in space but close in time have little impact, but in ST-DTW, their high temporal similarity is also added. This explains the overall increase in the ST-DTW results, with a greater increase when the intersection trajectories between patients were longer. For example, the similarity between *Patients 1222* and *1579* is 0.4 points higher in ST-DTW.

In summary, DTW and ST-DTW are two algorithms with strong restrictions regarding the order in which the points can be matched but not the maximum distance between the points to allow the match. Consequently, there are many pairs of

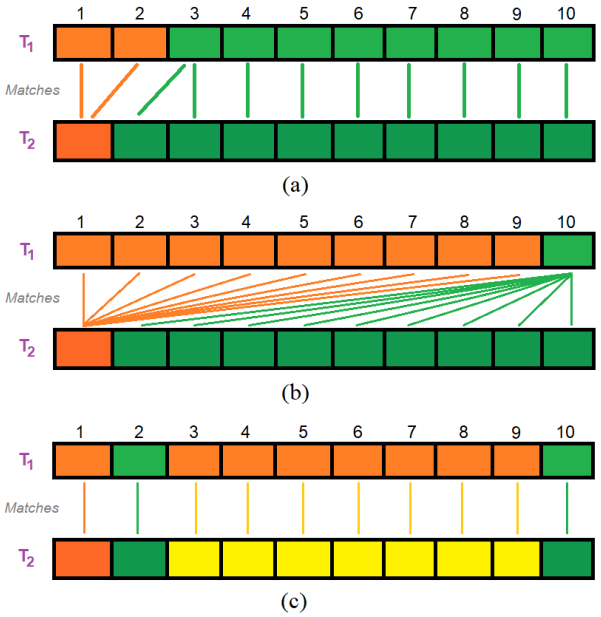


FIGURE 13. Representation of three cases where DTW is used to evaluate the spatiotemporal similarity between two trajectories, T_1 and T_2. Each trajectory is represented by a series of cells whose color represents the bed. We assume that the spatial distance between any two beds of the same color is the same and the lowest. The distance between yellow and green beds is the same as between yellow and orange beds, and it is lower than between green and orange beds.

patients with high similarity between them and low similarity with the rest, which causes the appearance of small clusters, particularly in DTW.

B. STLC AND JSTLC

STLC and JSTLC present a scheme in which each point of each trajectory must match at least once with a point of the other trajectory, without a match being conditioned by the previous ones. This implies that the longer the trajectories, the higher the similarity is. For example, patients *1222*, *1579*, *1585*, whereas *985* had some of the longest trajectories and highest similarities, and patients *1621* and *1640* had the opposite. Note that we do not consider it an issue: the more time and space two people share, the more an infection is likely to occur. However, evaluation of these dimensions can have a negative effect on the overall results. In STLC, the closest spatial and temporal points are searched separately for each point. In practice, the temporally closest point is always the one with the same timestamp. As the ITs between two trajectories have the same number of points and the same start and end dates, the temporal similarity is a measure of the length of the ITs, and only the spatial similarity is evaluated. The lack of restrictions in selecting the spatially closest point leads to, for example, $sim_{STLC}(T_1,T_2)$ being high and the same in Figs. 13(a) and 13(b). In the case of Fig. 13(c), as $sim_{STLC}(T_1,T_2)$ is calculated as the average of the similarities of each trajectory with the other, the high similarity between T_1 and T_2 would be compensated by the lower similarity between T_2 and T_1.

The JSTLC searches for the closest spatiotemporal point of each point. Thus, there must be a balance between both dimensions, and the proportion of points whose matches have a similarity close to the maximum should decrease. The overall similarity in Fig. 11(f) was slightly lower than that in Fig. 11(e). If we consider pairs of ITs with similar lengths, the ITs with the highest similarities were those that share most points that are close in space and time. However, like ST-DTW, there are points that are close in time, but not in space (and vice versa).

In both STLC and JSTLC, two facts (the requirement that all points must match, and the lack of limits on the distance to allow the match) lead to a medium-high overall similarity. This should not be considered as an issue. However, since there were no marked differences between the SVs of the patients, most of them were grouped in the same cluster, separating those that differed in other clusters. This is reflected in the fact that the clusters obtained from the results of STLC and JSTLC presented the largest average Euclidean distances between the SVs of the patients in the same cluster.

C. ST-LCSS AND ST-LCSS-WTW

Unlike the other TDMAs, ST-LCSS and ST-LCSS-WTW implement two parameters, ε and δ, to spatially and temporally limit when two points can match. The values for these parameters are essential because the effectiveness of the algorithm relies heavily on them and should be chosen based on the physical structure of the hospital and how the scientific literature characterizes the transmission of the infection.

If we consider the spatial dimension, we can take advantage of its hierarchical definition in the data model to fine-tune the evaluation of similarity between trajectories without changing the costs assigned to the edges of the graph database (see Fig. 1). For example, the first approach can be made at a superior level of the hierarchy, such as *the Floor*, because a lower level might result in most similarities between patients being zero, and therefore, not detecting a possible outbreak.If the most precise evaluation is needed, spatial matching can be restricted to *LogicZone* (a custom group of *Areas*), *Area* or *Service*. It is worth mentioning that the domain spatial hierarchy has several types of edges for neighborhood relations, such as *nextTo* and *opposite*. Thus, it is possible to represent different levels of proximity between points within the same *Floor*.

Regarding the results of the ST-LCSS and ST-LCSS-WTW, Figs. 11(c) and 11(d) show that the overall similarity between patients is medium-low. Given parameters ε and γ and that none of the points need to be matched, we can be more confident that a low or high similarity will have a relationship with points that are close in space and time within the set parameters. Similar to DTW and ST-DTW, the ST-LCSS conditions in which a new match cannot be made with points that are before the last match or that are already matched. In ST-LCSS-WTW, we attempted to lighten the first restriction and obtained a slight increase in the similarities between some patients, especially among those who already had a higher similarity than the others, differentiating them even more. An increase in the value of similarities benefits both the human analysis of the results and the clustering. ST-LCSS-WTW presents (except for DTW) clusters with the lowest average Euclidean distance within the cluster and an average distance with the other cluster similar to that of the other algorithms. Thus, clustering on the ST-LCSS-WTW results has one of the highest SC values (higher than that of ST-LCSS). In addition, the clusters in the ST-LCSS-WTW and ST-LCSS were the most similar to the patient groups observed in our visual interpretation of the SMs.

D. OTHER ASPECTS RELATED TO TDMAS

There are some aspects that are worth mentioning, although not directly related to how TDMAs work or their results.

From a computational cost perspective, ST-LCSS-WTW is the most expensive algorithm because it has branched execution that implies an order of exponential complexity. In addition, the branching level strongly depends on ω and the sampling frequency used to create trajectories. In contrast, STLC and JSTLC are algorithms with quadratic complexity orders. Despite its recursive definition, several optimizations have been studied for DTW (by extension, they can also be applied to ST-DTW) based on its implementation with dynamic programming [41], which presents a quadratic complexity order.

A matter of interest may also be the ease of modifying the algorithm in its implementation for parallel computation or adding new conditions to guide the search, such as setting minimal spatial and temporal similarities. In this sense, the JSTLC is an algorithm whose simplicity can facilitate these tasks. For example, the searches for the closest points of T_2 to each point of T_1 are independent so that they can be executed in parallel. Furthermore, when calculating the similarity between two points, we can add parameters ε and δ to check if the spatial and temporal similarities are above the defined minimum, and if not, then $sim_{ST}(p_1, p_2) = 0$. It is necessary to study whether a similar modification could be applied to DTW and ST-DTW.

With respect to this modified version of JSTLC with spatial and temporal constraints, its results would be like those of ST-LCSS-WTW, because both algorithms seek to maximize the matches between points and their total similarity. A relevant difference between the two algorithms is the number of matches that a point can have. In ST-LCSS-WTW, a point can match, at most, one point of the other trajectory. Meanwhile, in JSTLC, a point can be the closest to several points of the other trajectory, so there is a match between each pair. Weighing one alternative better depends on whether we consider that the probability of contagion only increases with a close and prolonged stay of both patients (ST-LCSS-WTW) or also with the susceptible patient remaining in a possibly contaminated environment (space, furniture and equipment, HCW), even if the other patient has left (modified JSTLC). Here, the characteristics of the pathogen are essential.

Although no explicit comparison has been made, the human interpretability of the results of each TDMA is also a relevant factor to consider, especially when considering the StESPT as a tool to support epidemiologists in detecting outbreaks among the patients. Apart from the similarity, some TDMAs return extra information for this purpose. ST-LCSS and ST-LCSS-WTW also return the number of matches between each trajectory and could return which these matches are and their similarity. This could help to determine the sources of transmission more precisely. Similarly, STLC and JSTLC can return, for each trajectory, which are the matches between their points and the closest points in the other. In the case of DTW, it is not possible to obtain this type of information.

E. TIME CONSIDERATIONS

Several aspects related to the definition of the time limits for comparing sampling points can be discussed.

When evaluating the similarity between two trajectories, T_1 and T_2, we define the intersection trajectories (IT) between them and start checking for matches from the first to the last points of each IT. However, we can propose checking points within one IT with points outside the other IT, but within the STPT. The following example illustrates a situation where this comparison could be beneficial: there are two patients, P_1 and P_2, whose hospitalizations start at different times, P_1 being the first. Therefore, T_1 will have some points before the IT between both patients (we consider that T_1 and T_2 are within the STPT). P_1 moves to another room when P_2 enters the hospital in the same room where P_1 was. The environment in which P_2 is located may be contaminated, but this situation is not considered when evaluating the similarity between their IT. A similar situation would occur if P_1's stay ended before P_2's, and P_2 was in a *Location* close to P_1's last *Location* when P_1 was no longer in the hospital.

We can set a margin to define the maximum allowed difference in time between two points when one is within the IT and the other is only within the STPT. Implementing this modification would be simple in STLC and JSTLC, because the matches between points are independent. For the rest of the algorithms, it is necessary to analyze whether these extra points that do not temporarily coincide with the other trajectory should be part of the *IT*, given that their benefits are situational. For example, matches (T_1^1, T_2^1) and (T_1^m, T_2^m) are mandatory in DTW and ST-DTW, respectively.

Another aspect to discuss is the definition of the endpoint of ITs between T_1 and T_2. It is defined based on the end of the STPT or the last point of one of the trajectories, depending on which time occurs first. However, the endpoint of the ITs can be defined as the sampling point at which the first positive *TestMicro* occurs in T_1 or T_2. This latter option would be appropriate for an epidemiological investigation focused on identifying which patients are most likely to have transmitted the infection to a particular patient. We want the study framework of this work to be broader, covering spatiotemporal relationships between patients beyond contagion, as they can be key in determining where the sources of the outbreak are located. For example, if a patient had a positive *TestMicro* while in a room but was later moved to another room, patients close to both rooms may be at risk of becoming infected.

F. CLUSTERING METHOD

When applying the k-means clustering algorithm, the results vary significantly depending on the TDMA used to generate the similarity matrices between patients. However, the six solutions have similar SCs ranging [0.13, 0.25].

K-means is an algorithm that creates K clusters by minimizing the sum of the distances (Euclidean distance) between each patient in the cluster and its centroid. This can be a drawback when we have a dataset, such as the SM of this work, in which there is no clear distinction to classify its elements, either because most of the data share similar values in each of the dimensions that define them (as, above all, in STLC and JSTLC) or because there are some small groups of data that are very similar to each other but are not very separated from the rest (as in DTW and ST-DTW). In search for a balance between the average distances between the patients in each cluster, patients can be divided into a few clusters that contain most patients or into many small clusters that divide patients into fine-grained levels. In practice, the relationship between cluster cohesion and separation (measured using SC) is similar for both options. Notably, clustering on the SM of DTW presents the highest SC. However, if we consider that

the STPT is eight days, that the number of patients among which to search for clusters is 17, and that they were in nearby locations, a smaller number of clusters would be more reasonable.

In the case of STLC and JSTLC, we could see that the overall medium-high similarities between patients and that the clustering algorithm groups the patients with the highest similarities apart from the rest as a sign that they are more specific for representing high similarities, zooming those patients that are particularly close. However, as shown in Figs. 11(c) and 11(d), the similarities between the patients forming these smaller clusters (*patients 1222*, *1632* and *1626*) were not the highest in each SV, and in Fig. 10, their trajectories do not seem to represent a group of patients who have been especially close during much of the STPT. Hence, in our experiment, we cannot be sure that the STLC and JSTLC could work as a ''loupe'' to highlight very closely related patients.

Concerning the clusters on ST-LCSS and ST-LCSS-WTW, we cannot determine whether the clusters found by k-means using their SMs are better than the SMs of the other TDMAs if we consider only the quantitative measures presented in Section IV-B.4. However, visually, we can notice that the clusters obtained with ST-LCSS and ST-LCSS-WTW are the most similar to the two groups of patients we deduced that could be spatiotemporally close in the previous step. Hence, based on the results when using ST-LCSS or ST-LCSS-WTW as TDMA, we concluded that the source of CD infection was the A&E service on *Floor 0*, and that, subsequently, two outbreaks occurred on *Floor 2*.

For all TDMAs, the clustering has an average SC of 0.2, which can be interpreted as low or ''weak.'' However, it has been reported that increasing the dimensionality of the data makes it difficult to obtain high SC values because the average distances between elements become similar: there is an increase in the space where the elements to clusters are, so they appear to be sparse, preventing the efficient detection of areas with similar properties [40]. Note that in our experiment, each patient is represented by a 17-dimensional vector (the similarity vectors have as many dimensions as patients to cluster). In addition, we normalized the spatial and temporal distances between two points from wider range to [0,1].

G. LIMITATIONS

To the best of our knowledge, there are no open repositories containing actual hospital data with the level of detail required to perform this type of study. A limitation can arise from the fact that the experiment presented herein was based on a synthetic dataset [37]. However, it is possible to generate different datasets using the same or different patient movement dynamics, bacterial transmission dynamics, and hospital layout. In addition, the values of the *cost* property in the edges of the spatial dimension in Fig. 1 represent distances in meters but are not based on actual measurements.

VI. CONCLUSION

This study aimed to support epidemiologists in investigating the epidemiological relationship between patients infected with bacterial HAI. For this purpose, we designed the StESPT methodology, where we defined a set of steps to quantify how spatiotemporally connected infected patients have been during their stay in the hospital. Finally, taking this as a basis, we try to determine if they are part of one or several outbreaks and which are the possible transmission routes. These steps are to obtain the infected patients and their trajectories, evaluate the similarity between two patient trajectories based on an equation that linearly combines the spatial and temporal similarities between their points, and cluster the patients based on their similarities with the k-means algorithm. We used the *silhouette coefficient* (SC) to measure the relationship between the cohesion and separation of the clusters.

We designed an experiment that simulates how epidemiologists could use StESPT to analyze the trajectories of several patients infected with *Clostridium difficile* in a hospital. Using the StESPT, we could determine the origin of the infection and the current existence of two outbreaks.

We modified three general-purpose TDMAs to calculate the similarity between two trajectories (DTW, STLC, and ST-LCSS) and created an extension with more significant changes for each one (ST-DTW, JSTLC, and ST-LCSS-WTW) to better adapt to the semantics of the problem.

The results of the equation used to evaluate the spatiotemporal similarity between points and the clustering algorithm depend on how each TDMA reflects the spatiotemporal relations between the trajectories, which are influenced by their characteristics.

We observed that evaluating the proximity between patients by combining spatial and temporal information provided better results than evaluating the spatial dimension alone (DTW and, in practice, STLC). We also realized two different cases when each point of one of the trajectories must match at least one point of the other trajectory: restrictions referring to that new matches must depend on the previous ones can lead to an over-increase in similarity between trajectories in some specific cases (DTW, ST-DTW), and the total absence of these restrictions can lead to a general similarity between all patients without significant variations (STLC, JSTLC).

Setting spatial and temporal limits to define when a match between points on two trajectories is allowed and, consequently, not forcing all points to be matched (ST-LCSS, ST-LCSS-WTW) enables the representation of similarities between trajectories that most closely reflect the spatiotemporal relationships between patients. When comparing ST-LCSS and ST-LCSS-WTW, the greater degree of independence in matching the points between two trajectories for ST-LCSS-WTW allows for greater accuracy in determining which points are spatiotemporally close. These spatial and temporal limits also enable the evaluation of trajectories at several precision levels. For example, spatial limits can be

supported by a flexible and multi-level model to represent the spatial dimension, which comprises several concepts and relations between them.

To conclude, in k-means, to achieve a high SC, data should not be uniformly dispersed (as in STLC and JSTLC) or with some data that have very similar characteristics (DTW, ST-DTW). Although with an average SC like the rest, the results from ST-LCSS and ST-LCSS-WTW led k-means to find the clusters of patients that most closely resemble what would have actually occurred based on the spatiotemporal relationships between the patients.

A. FUTURE WORKS

In the immediate future, we propose to improve the JSTLC by adding spatial and temporal parameters to limit the matches between points so that each point should not have at least one match. We want to compare the results of this modified version of the JSTLC with those of the ST-LCSS-WTW regarding their capabilities to represent spatiotemporal proximity between patients infected with diseases with different epidemiological characteristics.

We also plan to study the use of other types of clustering algorithms, such as hierarchical clustering, whose representation as a dendrogram of the merging of clusters can provide visual support for understanding the structure of the data; or fuzzy clustering, such as Fuzzy C-Means, in which an object in the dataset can be classified into more than one cluster. We plan to integrate the equation to calculate the spatiotemporal similarity between points and some TDMA with the clustering algorithm.

DATA AND SOFTWARE AVAILABILITY

The dataset and software used to execute the StESPT in the experiment are available in the following repository: github.com/LorenaPujante/StESPT. GraphDB was used as the RDF triplestore. The software used for the experiment was implemented in Python. The authors used the scikit-learn library for the clustering.

An extended description of the design of the hospital and the software to transform the output of [37] into a KG following the model from [10] are in the following repository: github.com/LorenaPujante/HospitalGeneratorRDF_V2.

REFERENCES

[1] C. Biondo, E. Ponzo, A. Midiri, G. B. Ostone, and G. Mancuso, "The dark side of nosocomial infections in critically ill COVID-19 patients," *Life*, vol. 13, no. 6, p. 1408, Jun. 2023, doi: 10.3390/life13061408.

[2] S. Szabó, B. Feier, D. Capatina, M. Tertis, C. Cristea, and A. Popa, "An overview of healthcare associated infections and their detection methods caused by pathogen bacteria in Romania and Europe," *J. Clin. Med.*, vol. 11, no. 11, p. 3204, Jun. 2022, doi: 10.3390/jcm11113204.

[3] F. Prestinaci, P. Pezzotti, and A. Pantosti, "Antimicrobial resistance: A global multifaceted phenomenon," *Pathogens Global Health*, vol. 109, no. 7, pp. 309–318, Oct. 2015, doi: 10.1179/2047773215y.0000000030.

[4] World Health Org. (2002). *Prevention of Hospital-Acquired Infections: A Practical Guide*. Accessed: Dec. 5, 2024. [Online]. Available: https://iris.who.int/handle/10665/67350

[5] N. Zhang, W. Chen, P.-T. Chan, H.-L. Yen, J. W.-T. Tang, and Y. Li, "Close contact behavior in indoor environment and transmission of respiratory infection," *Indoor Air*, vol. 30, no. 4, pp. 645–661, Jul. 2020, doi: 10.1111/ina.12673.

[6] P. Chen, D. Zhang, J. Liu, and I. Y. Jian, "Assessing personal exposure to COVID-19 transmission in public indoor spaces based on fine-grained trajectory data: A simulation study," *Building Environ.*, vol. 218, Jun. 2022, Art. no. 109153, doi: 10.1016/j.buildenv.2022.109153.

[7] Q. Zuo, J. Du, B. Di, J. Zhou, L. Zhang, H. Liu, and X. Hou, "Research on spatial–temporal spread and risk profile of the COVID-19 epidemic based on mobile phone trajectory data," *Frontiers Big Data*, vol. 5, Apr. 2022, Art. no. 705698, doi: 10.3389/fdata.2022.705698.

[8] Eur. Centre for Disease Prevention Control. (2017). *Surveillance of Healthcare-Associated Infections in Intensive Care Units*. [Online]. Available: https://data.europa.eu/doi/10.2900/833186

[9] *Guidelines for Clostridium Difficile (C. Diff) Outbreaks in Long-Term Care Facilities (LTCFs)*, West Virginia Bur. Public Health, 2013.

[10] L. Pujante-Otalora, M. Campos, J. M. Juarez, and B. Canovas-Segura, "Spatiotemporal data modelling for epidemiological research in hospitals," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 10, pp. 6280–6291, Oct. 2024, doi: 10.1109/JBHI.2024.3417224.

[11] H. Su, S. Liu, B. Zheng, X. Zhou, and K. Zheng, "A survey of trajectory distance measures and performance evaluation," *VLDB J.*, vol. 29, no. 1, pp. 3–32, Oct. 2019, doi: 10.1007/s00778-019-00574-9.

[12] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining Knowl. Discovery*, vol. 26, no. 2, pp. 275–309, Feb. 2012, doi: 10.1007/s10618-012-0250-5.

[13] I. M. Kamal, H. Bae, and K. H. Cho, "Trajectory linkage and spreader centrality for social epidemic networks," *IEEE Access*, vol. 8, pp. 210922–210934, 2020, doi: 10.1109/ACCESS.2020.3039260.

[14] L. A. Cowley et al., "Genomics, social media and mobile phone data enable mapping of SARS-CoV-2 lineages to inform health policy in Bangladesh," *Nature Microbiology*, vol. 6, no. 10, pp. 1271–1278, Sep. 2021, doi: 10.1038/s41564-021-00955-3.

[15] J. Erber, V. Kappler, B. Haller, H. Mijočević, A. Galhoz, C. Prazeres da Costa, F. Gebhardt, N. Graf, D. Hoffmann, M. Thaler, E. Lorenz, H. Roggendorf, F. Kohlmayer, A. Henkel, M. P. Menden, J. Ruland, C. D. Spinner, U. Protzer, P. Knolle, and P. Lingor, "Infection control measures and prevalence of SARS-CoV-2 IgG among 4,554 university hospital employees, Munich, Germany," *Emerg. Infectious Diseases*, vol. 28, no. 3, pp. 572–581, Mar. 2022, doi: 10.3201/eid2803.204436.

[16] A. Myall, J. R. Price, R. L. Peach, M. Abbas, S. Mookerjee, N. Zhu, I. Ahmad, D. Ming, F. Ramzan, D. Teixeira, C. Graf, A. Y. Weiße, S. Harbarth, A. Holmes, and M. Barahona, "Prediction of hospital-onset COVID-19 infections using dynamic networks of patient contact: An international retrospective cohort study," *Lancet Digit. Health*, vol. 4, no. 8, pp. e573–e583, Aug. 2022, doi: 10.1016/s2589-7500(22)00093-0.

[17] A. Myall, R. L. Peach, Y. Wan, S. Mookerjee, E. Jauneikaite, F. Bolt, J. Price, F. Davies, A. Weiße, A. Holmes, and M. Barahona, "Characterising contact in disease outbreaks via a network model of spatial–temporal proximity," *MedRxiv*, 2021, doi: 10.1101/2021.04.07.21254497. [Online]. Available: https://europepmc.org/article/PPR/PPR309081

[18] Y. Wan et al., "Integrated analysis of patient networks and plasmid genomes to investigate a regional, multispecies outbreak of carbapenemase-producing enterobacterales carrying both blaIMP and mcr-9 genes," *J. Infectious Diseases*, vol. 230, no. 1, pp. e159–e170, Jul. 2024, doi: 10.1093/infdis/jiae019.

[19] J. R. Price, S. Mookerjee, E. Dyakova, A. Myall, W. Leung, A. Y. Weiße, Y. Shersing, E. T. Brannigan, T. Galletly, D. Muir, P. Randell, F. Davies, F. Bolt, M. Barahona, J. A. Otter, and A. H. Holmes, "Development and delivery of a real-time hospital-onset COVID-19 surveillance system using network analysis," *Clin. Infectious Diseases*, vol. 72, no. 1, pp. 82–89, Jan. 2021, doi: 10.1093/cid/ciaa892.

[20] S. Pai, P. M. Polgreen, A. M. Segre, D. K. Sewell, and S. V. Pemmaraju, "Spatiotemporal clustering of in-hospital clostridioides difficile infection," *Infection Control Hospital Epidemiology*, vol. 41, no. 4, pp. 418–424, Apr. 2020, doi: 10.1017/ice.2019.350.

[21] A. C. Myall, R. L. Peach, A. Y. Weiße, S. Mookerjee, F. Davies, A. Holmes, and M. Barahona, "Network memory in the movement of hospital patients carrying antimicrobial-resistant bacteria," *Appl. Netw. Sci.*, vol. 6, no. 1, p. 34, Dec. 2021, doi: 10.1007/s41109-021-00376-5.

[22] P. P. Da Silva, F. A. Da Silva, C. A. S. Rodrigues, L. P. Souza, E. M. De Lima, M. H. B. Pereira, C. N. Candella, M. Z. De Oliveira Alves, N. D. Lourenço, W. S. Tassinari, C. Barcellos, M. Z. R. Gomes, V. P. R. Dutra, M. C. Da Silva, J. P. S. Tonhá, L. S. De Mello, M. M. Castro, Y. R. Mathuiy, and A. A. Da Silva Machado, "Geographical information system and spatial–temporal statistics for monitoring infectious agents in hospital: A model using klebsiella pneumoniae complex," *Antimicrobial Resistance Infection Control*, vol. 10, no. 1, p. 92, Dec. 2021, doi: 10.1186/s13756-021-00944-5.

[23] R. D'Mello, J. Melcher, and J. Torous, "Similarity matrix-based anomaly detection for clinical intervention," *Sci. Rep.*, vol. 12, no. 1, p. 9162, Jun. 2022, doi: 10.1038/s41598-022-12792-3.

[24] D. Brockmann and D. Helbing, "The hidden geometry of complex, network-driven contagion phenomena," *Science*, vol. 342, no. 6164, pp. 1337–1342, Dec. 2013, doi: 10.1126/science.1245200.

[25] D. Hu, L. Chen, H. Fang, Z. Fang, T. Li, and Y. Gao, "Spatio-temporal trajectory similarity measures: A comprehensive survey and quantitative study," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 5, pp. 2191–2212, May 2024, doi: 10.1109/TKDE.2023.3323535.

[26] J. Zhu, X. Niu, F. Li, Y. Wang, P. Fournier-Viger, and K. She, "STTraj2Vec: A spatio-temporal trajectory representation learning approach," *Knowl.-Based Syst.*, vol. 300, Sep. 2024, Art. no. 112207, doi: 10.1016/j.knosys.2024.112207.

[27] S. Shang, R. Ding, K. Zheng, C. S. Jensen, P. Kalnis, and X. Zhou, "Personalized trajectory matching in spatial networks," *VLDB J.*, vol. 23, no. 3, pp. 449–468, Jul. 2013, doi: 10.1007/s00778-013-0331-0.

[28] A. Mubayi, A. Pandey, C. Brasic, A. Mubayi, P. Ghosh, and A. Ghosh, "Analytical estimation of data-motivated time-dependent disease transmission rate: An application to ebola and selected public health problems," *Tropical Med. Infectious Disease*, vol. 6, no. 3, p. 141, Jul. 2021, doi: 10.3390/tropicalmed6030141.

[29] B.-K. Yi, H. V. Jagadish, and C. Faloutsos, "Efficient retrieval of similar time sequences under time warping," in *Proc. 14th Int. Conf. Data Eng.*, 1998, pp. 201–208, doi: 10.1109/ICDE.1998.655778.

[30] S. Shang, L. Chen, Z. Wei, C. S. Jensen, K. Zheng, and P. Kalnis, "Trajectory similarity join in spatial networks," *Proc. VLDB Endowment*, vol. 10, no. 11, pp. 1178–1189, Aug. 2017, doi: 10.14778/3137628.3137630.

[31] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, Dec. 2001, doi: 10.1023/A:1012801612483.

[32] D. Greene, P. Cunningham, and R. Mayer, "Unsupervised learning and clustering," in *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval* (Cognitive Technologies). Berlin, Germany: Springer, 2008, pp. 51–90, doi: 10.1007/978-3-540-75171-7_3.

[33] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[34] J.-O. Palacio-Niño and F. Berzal, "Evaluation metrics for unsupervised learning algorithms," 2019, *arXiv:1905.05667*.

[35] J. Codella, N. Safdar, R. Heffernan, and O. Alagoz, "An agent-based simulation model for *clostridium difficile* infection control," *Med. Decis. Making*, vol. 35, no. 2, pp. 211–229, Feb. 2015, doi: 10.1177/0272989X14545788.

[36] R. B. McFee and G. G. Abdelsayed, "Clostridium difficile," *Disease-Month*, vol. 55, no. 7, pp. 439–470, Jul. 2009, doi: 10.1016/j.disamonth.2009.04.010.

[37] D. Kim, B. Canovas-Segura, A. Jimeno-Almazán, M. Campos, and J. M. Juarez, "Spatial–temporal simulation for hospital infection spread and outbreaks of clostridioides difficile," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, Nov. 2023, doi: 10.1038/s41598-023-47296-1.

[38] S. R. Curry, M. T. Hecker, J. O'Hagan, P. K. Kutty, Y. K. Ng-Wong, J. L. Cadnum, H. Alhmidi, M. I. Gonzalez-Orta, C. Saldana, B. M. Wilson, and C. J. Donskey, "Incubation period of clostridioides difficile infection in hospitalized patients and long-term care facility residents: A prospective cohort study," *Antimicrobial Stewardship Healthcare Epidemiology*, vol. 4, no. 1, p. 144, Sep. 2024, doi: 10.1017/ash.2024.392.

[39] M. Skally, K. Bennett, H. Humphreys, and F. Fitzpatrick, "Rethinking clostridioides difficile infection (CDI) surveillance definitions based on changing healthcare utilisation and a more realistic incubation period: Reviewing data from a tertiary-referral hospital, Ireland, 2012 to 2021," *Eurosurveillance*, vol. 29, no. 6, Feb. 2024, Art. no. 2300335, doi: 10.2807/1560-7917.es.2024.29.6.2300335.

[40] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is 'nearest neighbor' meaningful?" in *Proc. 7th Int. Conf. Database Theory*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 1540, 1998, pp. 217–235, doi: 10.1007/3-540-49257-7_15.

[41] E. Mizutani and S. Dreyfus, "On using dynamic programming for time warping in pattern recognition," *Inf. Sci.*, vol. 580, pp. 684–704, Nov. 2021, doi: 10.1016/j.ins.2021.08.075.

LORENA PUJANTE-OTALORA received the B.Sc. and M.Sc. degrees in computer science engineering from the University of Murcia, Spain, in 2019 and 2020, respectively, where she is currently pursuing the Ph.D. degree. Her primary research interests include the graph-based representation of epidemiological information and the use of machine learning algorithms for its analysis to study how hospital-acquired infections spread.

MANUEL CAMPOS received the B.S. degree in computer science engineering and the Ph.D. degree in computer science from the University of Murcia, Spain, in 2001 and 2007, respectively. Currently, he is an Associate Professor with the University of Murcia. He was the Vice-Dean for External Relations of the Faculty of Computer Science, University of Murcia, from 2010 to 2012, and the Coordinator of Internships in the Studies Vicerrectorate, University of Murcia, from 2012 to 2017. He is the Co-Leader of the MedAILab Research Group. He has been the main researcher in a European Project, several Spanish research projects, and contracts focused on clinical applications of the research lines, with more than 100 national and international conferences, and journals. His research interests include clinical intelligence systems, medical informatics, and data analysis and mining.

JOSE M. JUAREZ received the B.Sc. and M.Sc. degrees in computer science engineering from the University of Murcia, Spain, in 2003, and the Ph.D. degree in artificial intelligence from the University of Murcia, in 2007. Currently, he is an Associate Professor with the University of Murcia. He was the Vice Dean of the Faculty of Computer Science, University of Murcia, from 2014 to 2017. In particular, he focuses on eXplainable Artificial Intelligence and clinical decision support systems, including a wide variety of AI methods. He is the Co-Leader of the MedAILab Research Group. He has more than 80 articles in scientific journals, conference proceedings, book chapters, and proceedings. He also led and participated in several European and Spanish research projects. His research interests include general, artificial intelligence, and its uses in healthcare clinical intelligent systems. He is a Board Member of the European Society of Artificial Intelligence in Medicine (AIME) and the Co-Founder and a Board Member of the Spanish Society of AI in Biomedicine (AI-BIOMED). He is also a fellow of the Spanish Society of Artificial Intelligence (AEPIA) and the Spanish Society of Health Informatics (SEIS). He is an Associate Editor of *Artificial Intelligence in Medicine* (Elsevier) and an Academic Editor of *PLOS One*.