UNIVERSIDAD DE MURCIA

Escuela de Doctorado

TESIS DOCTORAL

Combinación de modelado y razonamiento espacio-temporal con tecnologías de grafos aplicado a epidemiología de infecciones nosocomiales e infecciones multirresistentes

AUTOR/A Dña. Lorena Pujante Otálora

DIRECTOR/ES D. Manuel Campos Martínez
D. José Manuel Juárez Herrero

2025

UNIVERSIDAD DE MURCIA Escuela de Doctorado

TESIS DOCTORAL

Combinación de modelado y razonamiento espacio-temporal con tecnologías de grafos aplicado a epidemiología de infecciones nosocomiales e infecciones multirresistentes

AUTOR/A Dña. Lorena Pujante Otálora

DIRECTOR/ES D. Manuel Campos Martínez
D. José Manuel Juárez Herrero

2025

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TITULO DE DOCTOR/A

Aprobado por la Comisión General de Doctorado el 19 de octubre de 2022.

Yo, Dª. Lorena Pujante Otálora, habiendo cursado el Programa de Doctorado en Informática de la Escuela Internacional de Doctorado de la Universidad de Murcia (EIDUM), como autor/a de la tesis presentada para la obtención del título de Doctor/a titulada:

Combinación de modelado y razonamiento espacio-temporal con tecnologías de grafos aplicado a epidemiología de infecciones nosocomiales e infecciones multirresistentes

y dirigida por:

D.: Manuel Campos Martínez D.: José Manuel Juárez Herrero D.:

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

Murcia, a 27 de mayo de 2025

(firma)

Información básica sobre protección de sus datos personales aportados:	
Responsable	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad	Gestionar su declaración de autoría y originalidad
Destinatarios	No se prevén comunicaciones de datos
Derechos	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la quinta hoja, después de la portada de la tesis presentada para la obtención del título de Doctor/a.

Firmante: LORENA PUJANTE OTALORA; Fecha-hora: 27/05/2025 15:08:44; Emisor del certificado: C=ES,O=ACCV,OU=PKIACCV,CN=ACCVCA-120;

@dministración electrónica

Código seguro de verificación: RUxFMqfA-nkCAGVyR-8AOEf8kb-ftqBi3EX

COPIA ELECTRÓNICA - Página 1 de 1

Esta es una copia auténtica imprimible de un documento administrativo electrónico archivado por la Universidad de Murcia, según el artículo 27.3 c) de la Ley 39/2015, de 1 de octubre. Su autenticidad puede ser contrastada a través de la siguiente dirección: https://sede.um.es/validador/

The sun's out... This meteorological transformation is most splendid!
Like a felicitous twist of fate in the face of certain doom.

Fischl, Prinzessin der Verurteilung

Agradecimientos

Resulta casi irónico que, tras el arduo trabajo que ha conllevado la investigación plasmada en este documento, uno de los mayores retos sea el poder condensar en unas pocas líneas todo el agradecimiento que siento hacia aquellas personas que me han acompañado durante este camino.

A nivel institucional, quiero agradecer a la Universidad de Murcia y a la Escuela Internacional de Doctorado por la oportunidad de llevar a cabo esta investigación. De igual manera, a la Fundación Séneca por la financiación de este proyecto de tesis.

Como se acostumbra, he de mencionar a mis directores Manuel Campos y José Manuel Juárez. También, a Bernardo, que siempre ha brindado su ayuda en la parte más complicada de un artículo: su redacción.

En lo personal, me gustaría agradecer a mi familia y amigos por su apoyo continuo, por animarme en los momentos más duros, por confiar en que podría lograrlo, y por su infinita comprensión.

En especial, quiero agradecer a mis padres, Tomás y Josefa, quienes siempre han sido mi modelo a seguir. Fueron ellos quienes me inculcaron desde pequeña que nuestros objetivos sólo pueden cumplirse con perseverancia, esfuerzo y sacrificio. Me lo han dado todo y sé que, sin ellos, nada de esto habría podido ser posible.

A mi hermano Raúl, por nuestras anécdotas, bromas y risas. Llevamos años repitiéndolas pero, aún hoy, consiguen alegrarme el día.

A mi pareja Alberto. Mi compañero, con quien he compartido tanto risas y alegrías como los altibajos de la vida. Nadie mejor que él me conoce y nadie mejor que él podría haberme apoyado, escuchado, aconsejado y alentado a seguir. Él es ese pequeño trozo de mundo al que llamo paz mental.

A José Alberto. Siempre le digo entre risas que me salvó, pero no sabe cuán cierto es el significado de esas palabras.

A María José, por brindarme la confianza y oportunidad de descubrir el bonito camino de la enseñanza.

Para terminar, no podría olvidarme de todas mis mascotas, por su cariño y compañía, por ser quienes más me alegran cada día. En especial a mis gatos Salem, no podría quererlo más, pero nunca menos; a mi chiquitín Yago; a mi bolita Pochita; y al favorito de la casa, Pelufa. No podría no incluir a Jonesy, Bambi, Tarzán, Mona y Negrito. También mis lobas chicas Maya, Sally y Loba. Por último, mis gallinas Úrsula, Ruperta grande, Ruperta chica, Clotis, Pollito Negro, Pollo Patillas y Rogelio, Patricia, Aurelia y Claudia, Yuki y Cucú, Lady Susan y Gaviota, y Berta y Barbie. Sin olvidar a Julio, el pavo real y supervisor de obras. Llegaron de improvisto, pero nunca en mejor momento.

Financiación

Este trabajo fue financiado por la beca del Subprograma Regional de formación de personal investigador en universidades y organismos públicos de investigación de la Región de Murcia en los ámbitos académico y de interés para la industria integrada en el Programa Séneca 2020, Región de Murcia, España (Ref:21460/FPI/20).

Además, este trabajo fue parcialmente financiado por el proyecto SITSUS (Ref: RTI2018-094832-B-I00), financiado por el MCIU de España, la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER); el proyecto CONFAINCE (Ref: PID2021-122194OB-I00) por MCIN/AEI/10.13039/501100011033 y por "FEDER A way of making Europe".

Resumen

Esta tesis doctoral tiene como objetivo demostrar que el uso combinado de técnicas de modelado y razonamiento espacio-temporal con tecnologías basadas en grafos son efectivas para el análisis epidemiológico de infecciones nosocomiales.

Las infecciones nosocomiales, especialmente aquellas causadas por bacterias multirresistentes, representan un problema de salud pública global, debido a su rápida propagación y alto índice de mortalidad. En este contexto, planteamos como objetivo principal el diseño y formalización en forma de grafos de un modelo de datos y operacional que sirva de base para futuras investigaciones relacionadas con el análisis epidemiológico espacial y temporal basado en los movimientos y contactos entre pacientes dentro del hospital. Este modelo busca servir de base para tareas epidemiológicas fundamentales como la detección de brotes, de cadenas de transmisión entre pacientes y de potenciales fuentes de contagio (ya sean ubicaciones físicas o áreas dentro del hospital, o servicios o unidades del en los que se organiza el personal sanitario).

Hemos desarrollado la investigación como se detalla a continuación. En primer lugar, realizamos un análisis exhaustivo del estado del arte sobre el uso de redes en modelos computacionales aplicados a la propagación de brotes epidémicos, identificando tendencias en escalas espaciales y temporales, tipos de redes utilizadas y fuentes de datos. Posteriormente, proponemos un modelo de datos espacio-temporal que incluye una dimensión espacial jerárquica (estructura física del hospital y organización del personal) y una dimensión temporal basada en los eventos clínicos registrados en el Sistema de Información Hospitalario (SIH) o Historia Clínica Electrónica del paciente (HCE). A partir de este modelo, hemos diseñado y formalizado seis consultas que representan tareas epidemiológicas fundamentales en la vigilancia y detección de brotes nosocomiales. En cuanto a la formalización del modelo y las consultas, hemos evaluado dos tecnologías orientadas a grafos: grafos de propiedades y grafos de conocimiento (formatos RDF y RDF*). Concretamente, hemos evaluado el rendimiento (tiempo ejecución y consumo de memoria principal) de dos bases de datos orientadas a grafos representativas de estas tecnologías: Neo4j y GraphDB. Los resultados muestran diferencias significativas en términos de eficiencia y expresividad según la tecnología empleada, destacando la escalabilidad de GraphDB y la utilidad de RDF* (y su lenguaje de consulta, SPARQL*) en la implementación de relaciones con propiedades y en las ventajas que ofrece el ser estándares.

Finalmente, a partir del modelo y consultas propuestas, diseñamos y validamos un nuevo método denominado StESPT (*Spatio-Temporal Epidemiological Similarity based on Patient Trajectories*), orientado al descubrimiento de grupos de pacientes infectados espacial y temporalmente conectados. Este método consiste en cinco pasos: obtención

de los pacientes infectados potenciales de un brote y transformación de sus estancias hospitalarias en trayectorias, cálculo de la similitud espacio-temporal epidemiológica entre puntos y entre trayectorias, y aplicación de técnicas de clustering para identificar posibles brotes y rutas de contagio.

Validamos el modelo, consultas y método propuestos utilizando datos clínicos sintéticos.

En las conclusiones destacamos que la combinación de tecnologías de grafos y razonamiento espacio-temporal aporta una representación flexible, expresiva y eficiente para el análisis de infecciones nosocomiales causadas por bacterias en entornos hospitalarios. Este enfoque facilita tareas como la búsqueda de casos índice, la reconstrucción de brotes y el descubrimiento de patrones ocultos, contribuyendo así a la toma de decisiones clínicas y al diseño de estrategias preventivas. Además, la tesis ha sido desarrollada bajo principios de ciencia abierta, garantizando la reproducibilidad de los experimentos mediante la publicación de código y datos en repositorios públicos.

Abstract

This doctoral thesis aims to demonstrate that spatiotemporal modelling and reasoning techniques combined with graph-based technologies are effective for the epidemiological analysis of nosocomial infections.

Nosocomial infections, particularly those caused by multidrug-resistant bacteria, represent a global public health issue due to their rapid spread and high mortality rate. In this context, the main objective is the design and formalisation, in the form of graphs, of a data and operational model that serves as a foundation for future research related to spatio-temporal epidemiological analysis based on patient movements and contacts within the hospital. This model is intended to support fundamental epidemiological tasks such as outbreak detection, identification of transmission chains between patients, and recognition of potential sources of infection (whether physical locations or areas within the hospital or the services or units in which healthcare staff are organised).

The research has been developed in detail below. Firstly, a comprehensive analysis of the state of the art regarding the use of networks in computational models applied to the spread of epidemic outbreaks is conducted, identifying trends in spatial and temporal scales, types of networks used, and data sources. Subsequently, a spatio-temporal data model is proposed, which includes a hierarchical spatial dimension (the physical structure of the hospital and the organisation of the staff) and a temporal dimension based on clinical events recorded in the Hospital Information System (HIS) or the patient's Electronic Health Record (EHR). Based on this model, six queries representing fundamental epidemiological tasks for the surveillance and detection of nosocomial outbreaks have been designed and formalised. Regarding the formalisation of the model and the queries, two graph-oriented technologies have been evaluated: property graphs and knowledge graphs (RDF and RDF* formats). Specifically, the performance (execution time and primary memory consumption) of two graph-oriented databases representative of these technologies—Neo4j and GraphDB—has been assessed. The results show significant differences in terms of efficiency and expressiveness depending on the technology used, highlighting the scalability of GraphDB and the usefulness of RDF* (and its query language, SPARQL*) in the implementation of property-based relationships and in the advantages offered by being standard-compliant.

Finally, based on the proposed model and queries, a new method called StESPT (Spatio-Temporal Epidemiological Similarity based on Patient Trajectories) is designed and validated, aimed at discovering groups of infected patients who are spatially and temporally connected. This method consists of five steps: obtaining the potentially infected patients of an outbreak and transforming their hospital stays into trajectories, calculating the spatio-temporal epidemiological similarity between points and between

trajectories, and applying clustering techniques to identify outbreaks and transmission routes.

The validation of the proposed model, queries, and method is conducted using synthetic clinical data.

The conclusions highlight that the combination of graph technologies and spatio-temporal reasoning provides a flexible, expressive, and efficient representation for analysing nosocomial infections caused by bacteria in hospital environments. This approach facilitates tasks such as index case search, outbreak reconstruction, and discovering hidden patterns, thus contributing to clinical decision-making and the design of preventive strategies. Additionally, the thesis has been developed under open science principles, ensuring the reproducibility of experiments through the publication of code and data in public repositories.

Índice de Contenido

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1.1 Motivación

La epidemiología es el estudio de la distribución, frecuencia, magnitud y factores de riesgo de las enfermedades infecciosas que se presentan en una determinada población humana [168]. Entre las principales áreas del estudio epidemiológico se encuentran: describir las causas, transmisión e historia natural de las enfermedades, la investigación de brotes epidémicos y el desarrollo de estrategias para prevención y control [220, 223].

En los últimos años, las infecciones intrahospitalarias, también conocidas como infecciones nosocomiales (IN), se han convertido en un importante problema de salud pública. Las IN son aquellas que no están presentes ni en fase de incubación en el momento en que un paciente es hospitalizado. En general, se considera que una infección es nosocomial cuando es detectada a partir de las 48 horas desde el ingreso [89, 135].

Las IN son a menudo causadas por patógenos multirresistentes (MDR, del inglés *multidrug resistant*), especialmente bacterias. Las bacterias MDR son aquellas que han mutado en el tiempo, de manera que han dejado de ser susceptibles a los medicamentos comúnmente utilizados para tratarlos. Esto provoca su rápida propagación y un aumento en su gravedad y en el riesgo de muerte. El tratamiento contra las bacterias MDR puede considerarse crítico, pues el uso abusivo de ciertos antimicrobianos y antibióticos puede resultar en la generación de nuevas resistencias, viéndose reducido el número de tratamientos disponibles [218].

Se estima que, aproximadamente, 1,27 millones de las muertes a nivel mundial pueden ser atribuidas a bacterias MDR [147]. En el caso concreto de España, en un estudio liderado por la Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC) [164] se estimó que en el año 2023 hubo alrededor de 173.000 casos de pacientes infectados por bacterias MDR, que resultaron en la muerte de aproximadamente 24.500. En este estudio también se estima que la incidencia de casos de las IN en las que está involucrado una bacteria MDR se ha incrementado un 11 % desde 2018, y su mortalidad un 22 %. Entre las principales razones del aumento de la incidencia y prevalencia de las IN por bacterias MDR, a nivel mundial, se encuentra el hecho de que los hospitales atienden a un número cada vez mayor de pacientes, la

transferencia de patógenos del personal médico al paciente o del entorno al paciente, el aumento de la resistencia a los antibióticos, el incumplimiento o la falta de protocolos de saneamiento y el escaso énfasis en la prevención [199].

Las IN se han convertido en una de las principales causas del aumento de la morbilidad y mortalidad en pacientes hospitalizados. Los pacientes infectados con bacterias MDR tienden a presentar peores pronósticos, siendo estas una especial amenaza tras intervenciones quirúrgicas [20]. Las IN causadas por bacterias MDR también están relacionadas con un aumento significativo de los costes sanitarios debido a la necesidad de cuidados más intensivos y estancias de mayor duración [151].

En potencia, cualquier patógeno encontrado en un hospital puede causar una infección nosocomial. Sin embargo, la mayoría de ellas son causadas por un reducido grupo de bacterias, entre las que destaca [135]: *Methicillin-resistant Staphylococcus aureus* (MRSA), de especial importancia en UCI donde tiene una mayor incidencia así como severidad y provoca un mayor tiempo de estancia de los pacientes infectados [202]; *Klebsiella pneumoniae*, una bacteria de gran relevancia como causa en infecciones oportunistas en el tracto urinario y sistema respiratorio, sobre todo entre pacientes ingresados en UCI y en neonatos (se estima que el 10 % de las infecciones nosocomiales están causadas por ella) [107, 142]; y *Clostridium difficile* (*C. diff*). Esta última es de especial relevancia, debido al aumento de su incidencia, severidad y mortalidad en los últimos años, sobre todo en Europa y América del Norte [41]. Además, uno de sus aspectos más desafiantes es la recurrencia de la enfermedad tras completar con éxito el tratamiento, ya sea con una cepa diferente de *C. diff* o por la persistencia de la cepa responsable del episodio inicial [96]. Se ha convertido en una de las principales causas de infección nosocomial a nivel mundial y la principal causa de infección de colon y diarrea infecciosa en pacientes hospitalizados. Se trata de una bacteria que se transmite a través de esporas a través de una ruta fecal-oral. Las esporas pueden permanecer en un hospital por largos periodos de tiempo, pudiendo actuar como reservorio cualquier superficie, dispositivo o material (camas, inodoros, bañeras o termómetros rectales, entre otros) que haya sido contaminada puede servir como reservorio para las esporas de *C. diff*. La descontaminación de habitaciones y la higiene de manos de los trabajadores sanitarios son algunas de las estrategias más seguidas para reducir su transmisión, aunque aún sigue siendo un desafío su consecución [23].

Dentro de un ambiente hospitalario, los patógenos nosocomiales pueden transmitirse por contacto directo entre pacientes, por el contacto con un trabajador sanitario durante la atención al paciente, o por la exposición a un ambiente contaminado (aire, superficies, equipo médico infectado) [191, 221]. La propagación de los patógenos puede conducir a la aparición de brotes infecciosos o epidémicos, es decir, un aumento en el número de casos, por lo general en un período de tiempo relativamente corto, en un área concreta del hospital.

Debido a importancia clínica y epidemiológica de la *C. diff* y a su rápida transmisión de persona a persona o mediante el contacto con superficies contaminadas, hemos validado las propuestas presentadas en esta tesis doctoral mediante el análisis de brotes que se adhieren a las características de transmisión de *C. diff*. Se trata de una infección en la que el contacto directo e indirecto entre pacientes y el personal sanitario juega un papel crucial en su propagación.

Ante la problemática que supone la propagación de los patógenos nosocomiales, es fundamental que los sistemas sanitarios cuenten con recursos y herramientas adecuados para limitar la aparición y propagación de estas IN. En relación con ello, tanto la Organización Mundial de la Salud (OMS) [169, 222] como el Centro para el Control y la Prevención de Enfermedades [191] proponen una serie de pautas y recomendaciones para reducir la incidencia de IN por bacterias MDR. Entre estas recomendaciones se incluyen: la mejora en los sistemas de salud y la vigilancia (identificación precoz de focos infecciosos, la notificación al personal sanitario, y el seguimiento de los pacientes infectados), la mejora del uso de antimicrobianos, y el desarrollo de nuevos medicamentos y vacunas.

Asimismo, se destaca la importancia de analizar de forma continuada la frecuencia esperada de estas infecciones en cada área del hospital. Este indicador está influido por diversos factores, tales como el tamaño del centro, el perfil de los pacientes ingresados, su localización, las características del patógeno (modo de transmisión, estacionalidad, capacidad de resistencia, entre otros), y la efectividad de las medidas de prevención y control adoptadas por el hospital.

Se subraya que, a pesar de su importancia, la identificación de nuevos casos positivos de un patógeno nosocomial y el entrecruzado de las ubicaciones espaciales y temporales de estos pacientes se realiza mediante un proceso manual que implica la revisión diaria o semanal de los informes microbiológicos. Además, dado que no todos los pacientes pueden ser examinados de forma periódica, la detección de un brote puede tardar días, semanas o incluso meses, dependiendo del patógeno [37].

Los conjuntos de datos clínicos son, por naturaleza, espaciales y temporales, siendo un desafío la exploración y correlación entre diferentes fuentes de datos clínicas [190]. Como resultado, se han desarrollado herramientas destinadas a sintetizar datos provenientes de fuentes heterogéneas, en las que el espacio y el tiempo desempeñan un papel fundamental, especialmente en el ámbito de la epidemiología. En este campo, el análisis de la información referente a las dimensiones espacial y temporal, así como sobre las relaciones y contactos entre pacientes es esencial para analizar la distribución de una infección.

Tradicionalmente, el análisis epidemiológico hospitalario se ha apoyado en técnicas estadísticas univariadas y multivariadas, así como en modelos espaciales básicos y métodos de detección de cambios [110, 205]. Si bien estas aproximaciones han demostrado utilidad, presentan limitaciones importantes a la hora de integrar la creciente cantidad de datos clínicos heterogéneos disponibles, así como de capturar las dinámicas espacio-temporales complejas de los brotes en entornos hospitalarios

[132, 165]. En muchos casos, estas técnicas no logran modelar adecuadamente aspectos clave como la proximidad física entre pacientes, el movimiento dentro del hospital o las secuencias temporales de eventos clínicos [64].

En este contexto, la incorporación de técnicas de modelado espacio-temporal y el uso de tecnologías basadas en grafos surge como una alternativa prometedora. Estas metodologías permiten representar de forma más expresiva y eficiente las interacciones entre los distintos elementos que intervienen en la propagación de infecciones: pacientes, profesionales, espacios físicos (habitaciones, plantas, servicios), y eventos clínicos [184, 185]. Los grafos, en particular, proporcionan una estructura natural para modelar relaciones complejas y permiten aplicar algoritmos avanzados de análisis, visualización y minería de patrones [1, 51]. La combinación de estos enfoques permite no solo detectar patrones de contagio, sino también analizar de forma explicativa y visual las trayectorias espacio-temporales de los brotes, facilitando así la toma de decisiones por parte de los profesionales sanitarios [8, 113].

Pese a estos avances, persisten desafíos importantes. La mayoría de los modelos existentes están diseñados para escenarios geográficos amplios, y no abordan adecuadamente la estructura interna de los edificios hospitalarios, donde se produce la interacción directa entre agentes [207]. Además, la integración entre razonamiento formal, análisis de datos masivos, y sistemas clínicos sigue siendo escasa [15, 56, 90]. Esto evidencia una falta de soluciones tecnológicas especializadas que respondan a las necesidades concretas de los profesionales de la salud en la vigilancia y gestión de infecciones hospitalarias.

En este contexto, esta tesis plantea explorar las posibilidades que ofrece la combinación del modelado espacio-temporal junto con la tecnología de grafos para el análisis de infecciones nosocomiales. El objetivo es desarrollar una base que sirva para la investigación, tanto en esta tesis como a futuro, en el análisis de la transmisión de infecciones nosocomiales, mediante herramientas que permitan el descubrimiento de nuevo conocimiento, como detección de brotes, rutas de contagio y fuentes de infección.

1.2 Hipótesis y objetivos

La hipótesis planteada en esta tesis doctoral es demostrar que es posible combinar de forma efectiva las tecnologías basadas en grafos con el análisis espacio-temporal para la resolución de tareas epidemiológicas en el contexto de la vigilancia de infecciones nosocomiales causadas por bacterias multirresistentes.

Para probar esta hipótesis nos planteamos como objetivo principal el diseño y formalización en forma de grafos de un modelo de datos y operacional que sirva de base para futuras investigaciones relacionadas con el análisis epidemiológico espacial y temporal basado en el movimientos y contactos entre pacientes.

De acuerdo con la hipótesis de partida, hemos descompuesto este objetivo principal en los siguientes objetivos:

Objetivo 1: Análisis bibliográfico y estudio del uso de grafos en los modelos computacionales para la simulación de brotes epidémicos, ahondando también en las características espaciales y temporales de las simulaciones.

Objetivo 2: Modelado espacio-temporal para el análisis de contactos entre pacientes orientado a la transmisión epidemiológica de infecciones nosocomiales.

Objetivo 3: Diseño y formalización de las tareas epidemiológicas fundamentales en la vigilancia y detección de brotes nosocomiales.

Objetivo 4: Estudio y selección de tecnologías basadas en grafos para el almacenamiento e inferencia de información. Además, implementación del modelo espacio-temporal y las tareas epidemiológicas en la tecnología seleccionada.

Objetivo 5: Cuantificación de la similitud espacio-temporal entre pacientes infectados a partir del análisis de contactos derivado de sus movimientos en el hospital.

Objetivo 6: Agrupamiento de pacientes infectados en base a su conexión espacio-temporal para la detección de posibles brotes y potenciales rutas de transmisión.

Objetivo 7: Una garantía de la reproducibilidad de la investigación realizada mediante repositorios de acceso público.

1.3 Estructura de la tesis

Esta tesis consta de 5 capítulos, los cuales contribuyen a alcanzar los objetivos definidos en la Sección 1.2. Concretamente, estos capítulos son los siguientes:

Capítulo 1: este capítulo expone el contexto del trabajo desarrollado, presentando los fundamentos de las distintas partes que conforman la investigación. Además, se presenta la hipótesis y los objetivos de investigación que se persigue conseguir en esta tesis. Finalmente, se muestra la organización del documento.

Capítulo 2: en este capítulo se realiza un estudio del estado del arte sobre el uso de redes en modelos computacionales para la simulación y análisis de la propagación espacial y temporal de brotes epidémicos. Para ello, se pone en perspectiva los avances realizados entre los años 2010 y 2021 considerando las características espaciales, temporales y epidemiológicas de las simulaciones, las características de las redes utilizadas y la granularidad de los datos empleados para la creación de la población en la simulación.

Capítulo 3: en este capítulo se propone un modelo de datos que sea capaz de representar los movimientos de los pacientes hospitalizados dentro del hospital a lo largo del tiempo, sirviendo de base para la investigación epidemiológica nosocomial. Para ello, el modelo se compone de una dimensión espacial, definida con una jerarquía de los distintos elementos arquitectónicos de la estructura de un hospital y con una jerarquía de la organización de los trabajadores sanitarios, y de una dimensión temporal, en la que se registran a modo de eventos todas las acciones que le suceden a cada paciente. Estos eventos son el nexo entre ambas dimensiones.

También se definen seis consultas que, basadas en este modelo de datos, representan diferentes tareas clínicas en la investigación epidemiológica: detección de brotes, análisis de contactos entre pacientes e identificación del caso índice de un brote.

Además, se estudian las diversas alternativas para implementar el modelo de datos propuesto como un grafo en el que se mantenga la semántica definida. También se evalúa la expresividad y el rendimiento de dos bases de datos orientadas a grafos de uso común en la literatura científica tanto a nivel de almacenamiento como de ejecución de algunas de las consultas epidemiológicas propuestas sobre dichos grafos.

Capítulo 4: en este capítulo se propone un nuevo método, método StESPT, para el descubrimiento de grupos de pacientes hospitalizados que hayan sido infectados durante un posible brote, los cuales estén estrechamente relacionados espacio-temporalmente. Estos grupos de pacientes pueden ayudar en determinar la existencia de un brote epidémico en el hospital y a descubrir cuáles son sus posibles fuentes de transmisión. Para ello, y en base al en el modelo de

datos propuesto en el Capítulo 3, los movimientos dentro del hospital de los pacientes infectados son transformados en trayectorias. A continuación, se evalúa la similitud epidemiológica espacio-temporal entre las trayectorias de cada par de pacientes. Esta similitud entre trayectorias se basa en la definición de similitud espacio-temporal entre puntos que hemos diseñado. Finalmente, se utiliza un algoritmo de *clustering* para dividir los pacientes en grupos en base a la similitud de sus trayectorias.

Capítulo 5: este capítulo presenta las conclusiones de esta investigación, el trabajo futuro y los artículos publicados como resultado de esta tesis.

1.3.1. Relación entre los objetivos y la estructura de la tesis

En el Capítulo 2, se aborda el Objetivo 1, ya que realizamos un estudio del arte profundo sobre el uso de modelos de simulación de brotes en los que se utilizan redes o grafos, caracterizando semántica y cuantitativamente dichas redes. Además, se analiza cuáles son las dimensiones y tiempo que abarca la simulación y cómo se modela la evolución epidemiológica de la infección.

En el Capítulo 3, se tratan los Objetivos 2, 3 y 4. El Objetivo 2 y parte del Objetivo 4 son cubiertos mediante el diseño, formalización en forma de grafo de conocimiento e implementación en tecnología RDF* de un modelo de datos que nos permita describir espacio-temporalmente los movimientos de los pacientes a través de un hospital, añadiendo también información sobre las pruebas microbiológicas realizadas. En cuanto al Objetivo 3 y la parte restante del 4, se han diseñado y formalizado en SPARQL* un conjunto de seis consultas que, basadas en el modelo de datos del Objetivo 2, permiten el análisis de contactos entre pacientes y detección de brotes. Estos tres objetivos, especialmente el Objetivo 2, sirven de base para la realización de los Objetivos 5 y 6, abordados en el Capítulo 4.

Esta tesis doctoral ha sido concebida siguiendo los principios de la ciencia abierta (en inglés, *open-science*), tal como se refleja en el Objetivo 7, que promueven el uso de código abierto, recursos accesibles y conjuntos de datos públicos con el fin de hacer que nuestra investigación sea reproducible. En este sentido, se prima el uso de librerías software y bases de datos de código abierto, así como un almacenamiento de los datos que posibilite la interoperabilidad con otros sistemas.

La Tabla 1.1 proporciona un resumen de esta relación entre los objetivos propuestos y los diferentes capítulos en los que se divide esta tesis.

TABLA 1.1
Relación entre los objetivos propuestos y los capítulos de esta tesis.

	Capítulo 2	**Capítulo 3**	**Capítulo 4**
Objetivo *1*	×		
Objetivo *2*		×	
Objetivo *3*		×	
Objetivo *4*		×	
Objetivo *5*			×
Objetivo *6*			×
Objetivo *7*		×	×

Capítulo 2

Estado del arte del uso de redes en modelos computacionales para la propagación espacial y temporal de brotes en epidemiología

En este capítulo, hacemos un análisis de los modelos computacionales utilizados para la simulación y análisis de brotes epidémicos, más concretamente, en aquellos que usan redes para modelar algún aspecto relacionado con la propagación del brote. Para ello, hemos definido cinco preguntas que guían la investigación y hemos realizado una búsqueda exhaustiva en la literatura científica de los últimos años con el fin de darles respuesta. Podemos resumir las contribuciones de este análisis del estado en las siguientes: (1) identificación de los primeros modelos computacionales y tipos de redes desde un punto de vista semántico; (2) análisis de tendencias del uso combinado de ambos en función de la escala espacial y temporal que abarca la simulación; (3) análisis de los modelos epidemiológicos compartimentales con los que se describen las dinámicas de las infecciones; y (4) análisis de las fuentes de datos utilizadas para la validación de los modelos, centrándonos en su granularidad y disponibilidad. Los resultados de los estudios epidémicos suelen tener aplicación en el campo de la salud pública, definida como el conjunto de estrategias, intervenciones y políticas, tanto sanitarias como transversales, orientadas a la prevención de enfermedades y la promoción y recuperación de la salud de las personas, tanto a nivel individual como colectivo.

2.1 Introducción

El rápido avance de la medicina y la tecnología ha permitido cierto control sobre las epidemias (campañas de vacunación, tratamientos con antibióticos, mejoras en la higiene y las infraestructuras sanitarias). Sin embargo, la aparición de nuevos agentes infecciosos sigue representando un peligro constante, prueba de ello es que los brotes

de enfermedades infecciosas se han triplicado desde la década de 1980 [196]. Algunos factores sociales que favorecen este fenómeno pueden ser el crecimiento urbano masivo, la globalización y la movilidad internacional. Además, aún no se han desarrollado vacunas efectivas para algunas de las infecciones más comunes, y las poblaciones de organismos resistentes a antibióticos y antivirales están aumentando [216]. Todos estos factores contribuyen a la necesidad de estudiar modelos con los que predecir cómo se propagarán los brotes futuros y qué estrategias de intervención serían más eficientes dependiendo de la enfermedad y las condiciones de la población. Por tanto, estos modelos deben considerar tanto las dinámicas de transmisión de los patógenos como el contexto social.

En este capítulo, presentamos una visión general actualizada sobre las tendencias en modelos computacionales para la representación de la propagación de brotes infecciosos por contacto cercano en una población. Estos modelos pueden emplear técnicas y herramientas de distintas disciplinas, como las matemáticas, la estadística o la inteligencia artificial.

Hemos dedicado especial atención a los modelos que utilizan redes (en este contexto también pueden referirse con el nombre de grafos) para la representación de aspectos relacionados con la transmisión de la infección. Los brotes epidémicos son un claro ejemplo de sistemas complejos en los que se puede explotar el potencial de la teoría de redes. Las conexiones entre individuos durante un brote pueden definirse en una red en la que cada persona es un nodo, y las aristas representan situaciones en las que dos nodos han estado o probablemente estarán en contacto cercano. Estas redes permiten estudiar las posibles rutas de transmisión con el fin de mejorar la predicción de la evolución de la infección. Además, la dinámica y las fases de la enfermedad, representadas habitualmente mediante modelos compartimentales, también pueden analizarse a través de redes [49].

Este estudio se ha llevado a cabo mediante una revisión sistemática. Las contribuciones de esta revisión son las siguientes.

- Se enfatiza el uso de redes como herramientas. El modelado con grafos es actualmente un tema altamente activo, lo que queda demostrado por su amplio uso en diversos campos, como el modelado de propagación de información, sistemas de recomendación, inteligencia de negocio y predicción de enlaces, entre otros [67]. La representación mediante grafos, combinada con lógica, proporciona una base matemática para analizar, comprender y extraer conocimiento de sistemas complejos del mundo real [78].

- Se detallan las escalas espaciales y temporales utilizadas para analizar los datos. Los modelos y su funcionalidad pueden diferir dependiendo del tamaño del área de estudio: un edificio, una ciudad o un país. Asimismo, el análisis varía en función de la frecuencia con la que se recopilan los datos, ya sea diariamente o con una resolución temporal más detallada.

- Se ofrece un análisis de las fuentes de datos y su granularidad. La granularidad de los datos se refiere a si representan estadísticas agregadas para una población o información individualizada. La disponibilidad y el nivel de detalle de los datos tienen un impacto directo en la precisión y aplicabilidad de los modelos.

El resto del capítulo se estructura como sigue: en la sección 2.2 presentamos la metodología empleada para la realización de este capítulo, destacando las preguntas definidas para guiar el estudio (sección 2.2.1) y las consultas en las bases de datos (sección 2.2.2). En la sección 2.3 presentamos los resultados obtenidos, cuya discusión se encuentra en la sección 2.4. Finalmente, en la sección 2.5 exponemos las conclusiones obtenidas.

2.2 Metodología

El objetivo de este capítulo es el de obtener un estado del arte completo y coherente en relación del uso de redes en estudios epidemiológicos, centrándonos en aquellos que se enfocan en la propagación espacial y temporal de brotes sobre una población. Para ello, hemos llevado a cabo una revisión sistemática adoptando las directrices propuestas en *PRISMA 2009 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses. En español, Elementos de Informe Preferibles para Revisiones Sistemáticas y Meta-análisis)* [120, 143], junto con las sugerencias de Peterson et al. para la investigación bibliográfica en ingeniería del software [163].

En primer lugar, explicamos las preguntas definidas para guiar este estudio, así como las consultas realizadas para la búsqueda en las bases de datos. A continuación, mostramos los criterios definidos para la inclusión y exclusión de artículos en esta revisión.

2.2.1. Preguntas de investigación

Para guiar la revisión sistemática, definimos cinco preguntas de investigación:

- **PI1. ¿Cuáles son los modelos epidemiológicos más utilizados para estudiar la propagación de brotes (dentro del contexto de nuestro estudio)?**

  Hemos identificado qué modelos epidemiológicos se emplean para guiar la evolución de las simulaciones, analizando si es un tema de investigación recurrente en la actualidad.

- **PI2. ¿Cuáles son los principales modelos computacionales utilizados en la epidemiología espacio-temporal?**

  Dado que el modelado y simulación de la propagación de brotes epidémicos puede abordarse desde diferentes áreas de estudio (matemáticas, estadística, inteligencia artificial), hemos identificado cuáles son los enfoques más utilizados.

- **PI3. ¿Es común el uso de redes en el análisis epidemiológico espacio-temporal? ¿Cuáles son los tipos de redes más utilizados?**

  Hemos estudiado cómo se utilizan las redes en el modelado epidemiológico, centrándonos en los conceptos que representan sus nodos y aristas.

- **PI4. ¿Qué escalas espaciales y temporales se utilizan para analizar los brotes en este tipo de estudios?**

  Los brotes epidémicos pueden ocurrir tanto en el interior de un edificio como en áreas exteriores de diferente escala, como ciudades o países, por lo que hemos estudiado cuáles de estos casos son los más comunes de modelar. También hemos estudiado la granularidad temporal utilizada y si existe alguna relación entre las unidades de ambas dimensiones.

- **PI5. ¿Cuáles son las fuentes de datos usadas es este tipo de estudios? ¿Cuál es la granularidad de los datos?**

  Hemos identificado cuáles son los tipos de fuentes de datos más utilizadas para el desarrollo y evaluación de cada estudio. Además, hemos analizado su granularidad, es decir, si se componen de datos agregados o individuales.

2.2.2. Bases de datos y consultas bibliográficas

Para poder responder a las preguntas de investigación definidas en el apartado anterior, realizamos una búsqueda de artículos en cuatro bases de datos bibliográficas que artículos relacionados con ingeniería, informática y salud: ACM Digital Library, IEEE Xplore, PubMed y Scopus. Restringimos nuestra búsqueda a artículos de revistas y de conferencias, en inglés y que hubieran sido publicados entre los años 2010 y 2021 (año en el que se realiza la investigación asociada a este capítulo).

La definición de las cadenas de búsqueda se hace en base a la selección de un conjunto de palabras clave, cuya identificación se realiza en base al *Método PICO (Problema, Intervención, Comparación, Resultado* [en inglés, *Outcome*]*)*. En la Tabla 2.1, presentamos las palabras clave.

A partir de estas palabras clave construimos una cadena de búsqueda cuyo sentido general sería el siguiente:

"Nos gustaría encontrar aquellos artículos relacionados con la epidemiología humana (con especial énfasis en los brotes nosocomiales) que se centrasen en el razonamiento, modelado o simulación espacio-temporal (también solo espacial o temporal) utilizando redes (también denominadas grafos) como herramienta."

Las consultas creadas para cada base de datos se muestran en la Tabla 2.2

Tabla 2.1
Palabras clave para la búsqueda bibliográfica, en inglés y español (entre paréntesis).

Ámbito	**Palabras clave**
Problema	epidemic (epidemia), epidemiologic (epidemiológico), outbreak (brote), nosocomial (nosocomial), infectious (infeccioso)
Intervención	network (red), graph (grafo), spatiotemporal (espacio-temporal), spatial (espacial), temporal (temporal)
Comparación	simulation (simulación), simulate (simular), model (modelo, modelar), reason (razonar)

Tabla 2.2
Consulta para cada base de datos bibliográfica.

Base de datos	**Cadena de búsqueda**
ACM Digital Library	`(Title:(epidemic) OR Title:(epidemiologic) OR Title:(outbreak) OR Title:(infectious) OR Title:(nosocomial)) AND ((Title:(network) OR Title:(graph) OR Title:(spatial) OR Title:(temporal) OR Title:(spatiotemporal)) OR (Abstract:(network) OR Abstract:(graph) OR Abstract:(spatial) OR Abstract:(temporal) OR Abstract:(spatiotemporal))) AND ((Title:(simulation) OR Title:(simulate) OR Title:(model) OR Title:(reason)) OR (Abstract:(simulation) OR Abstract:(simulate) OR Abstract:(model) OR Abstract:(reason)))`
IEEE Xplore	`(\Document Title":epidemic OR \Document Title":epidemiologic OR \Document Title":outbreak OR \Document Title":infectious OR \Document Title":nosocomial) AND (\Document Title":network OR \Document Title":graph OR \Document Title":spatial OR \Document Title":temporal OR \Document Title":spatiotemporal) AND ((\Document Title":simulation OR \Document Title":simulate OR \Document Title":model OR \Document Title":reason) OR (\Abstract":simulation OR \Abstract":simulate OR \Abstract":model OR \Abstract":reason))`

PubMed	`(epidemic[Title] OR epidemiologic[Title] OR outbreak[Title] OR infectious[Title] OR nosocomial[Title]) AND (network[Title/Abstract] OR graph[Title/Abstract] OR spatial[Title/Abstract] OR temporal[Title/Abstract] OR spatiotemporal[Title/Abstract]) AND (simulation[Title/Abstract] OR simulate[Title/Abstract] OR model[Title/Abstract] OR reason[Title/Abstract])`
Scopus	`(TITLE(epidemic) OR TITLE(epidemiologic) OR TITLE(outbreak) OR TITLE(infectious) OR TITLE(nosocomial)) AND (TITLE(network) OR TITLE(graph) OR TITLE(spatial) OR TITLE(temporal) OR TITLE(spatiotemporal)) AND (TITLE-ABS-KEY(simulation) OR TITLE-ABS-KEY(simulate) OR TITLE-ABS-KEY(model) OR TITLE-ABS-KEY(reason)) AND PUBYEAR>2010 AND PUBYEAR<2022 AND (LIMIT-TO(SUBJAREA,"COMP")) AND (LIMIT-TO(DOCTYPE,"ar") OR LIMIT-TO(DOCTYPE,"cp")) AND (LIMIT-TO(LANGUAGE,"English")) AND (LIMIT-TO(SRCTYPE,"j") OR LIMIT-TO (SRCTYPE,"p"))`

2.2.3. Selección de artículos

Como resultado de las búsquedas realizadas hemos obtenido un total de **832** artículos. A continuación, hemos realizado dos fases de revisión de los artículos para su descarte o selección en base a unos criterios definidos. En la primera fase, hemos leído el resumen de los artículos, descartándose **640** artículos. En la siguiente fase, hemos leído el texto completo de los **192** artículos restantes, descartándose **80** y quedando **112** artículos para la extracción de información.

A continuación, se muestran los criterios de inclusión y exclusión que se han definido para las dos fases de revisión de los artículos. Hemos establecido como política de selección que se incluyan aquellos artículos que cumplan con al menos uno de los criterios de inclusión y ninguno de los criterios de exclusión.

Los criterios de inclusión son los siguientes:

- **CI1**: Artículos en los que se modela o simula la propagación espacial de un brote epidémico real.
- **CI2**: Artículos en los que se simula la propagación de un brote epidémico sintético mediante modelos computacionales (basados en disciplinas como las matemáticas o la inteligencia artificial).

Los criterios de exclusión son los siguientes:

- **CE1**: Artículos cuya extensión (excluyendo referencias y apéndices) fuese inferior a seis páginas. Se buscan estudios con una descripción completa de la metodología y los resultados, por lo que artículos cortos, como resúmenes extendidos o ponencias en conferencias, no cumplen con este requisito.

- **CE2**: Artículos sobre el uso de modelos epidemiológicos en otros contextos, como la difusión de *malware* en redes de telecomunicaciones o noticias falsas en redes sociales.
- **CE3**: Artículos que utilizan algoritmos de grafos en contextos no relacionados con la epidemiología.
- **CE4**: Artículos sobre epizootias (epidemias sobre una población animal, no humana).
- **CE5**: Artículos donde no se modelan ni simulan enfermedades cuya transmisión es "persona a persona" a través del aire o el contacto físico o con una superficie contaminada. Por tanto, para este estudio se excluyen los artículos sobre zoonosis y enfermedades transmitidas por vectores biológicos o agua. Tampoco se incluyen enfermedades de transmisión sexual.
- **CE6**: Artículos centrados en el análisis retrospectivo de brotes epidémicos en una región, o en la visualización de estadísticas relacionadas con su evolución.
- **CE7**: Artículos centrados en el uso de series temporales (por ejemplo, datos de incidencia o datos de localización) para modelar el pasado (por ejemplo, cadenas de infección o redes de contacto).
- **CE8**: Artículos en los que el modelado de los brotes no es el tema principal, sino una herramienta secundaria para el análisis de otros aspectos. Por ejemplo, para la optimización de recursos (vacunas, medicamentos, personal sanitario) durante un brote.

2.3 Resultados

Tras el proceso de búsqueda y selección de artículos, hemos obtenido un total de 112 artículos para revisar y extraer información relacionada con las preguntas de investigación. A continuación, se muestran los resultados de las preguntas de investigación PI1 a PI5.

2.3.1. Visión general de las características de los resultados

Si realizamos un análisis temporal de los resultados obtenidos (ver Figura 2.1), podemos observar que en los últimos dos años del período de estudio ha habido un aumento en la cantidad de artículos centrados en el modelado y simulación de la propagación de brotes. Concretamente, en 2020 y 2021 se publicaron 22 y 16 artículos respectivamente, en comparación con los años 2010 a 2019, donde las publicaciones anuales oscilaron entre 4 y 11.

Este hecho podría relacionarse con la aparición de la epidemia de COVID-19. De hecho, si nos fijamos en las enfermedades modeladas en los artículos, el COVID-19 es la más recurrente **(19/112)** (ver Figura 2.2). Las siguientes enfermedades más estudiadas son la gripe estacional **(15/112)** y la gripe A/H1N1 **(12/112)**. Otras enfermedades estudiadas, pero en menor medida, son el ébola **(3/112)**, SARS-CoV **(2/112)**, MRSA **(3/112)**, la enfermedad de mano-pie-boca (HFMD por sus siglas en inglés, *Hand-Foot-Mouth Disease*) **(1/112)**, y la tuberculosis **(1/112)**. En total, el 50 % **(56/112)** de los artículos se centran en una enfermedad en específico, mientras que en el resto se crea un modelo genérico con parámetros ajustables.

En cuanto a la finalidad de los estudios, encontramos que la mayoría de los estudios se enfocan en estudiar cómo se propaga un brote bajo diferentes escenarios (estadísticas demográficas, comportamiento de la población, aplicación de vacunas y otras medidas para evitar el contagio) y dinámicas del brote (tasa de transmisión, tasa de recuperación). También hay estudios cuyo objetivo principal es el análisis de un brote real.

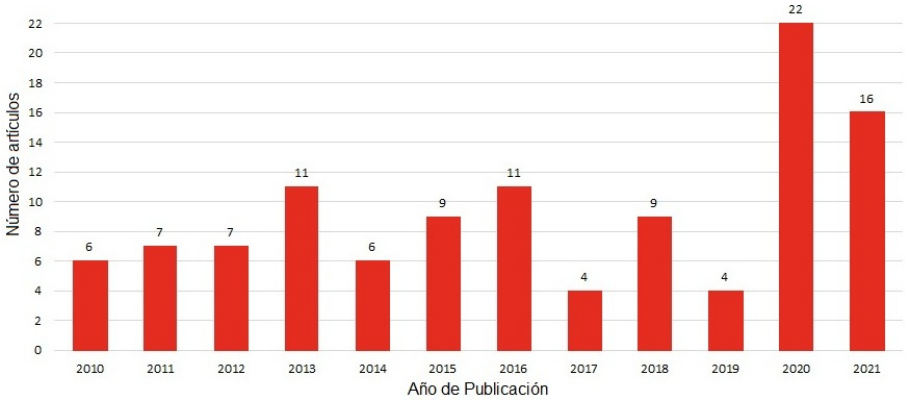


FIGURA 2.1. Número de artículos por año de publicación.

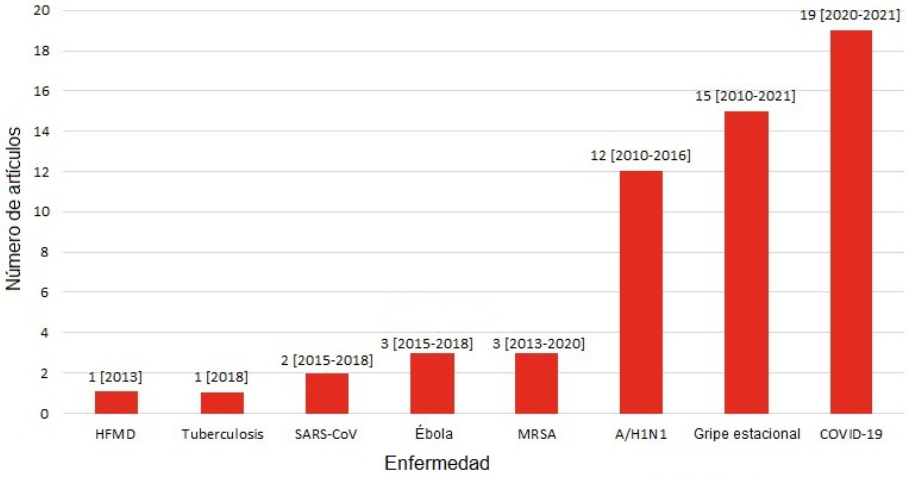


FIGURA 2.2. Número de artículos por enfermedad estudiada. También se muestra entre corchetes los años entre los que se publicaron los artículos sobre cada enfermedad.

2.3.2. PI1: ¿Cuáles son los modelos epidemiológicos más utilizados para estudiar la propagación de brotes (dentro del contexto de nuestro estudio)?

En el contexto de este estudio, vamos a definir los modelos epidemiológicos como modelos compartimentales. Estos son modelos matemáticos con los que se puede representar de forma simplificadas la evolución de un brote o epidemia sobre una población. En ellos, la población es dividida en compartimentos etiquetados o estados que representan las diferentes fases de una enfermedad infecciosa. Estos compartimentos son mutuamente excluyentes y los individuos pueden cambiar entre ellos, estando definidos en el modelo hacia qué estados se puede mover a partir de un determinado estado y con qué tasas de transmisión. Una representación gráfica habitual de estos modelos es en forma de grafo, donde los nodos representan los compartimentos y las aristas las posibles transiciones, estando etiquetadas con la tasa de transmisión entre el par de estados que conectan.

Para proporcionar una descripción más detallada de los modelos compartimentales epidemiológicos, en la Figura 2.3 se muestra el esquema de algunos de los modelos más citados por los artículos revisados. Las figuras 2.3.a, 2.3.b y 2.3.c representan tres modelos que hemos denominado como "originales", es decir, que están formados por estados comunes en la evolución de una enfermedad infecciosa y que son de uso común en la literatura desde décadas. A continuación, presentamos una descripción de cada modelo:

- **SIS** (*Susceptible-Infectado-Susceptible*). En este modelo, un individuo en el estado *Susceptible*, es decir, sano, puede pasar a *Infectado* con una cierta *tasa de transmisión*. Por su parte, un individuo *Infectado* puede recuperarse y volver al estado *Susceptible* con una *tasa de recuperación*, completando así un ciclo. Se ha de puntualizar que el estado *Infectado* representa que el individuo ha contraído la enfermedad y que es capaz de contagiar a otros individuos que sean *Susceptibles*.

- **SIR** (*Susceptible-Infectado-Recuperado*). Este modelo se diferencia del anterior en que un individuo *Infectado* no puede volver al estado de *Susceptible*, sino que tras un *período de infección* (o con una *tasa de recuperación*), pasa al estado *Recuperado*, permaneciendo inmune hasta el final del brote.

- **SEIR** (*Susceptible-Expuesto-Infectado-Recuperado*). Este modelo es una extensión de SIR, en la que se añade un nuevo estado *Expuesto* (también llamado *Latente*), que precede al estado *Infectado* para representar un período de latencia, durante el cual el individuo ya ha contraído la enfermedad, pero aún no es contagioso. Además, en algunos estudios se considera que el individuo es asintomático durante este estado.

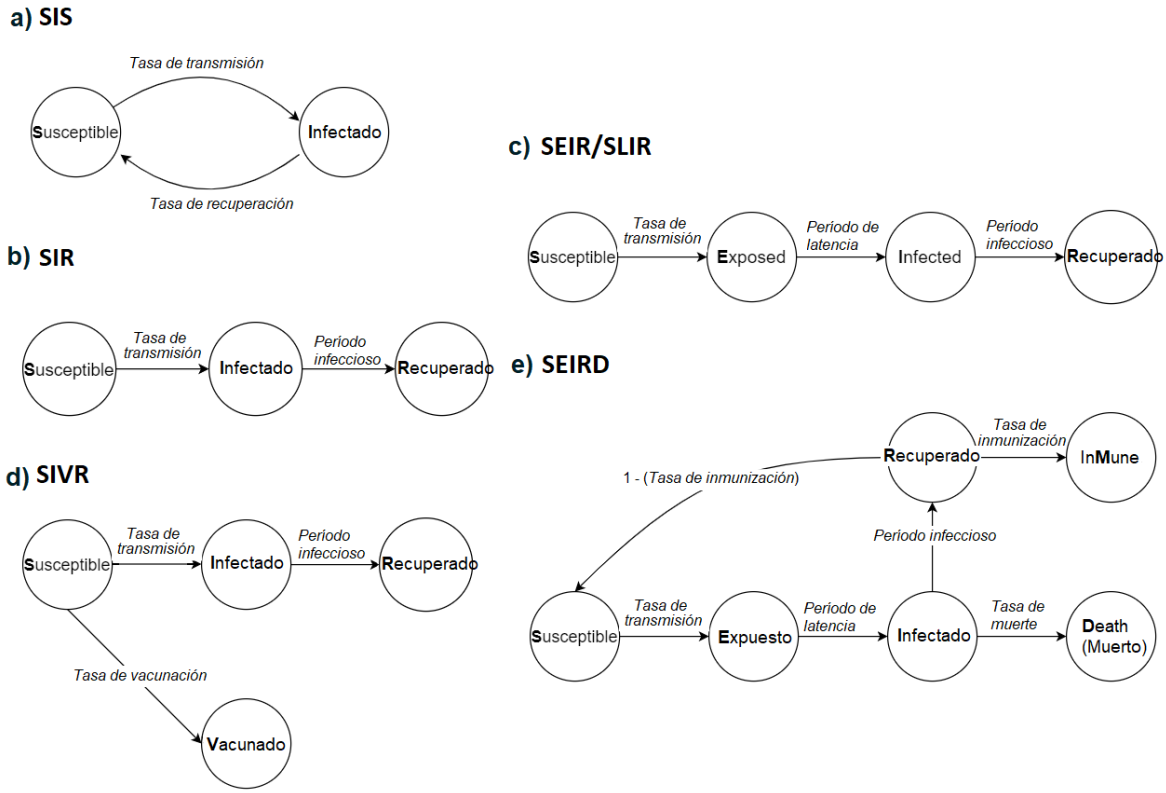


FIGURA 2.3. Esquema de algunos modelos compartimentales epidemiológicos encontrados en la investigación.

En las figuras 2.3.d y 2.3.e se representan dos variantes de los modelos SIR y SEIR. Hemos definido como "variantes" a aquellos modelos diseñados a partir de uno "original" mediante la adición de nuevos estados o la subdivisión de un estado existente en múltiples subestados.

- **SIVR** (*Susceptible-Infectado-Vacunado-Recuperado*). Este modelo es una variante de SIR donde se añade un nuevo estado *Vacunado*. Este estado permite representar que hay individuos *Susceptibles* que deciden vacunarse con una cierta *tasa de vacunación*, quedando inmunizados durante el brote.

- **SEIRD** (*Susceptible-Expuesto-Infectado-Recuperado-Fallecido* (del inglés, *Death*)). Este modelo es una variante de SEIR en la que se permiten ciclos, ya que el estado *Recuperado* ahora es un estado transitorio, desde el cual el individuo puede regresar al estado *Susceptible* o bien ser *Inmune* durante el resto del brote. Este modelo también introduce un nuevo estado para modelar que algunos individuos pueden fallecer al ser *Infectados*.

Tabla 2.3
Artículos por modelo epidemiológico compartimental.

Modelo Epidemiológico			**Variante**		
Modelo	**Artículos**	**Nº artículos**	**Variante**	**Artículos**	**Nº artículos**
SI	[19, 234]	2	-		
SIS	[2, 11, 38, 79, 84, 157, 158, 166, 181, 188, 224, 228]	12	*SIVS*	[92, 162]	2
			SIRS	[237]	1
			SIRVS	[195]	1
			SIS/AU	[73, 94, 180, 182, 239]	5
			SIRS/OIA	[206]	1
SIR	[10, 12, 17, 18, 27, 43, 46, 59, 61, 71, 99, 100, 112, 115, 116, 123, 125, 131, 133, 137, 138, 139, 146, 153, 174, 177, 179, 208, 210, 213, 226, 230, 235]	33	*SIVR*	[7, 103]	2
			SIRD	[3]	1
			SITR	[58]	1
			SIR/AU	[122, 236, 240]	3
SEIR	[5, 16, 33, 35, 36, 40, 42, 60, 63, 66, 72, 80, 86, 88, 105, 106, 111, 121, 124, 127, 128, 129, 130, 141, 161, 194, 203, 204, 225, 229, 233]	31	*SEIRD*	[74, 75, 95, 108, 119, 178, 219]	7
			SEIRDM	[68]	1
			SEIRDV	[104]	1
			SEIQR	[118]	1
			G-SEIV	[154, 155]	2
			SEIR/AU	[200]	1
SEI	[175]	1	-		

Modelos Ad-hoc		
Enfermedad	**Artículos**	**Nº artículos**
Para A/H1N1	[231]	1
Para COVID-19	[65]	1
Para Ébola	[136]	1

Hemos comprobado que todos los artículos revisados utilizan un modelo epidemiológico compartimental, siendo los denominados "originales" los más usados con un **73,2 % (82/112)** del total. De ellos, los modelos epidemiológicos más utilizados son SEIR y SIR, presentes en el **39,3 % (44/112)** y el **35,7 % (40/112)** de los artículos, respectivamente. El tercer modelo más empleado es SIS, con un 19,64 % (22/112) del total. Algunos estudios optan por diseñar su propio modelo epidemiológico *ad-hoc* para la enfermedad abordada, añadiendo nuevos estados específicos y relaciones que permiten una descripción más detallada de las fases de la enfermedad. La Tabla 2.3 muestra la relación entre cada artículo y el modelo epidemiológico utilizado, distinguiendo entre modelos "originales" y "variantes".

Cabe mencionar que la elección del modelo epidemiológico ha de basarse en la enfermedad a estudiar y el período de tiempo a simular. En la mayoría de los artículos revisados (independientemente de que traten una enfermedad concreta o no) se quiere modelar brotes de enfermedades respiratorias con un período de incubación o fase asintomática, tras el cual la persona enferma adquiere inmunidad durante el resto del brote. Los modelos SEIR y SIR, junto con sus variantes, han sido los más utilizados, como también podemos apreciar en la figura 2.4. SIR se emplea cuando el estado *Expuesto* es demasiado corto o se considera que los individuos pueden transmitir la enfermedad de igual modo independientemente de que tengan o no síntomas de la enfermedad.

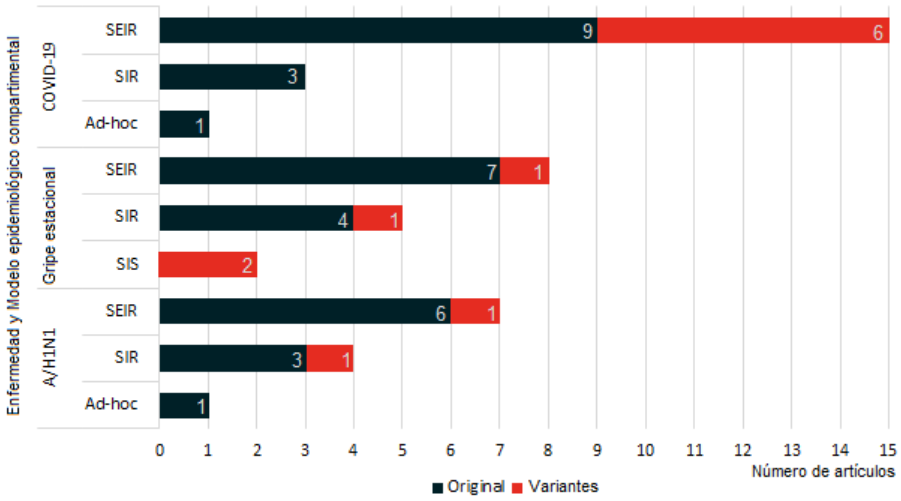


FIGURA 2.4. Número de artículos por enfermedad y modelo epidemiológico compartimental.

2.3.3. PI2: ¿Cuáles son los principales modelos computacionales usados en la epidemiología espacio-temporal?

Tras la revisión de los artículos seleccionados, podemos clasificar los modelos computacionales utilizados para la representación y simulación de la propagación de brotes epidémicos en tres grandes categorías: modelos deterministas, modelos estocásticos y modelos basados en agentes (MBA). En la Tabla 2.4 se muestra cada uno de los artículos revisados clasificados en una de las tres categorías definidas.

Una característica común de las dos primeras categorías, *modelos deterministas* y *modelos estocásticos*, es que ambos representan una aplicación directa de los modelos compartimentales epidemiológicos.

En el caso de los **modelos deterministas**, utilizados en el **25,89 % (29/112)** de los artículos, el modelo compartimental epidemiológico es descrito mediante un conjunto de ecuaciones diferenciales ordinarias (una para cada estado). Estos modelos asumen que los cambios en los tamaños de las poblaciones en cada estado son continuos en el tiempo, siendo el objetivo del modelo el cálculo del número de individuos en cada compartimento por cada paso de la simulación. Dado su componente determinista (dado un escenario definido con unos valores fijos para sus parámetros -*tasa de transmisión*, *tasa de recuperación*, etc., según el modelo-, múltiples ejecuciones sobre el mismo escenario generan siempre el mismo resultado), estos modelos se utilizan principalmente para la predicción de la evolución de la población, así como otros parámetros derivados de esta (por ejemplo, tasa de incidencia, número básico de reproducción y tasa de reproducción efectiva). Además, estos modelos han sido los más utilizados por aquellos estudios donde se quieren estudiar aspectos que pueden influir en la evolución del brote, como puede ser el comportamiento de la población o la aplicación de medidas preventivas del contagio.

Los **modelos estocásticos**, utilizados en el **55,36 % (62/112)** de los artículos, definen el modelo epidemiológico compartimental como un conjunto de ecuaciones diferenciales estocásticas. Estas ecuaciones comprenden un conjunto de variables aleatorias con las que se describe la probabilidad del cambio de estado de los individuos a lo largo del tiempo (*probabilidad de contagio*, *probabilidad de recuperación*, *probabilidad de infeccioso*, *probabilidad de muerte*, etc., según el modelo). Ha de indicarse que, en estos modelos, el estado de un individuo depende únicamente del estado inmediatamente anterior (cadena de Markov). Dada la aleatoriedad de los resultados, el enfoque de los estudios que utilizan modelos estocásticos es el siguiente: dado un escenario, el modelo se ejecuta varias veces, obteniendo un rango de resultados que deben ser analizados para obtener conclusiones sobre el comportamiento del brote. En el 5,36 % (6/112) de los artículos se indica que se ha seguido el método de Montecarlo. Los modelos estocásticos son utilizados para analizar la influencia de cada parámetro del modelo en la propagación del brote o para ajustar los valores de dichos parámetros para que los resultados se asemejen a un brote real.

La última categoría que hemos definido, **modelos basados en agentes** (**MBA**), proviene del campo de la inteligencia artificial (IA) y cuenta con el **18,75 % (21/112)** de los artículos. Un MBA comprende un conjunto de individuos autónomos (agentes) que, generalmente, pueden evolucionar y son capaces de interactuar entre sí y con su entorno. En el caso de los estudios revisados, los agentes representan a los individuos de la población sobre la que se propaga un brote.

Tabla 2.4
Número de artículos por modelo computacional.

Modelo computacional	**Artículos**	**Nº artículos**
Modelo estocástico	[2, 5, 16, 18, 27, 33, 38, 42, 43, 46, 58, 59, 61, 65, 71, 72, 73, 84, 88, 92, 94, 100, 104, 106, 112, 118, 119, 121, 122, 123, 127, 131, 133, 137, 138, 139, 141, 146, 154, 155, 157, 158, 161, 174, 175, 177, 182, 203, 206, 208, 224, 225, 226, 228, 229, 230, 233, 234, 236, 237, 239, 240]	62
Modelo determinista	[3, 10, 12, 17, 35, 60, 79, 80, 99, 103, 105, 115, 116, 125, 128, 153, 162, 166, 178, 180, 181, 188, 194, 200, 204, 210, 213, 219, 235]	29
Modelo basado en agentes	[7, 11, 19, 36, 40, 63, 66, 68, 74, 75, 86, 95, 108, 111, 124, 130, 136, 179, 195, 231]	21

2.3.4. PI3: ¿Es común el uso de redes en el análisis epidemiológico espacio-temporal? ¿Cuáles son los tipos de redes más utilizados?

En el contexto de esta investigación, definiremos las redes como estructuras de datos que son representadas como un grafo. De manera formal, un grafo G se define como un par $G = (N, A)$, donde:

- **N** es un conjunto de *nodos*, también llamados *vértices*. Los nodos representan conceptos u objetos indivisibles entre los que se quiere definir relaciones.
- **A** es un conjunto de *aristas*. Cada arista se define como un par (p, q), donde p y q son dos nodos distintos pertenecientes a N. Cada arista representa que existe una relación entre los dos nodos que comprende. Estas aristas pueden representar relaciones simétricas, donde el par (p, q) tiene el mismo significado que el par (q, p), o bien ser *dirigidas* y tener un sentido definido. También se puede definir una función de etiquetado sobre A para asignarle a las aristas una etiqueta que tenga un significado en el dominio asociado.

Las redes, o grafos, pueden clasificarse tanto por las características de su estructura como por su semántica, es decir, según qué representan sus nodos y aristas. En el caso de este estudio, hemos optado por abordar un enfoque semántico, dado que este está más enfocado en la finalidad de las redes dentro de los modelos con los que se representan los brotes epidémicos. De este modo, hemos clasificado las redes en tres categorías: redes de contacto, redes de relaciones y redes de metapoblaciones. También distinguimos un tipo especial de red debido a su estructura: la red multicapa. La Tabla 2.5 muestra un resumen con los artículos clasificados según el tipo de red utilizado.

Tabla 2.5
Número de artículos por tipo de red utilizada.

Tipo de red	**Artículos**	**Nº artículos**
Red de relaciones	[5, 27, 33, 43, 60, 63, 72, 80, 86, 92, 95, 99, 100, 103, 112, 115, 129, 131, 138, 146, 153, 154, 155, 162, 179, 180, 181, 188, 206, 210, 224, 228, 229, 230, 237]	36
Red de contactos	[7, 11, 36, 38, 46, 58, 84, 116, 119, 124, 125, 127, 141, 174, 175, 194, 233, 234, 239]	19
Red multicapa	[18, 59, 73, 94, 108, 122, 130, 157, 158, 182, 236, 240]	12
Red de metapoblación	[2, 12, 16, 17, 42, 61, 65, 68, 71, 79, 88, 104, 121, 133, 137, 161, 166, 177, 203, 204, 208, 213, 219, 225, 226]	26
No usa redes	[3, 10, 19, 35, 40, 66, 74, 75, 105, 106, 111, 118, 128, 136, 139, 178, 195, 200, 231]	19

Las **redes de relaciones** y las **redes de contactos** han sido los dos tipos de redes más utilizados con un **32,14 % (36/112)** y **16,96 % (19/112)** de los artículos, respectivamente. Si bien en ambos tipos de redes los nodos representan los individuos que conforman la población sobre la que se va a simular la propagación de un brote epidémico, el significado de sus aristas cambia. En las *redes de contacto*, cada arista simboliza que ha habido un contacto entre las dos personas representadas por los nodos que conecta, es decir, que ambas personas han estado físicamente lo suficientemente cerca como para que hubiese sido posible el contagio. En el caso de las redes de relaciones, las aristas tienen un significado más amplio: existe una conexión social entre dos personas, como pueden ser los lazos familiares, de amistad, de compañerismo en el trabajo o de vecindad. En este tipo de redes, no hay representación explícita de los contactos entre las personas, sino que, al haber una conexión social entre dos personas, es posible que entre ellas pueda darse un contacto en algún momento.

Las *redes de contactos* suelen utilizarse en estudios en los que para cada paso de la simulación se crea una nueva red para representar los contactos que se han producido durante dicho paso. Normalmente, los nuevos contagios que se dan en cada paso dependen únicamente de los contactos ocurridos durante ese paso y no tienen en cuenta los contactos de los pasos anteriores. En el caso de las *redes de relaciones*, se suele utilizar una única red que no cambia durante toda la simulación.

Un tipo particular de las redes de relaciones son las **redes multicapa**, cuya estructura consiste en dos grafos (las capas) que comparten los mismos nodos, pero no las aristas. Los nodos siguen representando a personas en ambos grafos, pero el significado de las aristas es diferente en cada capa. La primera capa se correspondería con una red de relaciones como las descritas y recibe el nombre de "capa de propagación del brote" (en inglés, *outbreak propagation layer* (*OPL*)). Mientras que, en la segunda capa, llamada "capa de propagación de la información" (en inglés, *information propagation layer* (*IPL*)), las aristas indican que ha habido un contacto virtual entre las personas, como pueden ser las llamadas telefónicas, mensajes de texto o chats en línea. La figura 2.5 muestra un ejemplo de una red multicapa. Estas redes se utilizan en estudios donde también se quiere modelar el comportamiento de los individuos en función de la cantidad de información que disponen sobre el brote. Aquellos pares de nodos que tengan una arista en la IPL, aunque tengan o no un contacto físico (esto depende de si tienen o no una arista en la OPL), sí mantendrán conversaciones, por lo que es posible que hablen entre sí sobre el brote. Esto puede influir en su comportamiento, por ejemplo, generándoles miedo y haciendo que sean más precavidos para salir de casa. Este tipo de redes han sido usadas en el **10,71 % (12/112)** de los artículos.

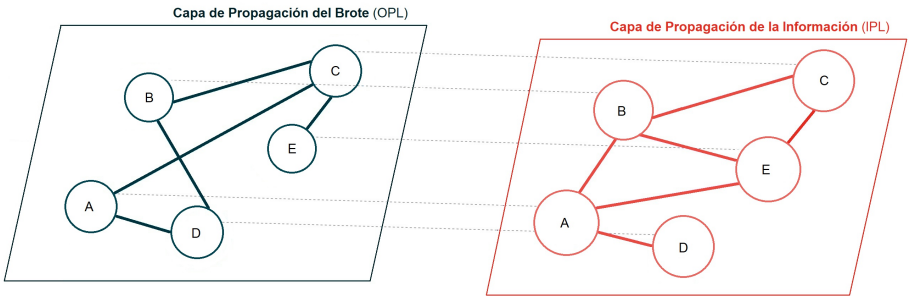


FIGURA 2.5. Ejemplo de una red multicapa.

Como último tipo de red tenemos las **redes de metapoblaciones**. Introducidas por Richard Levins [114] en 1969 para describir la dinámica poblacional de las plagas de insectos en campos agrícolas, la finalidad de estas redes se ha ampliado: se utilizan para estudiar la dinámica poblacional (tamaño y composición por edad) de varias poblaciones espacialmente separadas de la misma especie animal y cómo se relacionan entre sí. En el caso de los estudios revisados, los nodos representan una ubicación (por ejemplo, una ciudad o una región) y las aristas representan el movimiento de personas entre ellas (por ejemplo, por motivos laborales, por transporte de mercancías o vacacionales). Este tipo de redes son utilizadas en el **23,21 % (26/112)** de los artículos.

Finalmente, también hemos encontrado artículos en los que no se utilizan redes, siendo estos el **16,96 % (19/112)** del total. La mayoría de estos artículos trabajan con MBA, por lo que la información de las relaciones entre los individuos y poblaciones las almacenan los propios agentes. También hay estudios en los que se define un

espacio bidimensional, como puede ser una cuadrícula, sobre el que se van moviendo los individuos de la población en cada paso de la simulación. Los nuevos contagios se obtienen en base a las posiciones de los individuos.

2.3.5. PI4: ¿Qué escalas espaciales y temporales utilizan para analizar los brotes en este tipo de estudios?

Desde una perspectiva espacial, los brotes epidémicos pueden modelarse en el interior de un edificio o en áreas geográficas de diferentes tamaños. También los hay que **no indican explícitamente el espacio** sobre el que se simula el brote, enfocándose en los contactos entre personas que tienen una relación social. Estos artículos son el **39,29 % (44/112)** del total. De los restantes, la mayoría están enfocados en la propagación de un brote sobre un área geográfica (los hemos llamado "**simulaciones en exteriores**") con el **53,57 % (60/112)** de los artículos, mientras que las "**simulaciones en interiores**" son el **7,14 % (8/112)**. Estos últimos han estudiado la propagación de un brote dentro de un hospital (5/112), una residencia estudiantil (2/112) y una escuela (1/112). La Tabla 2.6 muestra la relación entre cada artículo y su escala espacial.

En las *simulaciones exteriores* hemos agrupado las dimensiones de las áreas que abarcan en las siguientes escalas:

- **Ciudad**. Es la escala más pequeña, cubriendo una ciudad completa (también podría ser un pueblo) o algunos de sus barrios. Estos estudios se centran en los desplazamientos de personas entre edificios. Esta escala comprende el **18,75 % (21/112)** de los artículos.
- **Región**. Esta escala puede abarcar diversas divisiones administrativas, como *provincia*, *departamento*, *distrito* o *municipio*. Estos estudios se centran en los desplazamientos de personas entre localidades. Esta escala comprende el **19,64 % (22/112)** de los artículos.
- **País**. Esta escala cubre un país y un conjunto de países. Estos estudios analizan los desplazamientos de las personas entre grandes áreas. Esta escala comprende el **15,18 % (17/112)** de los artículos.

En cuanto a la dimensión temporal, queremos identificar cuáles son las principales unidades de tiempo que representan cada paso de las simulaciones. Sin embargo, el **57,14 % (64/112)** de los artículos no proporciona esta información. En aquellos que sí la incluyen podemos distinguir dos escalas principales (ver Tabla 2.7):

- **Un día**. El **28,57 % (32/112)** de los artículos utiliza esta escala.
- **Menos de un día**. Esta escala va desde el minuto hasta varias horas. Estos trabajos representan el **14,29 % (16/112)** del total.

Tabla 2.6
Artículos por escala espacial.

Escala espacial	**Artículos**	**Nº artículos**
Edificio	[19, 38, 46, 58, 124, 127, 175, 206]	8
Ciudad	[11, 59, 66, 68, 75, 95, 105, 111, 118, 129, 131, 139, 141, 146, 153, 179, 194, 213, 226, 231, 233]	21
Región	[2, 10, 12, 17, 35, 36, 60, 61, 63, 71, 74, 79, 86, 100, 108, 123, 133, 137, 166, 177, 208, 219]	22
País	[16, 40, 42, 65, 88, 104, 106, 121, 127, 128, 136, 161, 178, 195, 200, 203, 204, 229]	17
Sin escala espacial	[3, 5, 7, 18, 27, 33, 43, 72, 73, 80, 84, 92, 94, 99, 103, 112, 115, 116, 122, 125, 130, 138, 154, 155, 157, 158, 162, 174, 180, 181, 182, 188, 210, 224, 225, 228, 230, 234, 235, 236, 237, 239, 240]	44

Tabla 2.7
Artículos por escala temporal.

Escala temporal	**Artículos**	**Nº artículos**
Un día	[7, 16, 36, 46, 59, 72, 74, 88, 105, 111, 112, 118, 119, 121, 123, 128, 129, 130, 139, 161, 166, 175, 179, 194, 200, 203, 204, 208, 229, 231, 233]	32
Menos de un día	[19, 35, 40, 42, 58, 63, 66, 68, 95, 104, 108, 127, 141, 206, 225, 226]	16
Sin información sobre la escala temporal	[2, 3, 5, 10, 11, 12, 17, 18, 27, 33, 38, 43, 60, 61, 65, 71, 73, 75, 79, 80, 84, 86, 92, 94, 99, 100, 103, 106, 112, 115, 116, 122, 124, 125, 131, 133, 136, 137, 138, 146, 153, 154, 155, 157, 158, 162, 174, 177, 178, 180, 181, 182, 188, 195, 210, 213, 219, 224, 228, 230, 234, 236, 237, 239, 240]	64

En el caso de las *simulaciones exteriores*, la unidad de tiempo más utilizada ha sido un día, encontrando unidades más pequeñas como medio día y otras fracciones de un día. En el caso de las *simulaciones interiores*, no es posible identificar una tendencia clara respecto a la dimensión temporal: la mitad de los estudios emplea un día como unidad temporal, mientras que la otra mitad utiliza unidades del orden de minutos. En la Tabla 2.8 se muestra la relación entre las dimensiones espacial y temporal.

Tampoco es común que se indique el tiempo total o número de pasos que abarcará la simulación.

Tabla 2.8
Relación entre escala espacial y temporal.

		Escala temporal	
		Menos de 1 día	*1 Día*
Escala espacial	*Edificio*	[19, 58, 127]	[46, 175, 206]
	Ciudad	[66, 68, 141, 226]	[59, 75, 105, 111, 118, 130, 139, 179, 194, 231, 233]
	Región	[35, 63, 95]	[36, 74, 123, 166, 208]
	País	[40, 42, 104]	[16, 88, 121, 128, 161, 200, 203, 204, 229]
	Sin escala espacial	[95, 225]	[7, 72, 112, 119, 130]

2.3.6. PI5: ¿Cuáles son las fuentes de datos usadas en este tipo de estudios? ¿Cuál es la granularidad de los datos?

Tras revisar los artículos hemos identificado el uso de diversas fuentes de datos, dependiendo del uso que se hace de ellas. Algunas se utilizan como referencia para crear poblaciones y localizaciones con características similares a la realidad, como es el caso de los **censos oficiales**, para el diseño de ciudades y regiones con unas estadísticas demográficas concretas (edad, sexo, empleo, densidad poblacional) en los MBA y redes de relaciones; los **estudios de transporte** para determinar la proporción de personas que se desplazan entre ubicaciones en los modelos MBA y redes de metapoblaciones; las **encuestas de población** para modelar el comportamiento social de las personas en su vida cotidiana y durante un brote; y otras fuentes para el mismo propósito, como perfiles de Facebook [131, 206], correos electrónicos [131] y publicaciones en Foursquare [11].

Para la **validación de los resultados** de las simulaciones se han utilizado los **datos oficiales de incidencia epidémica**. Por su parte, para la **creación de redes de contactos realistas** se han utilizado varias fuentes: registros de **coordenadas GPS** [11], registros de **localización de llamadas telefónicas** [213]; **datos de ingreso y alta hospitalaria** obtenidos de los Sistemas de Información Hospitalaria (SIH) y la Historia Clínica Electrónica (HCE) para modelar el movimiento de pacientes dentro de un hospital; **registros de localización mediante Bluetooth** de estudiantes alojados en una residencia universitaria [58]; y registros de personas equipados con **dispositivos de identificación por radiofrecuencia** (RFID) de cuándo han estado a un mínimo de distancia. De este último tipo de fuente de datos distinguimos los datos de acceso abierto del proyecto Sociopatterns [43, 124, 235], y los registros de movimientos de personal sanitario, pacientes hospitalizados [38] y acompañantes [127].

TABLA 2.9
Artículos por fuentes de datos utilizadas.

Artículos	**Datos agregados**				**Datos individuales**					
	Censo	**Estudio de transportes**	**Encuesta de población**	**Datos oficiales de incidencia**	**Redes sociales**	**Datos de ingreso y alta de SIH y HCE**	**Dispositivos portátiles**	**Otros**	**Datos sintéticos**	**Nº artículos**
[35, 36, 40, 106, 108, 137, 194, 231]	✓									8
[16, 86, 88, 130, 204]	✓	✓								5
[42]	✓	✓	✓			✓				1
[65, 104, 203]	✓	✓	✓	✓						3
[213]	✓	✓	✓				✓			1
[111, 233]	✓		✓							2
[118]	✓		✓	✓						1
[131]	✓		✓	✓	✓					1
[61, 68, 74, 75, 226]		✓	✓							5
[121, 161]		✓	✓	✓						2
[43]		✓	✓				✓			1
[153]			✓							1
[206]			✓		✓					1
[210]			✓						✓	1
[60, 112, 128, 178, 195, 229]			✓	✓						6
[59]			✓	✓					✓	1
[11]					✓					1
[46, 105, 166, 175, 234]						✓				5
[38, 124, 127, 235]							✓			4
[58]							✓	✓		1
[7, 63, 136, 141]								✓		4
[2, 3, 5, 10, 12, 17, 18, 19, 27, 33, 61, 66, 71, 72, 73, 79, 80, 84, 92, 94, 95, 99, 100, 103, 115, 116, 119, 122, 125, 130, 133, 139, 146, 155, 157, 158, 162, 174, 177, 179, 180, 181, 182, 188, 200, 208, 219, 224, 225, 228, 230, 236, 237, 239, 240]									✓	56
Nº artículos	22	18	6	14	3	6	7	5	58	

En el **51,79 % (58/112)** de los artículos se utilizan **datos sintéticos** generados sin utilizar información extraída de fuentes de datos para la creación de distintos tipos de redes. Es estos casos destaca la creación de redes de relaciones siguiendo modelos para la generación de redes complejas, como el modelo de red aleatoria de Erdös-Rényi y el modelo de red libre de escala de Barabási-Albert.

Un resumen de las categorías de datos utilizadas en cada trabajo analizado en esta investigación se presenta en la Tabla 2.9

En cuanto a la granularidad, las fuentes de datos pueden clasificarse en:

- **Datos agregados**. Estas fuentes de datos resumen el número de casos incluidos en diferentes categorías. Aquí se incluyen censos oficiales, estudios de transporte y datos oficiales de incidencia epidémica. También se incluyen aquí las respuestas agregadas de las encuestas de población.
- **Datos individuales**. Estas fuentes de datos contienen información detallada de cada registro. Aquí se incluyen datos de redes sociales, registros de SIH y HCE, registros RFID y el resto de las fuentes incluidas en la categoría "otros" en la Tabla 2.8.

Los **datos agregados** son la única fuente de datos en el **27,68 % (31/112)** de los artículos, mientras los artículos que solo utilizan **datos individuales** representan el **16,07 % (18/112)**. El **8,04 % (9/112)** de los trabajos presentan el **uso combinado** de ambos tipos de datos, utilizando censos y encuestas de transporte junto con encuestas de población y redes sociales.

Otro aspecto relevante sobre las fuentes de datos es su **accesibilidad**, siendo una característica cada vez más importante para garantizar la reproducibilidad de la investigación. Algunas fuentes de datos son oficiales y públicas, como el caso de los censos, estudios de transporte, datos de incidencia y algunas encuestas [42, 111, 153, 210], algunos conjuntos de datos de redes sociales [11, 131]. El 38,4 % (43/112) de los trabajos utilizan únicamente datos de acceso abierto, mientras que solo el 10,71 % (12/112) utilizan datos de fuentes privadas, como los registros de SIH y HCE, la mayoría de los registros RFID y algunas encuestas.

Sin embargo, tanto en los estudios que se utilizan fuentes públicas como privadas como en los que crean datos sintéticos sin referencias, no es habitual proporcionar las redes creadas (o los agentes en MBA) para las simulaciones. Esto afecta negativamente a la reproducibilidad de la investigación, un tema clave en la fiabilidad de los estudios científicos. En relación con este tema, podemos incluir el acceso al código implementado. Solo cuatro artículos [72, 104, 200, 230] han proporcionado acceso al código mediante repositorios de acceso abierto.

2.4 Discusión

2.4.1. Uso de los modelos epidemiológicos compartimentales

Los modelos epidemiológicos compartimentales surgieron como una herramienta con la que simplificar la modelización matemática de las enfermedades infecciosas. Desde el primer modelo epidemiológico compartimental determinista, SIR, propuesto por Kermack y McKendrick [97] en 1927, este campo sigue siendo objeto de estudio. Por ejemplo, se ha estudiado la inclusión de patrones de comportamiento en el modelo, como ocurre con el par de estados *Consciente-Inconsciente* (del inglés, *Aware-Unaware, AU*) para indicar si un individuo está informado del brote y, por ello, adquiere más precaución en su vida social. Estos estados no se usan como estados independientes, sino en combinación con los otros estados del modelo. Por ejemplo, una persona en el estado *US* (*Unaware* y *Susceptible*) podría pasar a *AS* (*Aware* y *Susceptible*) o a *AI* (*Aware* e *Infectado*).

También hay estudios que analizan la dinámica de brotes reales para conseguir modelos más precisos para enfermedades concretas. Estos modelos suelen incluir nuevos estados que definen diferentes niveles de gravedad dentro del estado *Infectado*, como *Asintomático*, *Sintomático Leve* y *Sintomático Grave*, o acciones de intervención, como *Hospitalizado* y *Cuarentena en casa*. En contraposición, en [154] y [155] se intenta crear un modelo, *G-SEIV* (Generalización de Susceptible-Expuesto-Infectado-Vigilante), que sea genérico y adaptable a enfermedades infecciosas con distintas dinámicas.

En general, es preferible trabajar con modelos que se ajusten únicamente a la dinámica requerida, en lugar de modelos complejos con múltiples estados y probabilidades de transición entre ellos. Debido a la gran cantidad de parámetros involucrados, estos modelos detallados son difíciles de resolver analíticamente y, aunque actualmente son factibles a nivel computacional, implican simulaciones con una gran incertidumbre y dificultad en la interpretación. Por tanto, su utilidad para fines teóricos puede ser limitada.

2.4.2. Relación entre modelos computacionales, tipos de redes y dimensión espacial

Hemos encontrado una estrecha relación entre los resultados de PI2, PI3 y los referidos a la dimensión espacial de PI4 (ver Tabla 2.10), mostrándose a continuación los hallazgos más destacados.

- En la mayoría de los artículos que modelan un brote en el **interior de un edificio** han utilizado un **modelo estocástico** en el que los posibles contagios se restringen en función de **redes de contacto**. Estos estudios se centran en los contactos diarios dentro de un edificio, principalmente hospitales. Cabe destacar que para las simulaciones interiores no se ha utilizado MBA. Tampoco se han

utilizado redes multicapa. Sin embargo, consideramos que podrían ser útiles pues su IPL permitiría simular políticas como el aislamiento de pacientes: si un paciente está infectado y es aislado, el personal sanitario lo sabría y reduciría su contacto físico con el paciente, disminuyendo así su probabilidad de infección.

- A medida que aumenta la escala espacial de las simulaciones, se utilizan distintos tipos de redes. Las **redes de contacto** se emplean a nivel de **ciudad**. A diferencia de otras escalas mayores, a nivel de ciudad es computacionalmente viable la simulación de contactos individuales entre personas y es posible encontrar información real sobre contactos: hay individuos que comparten los mismos espacios, como los hogares y los centros de estudio o trabajo, y puede haber interacciones fortuitas en el transporte público y comercios. El modelado de estos contactos tan específicos se realiza principalmente con MBA y modelos estocásticos.

- Las **redes de metapoblaciones** se han utilizado principalmente en escalas grandes, como **región y país**. En estas redes, los nodos representan localidades e incluyen propiedades, como el número de habitantes. Las aristas son dirigidas y representan el número de personas que se trasladan diariamente entre ellas. Estos estudios, que utilizan principalmente un **modelo estocástico**, se utilizan para estudiar si es posible la transmisión de un brote entre distintas localidades debido el movimiento de personas entre ellas. Para cada localidad se define un modelo epidemiológico (en la mayoría de los casos, sin redes de contacto o de relaciones) para simular la evolución del brote en ella.

- Los estudios que **no indican su escala espacial** utilizan, principalmente, **redes de relaciones y multicapa**. En estos artículos se suelen analizar aspectos concretos, como la eficacia de algunas medidas de intervención (planes de vacunación, cuarentenas, confinamientos, distanciamiento social) y el comportamiento de las personas. En estos casos, el modelo más utilizado es el estocástico, seguido del determinista.

- Los **MBA** se utilizan principalmente a nivel de **ciudad y región**, donde cada agente guarda los contactos o relaciones que tiene (no se usan redes) o bien se crea una red global para toda la simulación, principalmente de redes de relaciones o multicapa.

TABLA 2.10
Relación entre modelo computacional, escala espacial y tipo de red.

		Modelo computacional		
Escala espacial	**Tipo de red**	**Modelo estocástico**	**Modelo determinista**	**Modelo basado en agentes**
Edificio	*Red de relaciones*	[206]	-	-
	Red de contactos	[38, 46, 58, 127, 175]	-	[124]
	Red multicapa	-	-	-
	Red de metapoblación	-	-	-
	No usa redes	-	-	[19]
Ciudad	*Red de relaciones*	[131, 146]	[153]	[129, 179]
	Red de contactos	[141, 233]	[194]	[11]
	Red multicapa	[59]		
	Red de metapoblación	[226]	[213]	[68]
	No usa redes	[118, 139]	[105]	[66, 75, 111, 231]
Región	*Red de relaciones*	[100]	[60]	[63, 86]
	Red de contactos	-	-	[36]
	Red multicapa	-	-	[108]
	Red de metapoblación	[2, 61, 71, 121, 123, 133, 137, 177, 203, 208]	[12, 17, 79, 166, 219]	-
	No usa redes	-	[10, 35]	[74]
País	*Red de relaciones*	[229]	-	-
	Red de contactos	-	-	-
	Red multicapa	-	-	-
	Red de metapoblación	[16, 42, 65, 88, 104, 161]	[204]	-
	No usa redes	[106]	[128, 178, 200]	[40, 136, 195]

Sin escala espacial	*Red de relaciones*	[5, 27, 33, 43, 72, 92, 112, 138, 154, 155, 224, 228, 230, 237, 239]	[80, 99, 103, 115, 162, 180, 181, 188, 210, 235]	[95]
	Red de contactos	[84, 119, 174, 234, 239]	[116, 125]	[7]
	Red multicapa	[18, 73, 94, 122, 157, 182, 236, 240]	-	[130]
	Red de metapoblación	[225]	-	-
	No usa redes	-	[3]	-

2.4.3. Propiedades de las redes

Podemos analizar la estructura de las redes en base a tres características principales: tamaño (número de nodos), topología y dinámica (su estructura cambia a lo largo de la simulación).

El **tamaño** de las redes se incrementa con la escala espacial. Las redes utilizadas en interiores tienen del orden de 10^2 a 10^3 nodos, a nivel de ciudad y región tienen de 10^3 a 10^4 nodos y a nivel de país tienen de 10^6 nodos. En los estudios sin escala espacial se utilizan redes pequeñas de 10^2 a 10^3 nodos.

En cuanto a su **topología**, las *redes complejas* son el tipo topológico más utilizado para modelar las redes de relaciones y multicapa, ya que permiten ajustar propiedades como la asortatividad entre nodos, el coeficiente de agrupamiento y la distribución de los grados de los nodos. Esto permite modelar múltiples aspectos sociales de la población, como grupos familiares y de amistad o estratos de edad. Se ha identificado una relación entre la topología y el uso de las redes, basada en los conceptos que representan sus nodos y aristas. Las redes de relaciones se emplean principalmente cuando no hay información espacial. Representan poblaciones con distintas propiedades, como aspectos sociales (hogares familiares, grupos sociales, grado de sociabilidad de las personas) y los efectos de las medidas de intervención sobre la población.

Hemos encontrado estudios en los que se comparan distintos métodos de generación de redes complejas para identificar cuáles de ellas modelan mejor la población o fenómeno que quieren representar. Entre los tipos de redes complejas citadas se encuentran: las *redes libres de escala de Barabási-Albert*, las *redes aleatorias de Erdös-Rényi* y las *redes de mundo pequeño de Watts y Strogatz*. Estos tipos topológicos de redes se utilizan tanto para redes de relaciones y multicapa, como de contactos.

En el caso de las redes de metapoblaciones, no se referencian topologías específicas, sino que se definen como redes estáticas, dirigidas y ponderadas.

En cuanto a la **dinámica** de las redes, podemos hallar una relación entre el tipo de red y cómo cambian sus aristas. Las redes de relaciones son estáticas, es decir, no varían los grupos sociales a lo largo de la simulación. Por el contrario, en las redes de contactos encontramos que se modifican las aristas para cada paso de la simulación, estando algunas ponderadas con el número de contactos ocurridos o la suma de sus duraciones. Estos cambios pueden aplicarse directamente a la red o se puede construir una nueva red estática en cada paso (red de variación temporal). En el caso de las redes multicapa, algunos estudios crean una OPL dinámica con el fin de representar la reducción de los contactos físicos entre personas a medida que el brote evoluciona.

Con relación al número de nodos, este permanece constante en todos los estudios analizados, ya que los períodos de simulación son lo suficientemente cortos como para que los cambios demográficos no afecten al brote. Por tanto, no es necesario añadir o eliminar nodos que representen nacimientos o muertes naturales o movimientos migratorios. Los individuos que mueren a causa de la enfermedad se representan con un estado en el modelo epidemiológico compartimental.

Consideramos que el uso de aristas y nodos dinámicos es apropiado en simulaciones en interiores, ya que estos escenarios presentan un alto nivel de contactos físicos y entradas y salidas del edificio en pequeños intervalos de tiempo. Un caso práctico podría ser un hospital, donde hay nuevos ingresos y altas de manera continuada, y los pacientes pueden cambiar de habitación en intervalos de varias horas.

2.4.4. Características de los datos

Los datos son un aspecto fundamental en el diseño y la evaluación de un modelo computacional. Dos aspectos relacionados con los datos a discutir son: la capacidad de análisis y la disponibilidad.

La disponibilidad de datos es limitada, lo que implica que no siempre sea posible un análisis que combine las perspectivas espacial y temporal. Desde un punto de vista espacio-temporal, los datos individuales permiten realizar análisis más detallados y sólidos que los datos agregados. Además, tienen la ventaja de que pueden agregarse en cualquier nivel temporal y espacial, mientras que los datos agregados solo se utilizan como parámetros para crear poblaciones (incluidos las relaciones sociales y los trayectos entre ubicaciones) con unas características demográficas que se ajusten a una población real. En este sentido, fuentes como los censos y estudios de transporte son las más limitadas.

A nivel temporal, los datos agregados no tienen un detalle menor a un día, por lo que no pueden emplearse para reconstruir redes de contacto o de relaciones. En estos casos, los contactos son estimados, lo que introduce una alta incertidumbre sobre las condiciones de transmisión del brote simulado. Además, cuando se emplean datos agregados sin información espacial, las conclusiones extraídas de las simulaciones podrían ser comparables a las obtenidas mediante modelos de series temporales. Aunque cabe mencionar que aún presentarían como ventaja respecto a las series

temporales el poder calcular y estimar características del brote en base al modelo epidemiológico utilizado, permitiendo una mejor interpretación de los resultados.

El uso de datos individuales como los perfiles de redes sociales puede permitir la reconstrucción de redes de relaciones (e incluso de contactos) más realistas, mejorando la capacidad de análisis y la calidad de las conclusiones (respecto al uso de datos agregados a nivel de alto nivel espacial). Sin embargo, este tipo de datos se usa principalmente como un complemento para refinar aspectos del comportamiento social de la población. Su menor uso puede deberse a que estos perfiles no suelen ser públicos y muchas personas no consienten el análisis de sus perfiles privados. Además, es importante mencionar que los perfiles en redes sociales pueden no ser una fuente fiable de información sobre los contactos reales de una persona en su vida fuera del ámbito digital.

Una mayor precisión en el nivel de detalle de los datos eliminaría la necesidad de generar datos sintéticos para ajustar los modelos. Por ejemplo, el rastreo periódico de la posición de los individuos permitiría establecer sus redes de contacto. Cuando existe un alto nivel de detalle, el análisis espacial puede realizarse tanto a pequeña escala (el interior de un edificio) como a gran escala (una ciudad o una región). También sería posible hacer simulaciones con diferentes escalas temporales, desde segundos a días. Sin embargo, solo algunos estudios han alcanzado este nivel de detalle.

Entre los estudios con un mayor nivel de detalle destacan aquellos que modelan la propagación de una infección dentro de un hospital. Estos estudios basan la simulación en la reconstrucción de una red de contactos en base a los registros de ingreso y alta del SIH y el registro de sensores RFID llevados por el personal sanitario y los pacientes hospitalizados. Sin embargo, este tipo de datos no suelen estar disponibles públicamente, ya que requieren la autorización de los participantes y la administración sanitaria. Bases de datos clínicas como MIMIC-III y el uso de datos sintéticos podrían facilitar el desarrollo y evaluación de estos estudios, ya que evitarían las dificultades de acceso a datos reales.

En el ámbito de la salud, también existe una tendencia creciente hacia el uso secundario de datos de HCE con fines más amplios que la simple atención al paciente. Actualmente, algunas HCE registran el historial de servicios y camas ocupadas por cada paciente, como salas de radiografía. Además, esta información puede proporcionarse con el nivel de detalle espacial y temporal requerido, por ejemplo, para construir una red diaria de contactos entre pacientes.

2.5 Conclusiones

En este capítulo hemos analizado el estado de los modelos computacionales para la representación de la transmisión de brotes epidémicos, centrándonos en el uso el uso de redes y el análisis espacio-temporal.

Desde el punto de vista epidemiológico, encontramos que las enfermedades infecciosas más estudiadas han sido enfermedades virales, como la gripe estacional o el COVID-19, que suelen tener un periodo de latencia y que, al recuperarse de la infección, el individuo obtiene un estado inmune hasta el final de la simulación. Este tipo de infecciones son representadas principalmente con los modelos compartimentales SEIR y SIR, así como con variaciones de estos para adaptarlos a escenarios más complejos y análisis de otros aspectos relacionados con el brote.

Hemos identificado un creciente interés en el uso de redes como apoyo para la simulación de los brotes. Desde el punto de vista semántico, las redes más utilizadas son las redes de relaciones y las redes de metapoblaciones. Ambos tipos de redes suelen ser estáticas. Por su parte, las redes de contactos suelen ser dinámicas, cambiando en cada paso de la simulación. Desde un punto de vista topológico, hemos encontrado una preferencia en el uso de redes complejas para modelar redes de relaciones y redes multicapa, ya que cuentan con varios parámetros ajustables que facilitan la representación de conexiones sociales como núcleos familiares y grupos de amigos.

Si relacionamos el uso de los modelos computacionales, el tipo de red y la escala espacial a simular, destacan las siguientes combinaciones:

- En interiores se suelen simular brotes mediante modelos estocásticos en los que los contactos entre los individuos vienen dados por una red de contactos.
- En las simulaciones en ciudad encontramos que se han estudiado todas las combinaciones posibles de modelo computacional y tipo de red. Podemos destacar el uso de MBA para rastrear los contactos entre personas en escalas temporales reducidas.
- Al aumentar la escala espacial, las redes de metapoblaciones en combinación con los modelos estocásticos son predominantes.
- Cuando las simulaciones se realizan sin modelar un espacio concreto, suelen enfocarse en otros aspectos de la epidemia, como la difusión de la conciencia sobre el brote o la efectividad de distintas medidas de intervención. En estos estudios destaca el uso de tanto modelos estocásticos como deterministas, así como las redes de relaciones, multicapa y de contactos.

En cuanto a la información temporal, no se suele indicar cuál es la unidad de tiempo que representa cada paso de las simulaciones, aunque la más mencionada ha sido un día o fracciones de día.

Finalmente, con relación a los datos, queremos destacar la falta de uso datos detallados sobre los movimientos y contactos de la población, tanto para simulaciones interiores como exteriores. Lo más común es el uso de datos agregados a gran escala, como censos y estudios de transporte, para obtener estadísticas demográficas deseables para la población de la simulación. Además, a nivel hospitalario, encontramos dificultades en el acceso a fuentes de datos relevantes, como los registros de SIH y HCE, así como una escasez en modelos de simulación para movimientos hospitalarios.

Capítulo 3

Modelado espacio-temporal para la investigación epidemiológica de infecciones nosocomiales

En este capítulo proponemos un modelo de datos epidemiológico que permita una descripción precisa y a diferentes escalas de la distribución de un hospital, así como el registro de los movimientos de los pacientes a través de este. El objetivo del modelo es permitir razonar sobre los contactos entre pacientes y, en base a ellos, las posibles rutas de transmisión de la infección. Queremos encontrar un equilibrio entre generalidad y detalle, de manera que el modelo pueda servir de base para futuras investigaciones epidemiológicas intrahospitalarias. Además, hemos desarrollado seis consultas que, basadas en el modelo, consigan realizar tareas epidemiológicas básicas, como la detección de brotes.

3.1 Introducción

Aunque ya hemos destacado anteriormente la importancia del estudio de las infecciones nosocomiales, que tienen lugar dentro de los hospitales, en base a la información presentada en el Capítulo 2, extraemos que no hay tanta investigación epidemiológica basada en los contactos de los individuos en espacios cerrados como para grandes espacios abiertos. En el caso concreto de ambientes hospitalarios, hasta donde sabemos, no existen trabajos que propongan un modelo de datos general que permita realizar un razonamiento espacial y temporal para la investigación epidemiológica dentro de hospitales. Además, la mayoría de los trabajos obtienen datos de sensores RFID que registran cuándo dos personas han estado a una distancia máxima entre sí. Por lo tanto, no existe un registro de los lugares donde ocurrió el contacto.

Proponemos un modelo de datos que pueda servir como base para el razonamiento espacial y temporal en futuras investigaciones epidemiológicas dentro de hospitales. El objetivo es que nuestro modelo sea lo suficientemente general, encontrando un equilibrio

entre una representación detallada y una escalabilidad computacional eficiente. Con nuestro modelo espacio-temporal pretendemos poder representar la disposición espacial de cualquier hospital y el movimiento de sus pacientes dentro de sus instalaciones, obteniendo esta última información de los Sistemas de Información Hospitalaria (SIH) o de la Historia Clínica Electrónica (HCE) de los pacientes. Las dimensiones espacial y temporal se describen por separado, pero están interconectadas, lo que permite conocer en todo momento la ubicación exacta de un paciente. También evaluamos la idoneidad de nuestro modelo mediante la definición de seis consultas que explotan su semántica para representar distintas tareas en la investigación epidemiológica, como la detección de brotes y el análisis de su propagación.

Vamos a realizar una selección tecnológica que permite implementar eficientemente el modelo como un grafo y las consultas. La validación del modelo espacio-temporal y las consultas se realiza mediante el diseño de dos experimentos en los que analizamos dos brotes de *C. diff* utilizando diferentes conjuntos de consultas epidemiológicas. Para estos experimentos empleamos datos sintéticos generados a partir de un simulador realista de infecciones hospitalarias.

A continuación, resumimos las principales contribuciones de este capítulo:

- **Definición de un modelo de datos espacio-temporal para la investigación epidemiológica en hospitales**, basado en los movimientos y contactos de los pacientes. Lo hemos representado mediante UML (*Unified Modeling Language*; en español, Lenguaje de Modelado Unificado).
- **Definición de seis consultas para la investigación epidemiológica** basadas en el modelo espacio-temporal.
- **Un estudio empírico del rendimiento de las bases de datos orientadas a grafos**, concretamente a grafos de propiedades y grafos de conocimiento, basado en los dos primeros puntos.

El resto del capítulo se estructura como sigue: en la Sección 3.2 presentamos nuestra propuesta para que sirva de base para la investigación espacio-temporal en hospitales, la cual está formada por un modelado de datos (Sección 3.2.1) con una descripción de la dimensión espacial (Sección 3.2.1.a) y temporal (Sección 3.2.1.b), y por un conjunto de consultas que representan tareas epidemiológicas (Sección 3.2.2). Además, en la Sección 3.3 analizamos la tecnología para almacenamiento y procesamiento de datos basados en grafos, seleccionando aquella con la que implementar el modelo y consultas propuestos. En la Sección 3.4 validamos el modelo espacio-temporal y las consultas epidemiológicas, siendo los hallazgos más destacados discutidos en la Sección 3.5. Finalmente, presentamos nuestras conclusiones en la sección 3.6.

3.2 Propuesta

3.2.1. Modelado del dominio

En el capítulo anterior, Capítulo 2, observamos que, si bien se ha estudiado el uso de redes y grafos para el análisis epidemiológico y simulación de brotes, estas estructuras suelen carecer de complejidad semántica y juegan un papel secundario en el análisis. Todos sus nodos y aristas representan un mismo concepto, en muy pocos casos añadiendo propiedades a alguno de estos elementos. Se utilizan como una estructura para el almacenamiento de datos conexos (personas relacionadas física o virtualmente, o regiones con movimiento de personas) durante la ejecución de la simulación del brote. En nuestra opinión, no se explotan al máximo los beneficios que puede proporcionar el uso de redes, como puede ser el acceso eficiente a grandes cantidades de datos altamente conectados que representan distintos conceptos o subcategorías de un mismo concepto. También se podrían aplicar diversas técnicas de análisis de grafos, como puede ser la detección de patrones, de comunidades o de nodos altamente conectados, cuyos resultados pueden ayudar a identificar factores que hayan favorecido la propagación de la infección.

También observamos que el grueso de la investigación en propagación de brotes se centra en la simulación de brotes sobre grandes espacios, como una ciudad o una región. La investigación dentro de edificios fue minoritaria. Los hospitales son un caso de edificio donde la aparición y propagación de brotes infecciosos entre pacientes es común y, por consiguiente, un problema. En el interior de un hospital, la cercanía y el contacto entre pacientes puede definirse a distintos niveles de extensión superficial: desde una planta hasta estar en camas dentro de la misma habitación, pasando por diversos niveles intermedios. Los conceptos que definen estos niveles y las relaciones espaciales que se establecen entre ellos son propicios de modelarse en forma de red. Como identificamos en el Capítulo 2, un tipo semántico de red utilizado en la literatura son las redes de contactos, con las que se registra qué pares de individuos han estado en contacto durante un tiempo determinado. Sin embargo, en estas redes no se incluye la información relativa al espacio y el tiempo que describe el contacto. Partiendo de la idea propuesta con las redes de contactos y uniéndola con la estratificación de la estructura física de un hospital, podemos crear una red donde se detallen espacio-temporalmente los contactos entre pacientes. Sin embargo, podemos profundizar aún más en la semántica temporal de la red: en lugar de registrar contactos, se puede utilizar para almacenar los movimientos de cada uno de los pacientes. De esta manera, el contacto entre pacientes ya no queda definido por una arista entre sus nodos, sino a través de los nodos que representan el espacio y el tiempo donde han estado. Por tanto, es posible definir contactos con distintos niveles de precisión tanto espacial como temporal. Además, esta información puede ser persistida en una base de datos orientada a grafos, posibilitando su uso para en diferentes tipos de análisis: la información sobre la ubicación de los pacientes es crucial para investigar cómo se propaga un brote o dónde se encuentra la fuente de contagio.

En este capítulo proponemos un **modelo de datos orientado a grafos** que pueda servir como base para el **razonamiento espacial y temporal** en futuras **investigaciones epidemiológicas dentro de hospitales**. Más concretamente, presentamos un **modelo espacio-temporal** con el que representar la **disposición espacial de cualquier hospital** y el **movimiento de sus pacientes** dentro de sus instalaciones. Planteamos como objetivo que el modelo sea general, encontrando un equilibrio entre una representación detallada y una escalabilidad computacional eficiente. En cuanto a los datos con los que poblar el modelo, nuestro objetivo es que se utilice información procedente de los Sistemas de Información Hospitalaria (SIH) o de la Historia Clínica Electrónica (HCE) de los pacientes.

En nuestro modelo, **las dimensiones espacial y temporal se definen por separado, aunque se encuentran interconectadas**. Esto nos permite representar con precisión los desplazamientos de los pacientes, además de poder inferir relaciones entre ambas dimensiones. Hemos formalizado nuestro modelo espacio-temporal mediante un diagrama de clases UML (Figura 3.1), en el cual se diferencian claramente los conceptos de las dimensiones espacial y temporal. En las siguientes secciones detallamos cada una de las dimensiones.

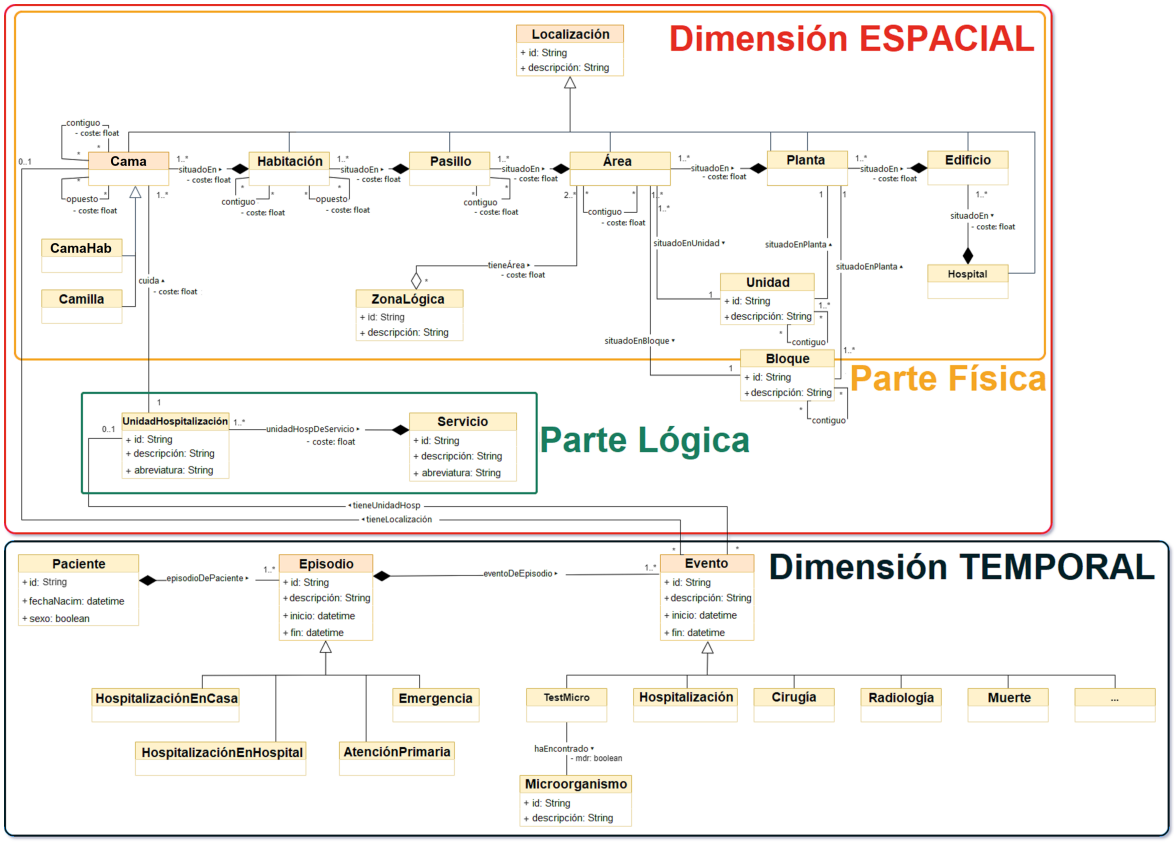


FIGURA 3.1. Diagrama de clases UML del modelo epidemiológico espacio-temporal.

3.2.1.a. Dimensión espacial

La dimensión espacial se utiliza para describir la estructura de un hospital, definiendo varios niveles que se estructuran jerárquicamente. Proponemos utilizar el concepto *Localización* para identificar dónde se encuentra un paciente dentro del hospital. Una *Localización* es una interfaz que representa cualquier elemento arquitectónico de un edificio, como una *Habitación* o un *Pasillo*. También lo hemos extendido a otros niveles conceptuales más y menos concretos, como *Cama* o *Área*. Cada *Localización* tiene un identificador interno y una descripción para facilitar su reconocimiento por parte de los usuarios.

Las *Localizaciones* están organizadas jerárquicamente mediante la relación *situadoEn*, de manera que los niveles superiores agrupan entidades del nivel inmediatamente inferior:

- El nivel más bajo es ***Cama***, una interfaz que indica un mueble ocupado por un paciente, como una *Cama* para hospitalizaciones, una *Camilla* de quirófano o un *Box* de UCI.
- El siguiente nivel es ***Habitación***, que puede contener varios *Camas*.
- ***Pasillo*** representa un conjunto de *Habitaciones* situadas en un mismo pasillo o sección de un pasillo.
- El concepto de ***Área*** nos permite agrupar pasillos cercanos. Para ello, hemos dividido cada *Planta* en una cuadrícula con coordenadas, en la que cada celda representa un *Área*. Las filas y columnas de la cuadrícula se representan con los conceptos ***Unidad*** (fila) y ***Bloque*** (columna), de manera que podemos localizar cada *Área* dentro de una *Planta* mediante un par *(Bloque, Unidad)*.
- Los tres niveles superiores de la jerarquía son ***Planta***, ***Edificio*** y ***Hospital***.

Esta jerarquía de *Localizaciones* permite definir diferentes niveles de granularidad para buscar pacientes y determinar su proximidad espacial de manera cualitativa. Dado que los nodos *Paciente* están conectados a las *Camas* mediante la relación *situadoEn* definida entre este nivel inferior de la jerarquía y los *Eventos* que definen la dimensión temporal, es posible descubrir qué otros pacientes están conectados a la misma *Localización* (*Habitación*, *Pasillo*, *Área*, etc.).

Aún podemos dotar de más complejidad semántica a las relaciones entre *Localizaciones*, dado que además de la relación de pertenencia *situadoEn*, podemos establecer relaciones entre elementos de un mismo nivel, como es el caso de *contiguo* y *opuesto* (dos *Camas* enfrentadas en la misma *Habitación*, dos *Habitaciones* opuestas separadas por un pasillo), permitiendo distinguir pacientes con mayor o menor proximidad más detalladamente. También hemos introducido el concepto *ZonaLógica* (*ZL*) que nos permite agrupar *Áreas* conectadas por razones ajenas a la disposición arquitectónica del hospital.

Además de por la cercanía física entre los pacientes (estar a una corta distancia), en un hospital también se pueden encontrar medios de contacto indirecto. Uno de estos casos puede ser el personal sanitario (PS). Para registrar qué PS ha estado con un paciente, utilizamos el término *UnidadHospitalización* (*UH*), que representa un grupo de PS que trabajan en la misma unidad de atención sanitaria. Cada *UH* es responsable del cuidado de un conjunto de *Camas*, lo cual se representa mediante la relación *cuida*. A su vez, las *UH* se agrupan en *Servicios*.

A modo de resumen, podemos describir la dimensión espacial del modelo como la unión de dos partes:

- La **parte física**, que abarca los elementos que conforman la estructura arquitectónica del hospital (jerarquía de *Localizaciones*, *ZL*, *Bloque* y *Unidad*).
- La **parte lógica**, que representa la organización del personal sanitario (*UH* y *Servicio*).

Con el fin de permitir una evaluación no solo cualitativa de la distancia entre pacientes, sino también cuantitativa, hemos añadido la propiedad ***coste*** a las aristas que nos permiten unir los conceptos físicos y lógicos compartidos por dos pacientes: *situadoEn*, *contiguo*, *opuesto*, *tieneÁrea* y *cuida*, *unidadHospDeServicio*.

El valor de la propiedad *coste* puede establecerse de manera unitaria para cada arista (por ejemplo, se conocen las distancias reales entre cada para de *Localizaciones*), de manera global con un mismo valor para todas las aristas (por ejemplo, se les puede poner *coste=1* a todas las aristas, de manera que la distancia entre dos pacientes se calcularía con un algoritmo de camino más corto basado en el número de saltos dados), o aplicando un mismo valor a cada tupla (tipo de relación, concepto del nodo de origen, concepto del nodo destino). Para este último caso, se podrían definir dos valores para las aristas *contiguo* entre *Habitaciones*: un valor menor para cuando ambas habitaciones tienen su puerta contigua y otro valor para cuando están físicamente juntas pero su puerta está orientada a distintos pasillos. De igual modo, también se podría querer rebajar el nivel de detalle y establecer el mismo peso para las relaciones *opuesto* y *contiguo* entre *Localizaciones* del mismo nivel jerárquico.

Ha de notarse que podemos trazar dos caminos para definir la vecindad entre dos *Áreas*: uno a través de la arista *contiguo* y otro a través de los conceptos *Unidad* o *Bloque*. Sin embargo, el uso del segundo camino podría dar caminos erróneos, por lo que hemos decidido no incluir los elementos que lo componen dentro del subgrafo que se definiría para calcular la distancia espacial entre dos localizaciones. Tampoco se le ha añadido la propiedad *coste* a sus aristas. Un ejemplo del uso erróneo de este segundo camino puede verse en la Figura 3.2. En ella encontramos en la esquina superior derecha la representación en forma de cuadrícula de una *Planta* formada por seis *Áreas* divididas entre dos *Unidades* y tres *Bloques*. Un diagrama en forma de grafo representa las relaciones entre las *Áreas*, *Unidades* y *Bloques*. El *Área 0A* sería vecina de las *Áreas 1A* y *0B*. Sin embargo, a través de las relaciones *contiguo* entre las *Unidades* y *Bloques*

se podría llegar a establecer la vecindad con cualquier *Área* situada en la misma fila o columna, por ejemplo, con el *Área 2A*.

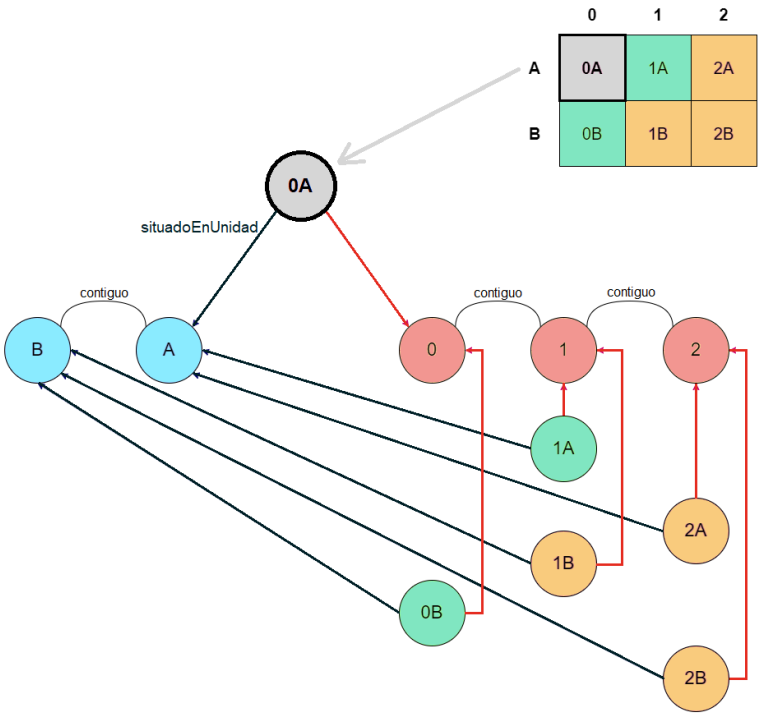


FIGURA 3.2. Ejemplo de una planta dividida en una cuadrícula y su representación como grafo.

3.2.1.b. Dimensión temporal

La dimensión temporal se utiliza para registrar el momento en el que le ocurre cualquier suceso a un paciente durante su hospitalización. Modelamos estos sucesos mediante el concepto ***Evento***, siendo de gran relevancia aquellos que nos permitan seguir los movimientos de los pacientes a través del hospital.

Trabajamos con dos tipos de datos temporales, que se diferencian según su duración:

- **Intervalos**: representan **acciones que se extienden en el tiempo**. Por ejemplo, una hospitalización en una cama, una cirugía, un examen médico o el padecimiento de una enfermedad.
- **Puntos**: representan acciones que ocurren en un instante específico. Por ejemplo, una admisión, el resultado de un análisis de sangre, una prueba microbiológica positiva o un fallecimiento.

Esta división del tiempo nos permite representar lo que les sucede a los pacientes como una secuencia de acciones donde se permite la superposición en el tiempo. Nuestro

modelado del tiempo orientado a eventos ya ha sido estudiado en trabajos como [189] y [102].

Aunque hemos definido dos tipos de *Eventos* en función de su duración, *intervalos* o *puntos*, ambos tipos se representan bajo una misma clase en nuestro modelo. La diferencia entre ellos está en los valores de sus propiedades *inicio* y *fin*: en los intervalos, estos valores son diferentes, mientras que, en los puntos, son idénticos. Los valores de estas dos propiedades están formados por la combinación de una fecha y el tiempo en formato `hora:minuto:segundo`. Para simplificación del modelo, los *Eventos* solamente se clasifican semánticamente según la acción que representan, creando una jerarquía.

Un tipo particular de *Evento* es ***TestMicro***, que representa una prueba microbiológica que ha detectado una o más especies de microorganismos infecciosos. Estos *Eventos* son útiles para determinar cuándo hay pruebas empíricas de que un paciente está infectado (ver Sección 2.3.2). La relación *haEncontrado* conecta un *TestMicro* con cada uno de los *Microorganismos* detectados, y su propiedad *mdr* (del inglés, *MultiDrug-Resistant organism*) indica si se trata de un patógeno multirresistente (es decir, resistente a múltiples fármacos).

Los *Eventos* son el punto de unión entre las dimensiones temporal y espacial. Las relaciones *tieneLocalización* y *tieneUnidadHosp* nos permiten conectar los *Eventos* con cada una de las partes que forman la dimensión espacial de manera independiente. Por ejemplo, es posible representar que un paciente se sometió a una cirugía en un quirófano y qué UH llevó a cabo la operación. También sería posible definir que mientras un paciente está hospitalizado en una *Cama* y atendido por una *UH*, se le puede realizar un examen médico por parte de otra *UH* del mismo *Servicio*. Además de los *Eventos*, en la dimensión temporal también hemos introducido el concepto de ***Episodio*** para agrupar todos los *Eventos* que le ocurren a un *Paciente* desde su admisión hasta su alta hospitalaria. Por tanto, representamos los *Episodios* como intervalos. Tanto los *Eventos* como los *Episodios* incluyen la propiedad *descripción*, que permite asignarles un nombre legible para los usuarios. El concepto de *Paciente* solo contiene propiedades de identificación y datos demográficos.

3.2.1.c. Definición de contacto

Además de la definición espacial y temporal de los movimientos de los pacientes en el hospital, en un análisis epidemiológico también es necesaria la definición de **contacto**. En esta sección aportamos una definición genérica en la que un contacto es la situación de que dos pacientes han estado la suficientemente cerca en espacio (física o lógicamente) y tiempo como para que sea posible la transmisión de una infección. La definición de contacto es intencionadamente flexible, de manera que nos permita identificar rutas desconocidas por las que se han propagado los brotes. Específicamente, consideramos que ha ocurrido un contacto entre dos pacientes cuando:

- Han estado en la misma *Localización*. Se consideran las *Localizaciones* desde el nivel de *Cama* hasta el nivel de *Área*.

- Han estado en dos *Habitaciones* contiguas que pertenecen a diferentes *Áreas*.
- Han estado en dos *Áreas* distintas que están incluidas en la misma *ZL*.
- Han sido atendidos por la misma *UH*.

Aunque el tiempo se representa mediante la combinación de la fecha y hora con una granularidad de segundos, para la definición de contacto utilizamos una representación discreta del tiempo con granularidad de días. Consideramos que ocurre un contacto cuando dos pacientes cumplen alguna de las condiciones anteriores dentro del mismo día, independientemente de la hora. Por ejemplo, consideramos que hay un contacto cuando dos pacientes hayan estado en la misma *Habitación*, pero uno en la mañana y otro en la tarde. Para modelar cuándo dos eventos coinciden en el tiempo utilizamos las relaciones intervalo-intervalo definidas en el álgebra de intervalos de Allen [4] y las relaciones punto-intervalo definidas en la extensión de Vilain & Kautz [209] (ver Figura 3.3).

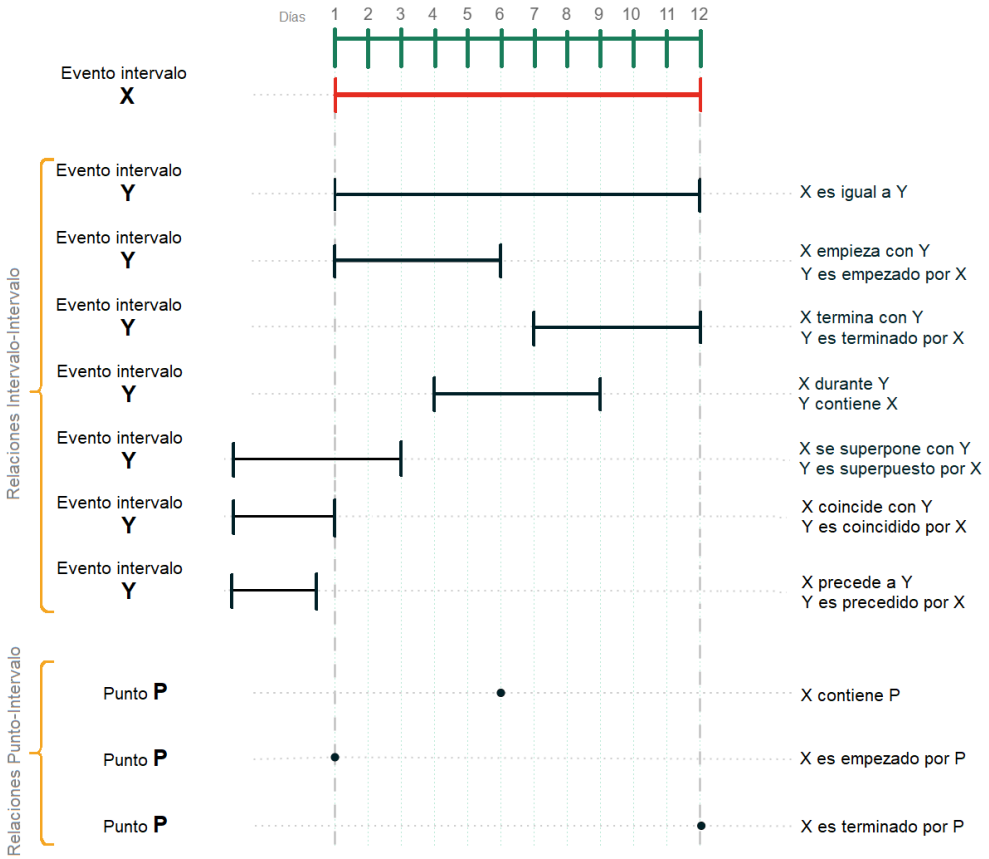


FIGURA 3.3. Relaciones intervalo-intervalo y relaciones punto-intervalo.

3.2.2. Consultas epidemiológicas

En este capítulo, proponemos seis consultas que representan diferentes tareas clínicas en la investigación epidemiológica. Hemos denominado estas consultas como *consultas epidemiológicas* (*CE*). El propósito de las dos primeras es detectar brotes, mientras que el de la tercera, cuarta y quinta es el análisis de contactos. La última consulta se centra en la identificación del caso índice de un brote.

Nuestro objetivo es que las CE exploten la semántica del modelo espacio-temporal epidemiológico definido en la Sección 3.2.1 para inferir las relaciones espaciales y temporales entre los pacientes. Para cada consulta, presentamos la tarea clínica que representa, una descripción basada en el modelo espacio-temporal, sus parámetros, los caminos que recorre y el resultado obtenido.

Un camino es la ruta ordenada de nodos y aristas que se recorren durante la ejecución de la consulta. Ha de indicarse que, aunque las aristas tengan dirección, en la consulta pueden ser recorridas en ambos sentidos. Una consulta puede tener varios caminos, ya que la ruta puede bifurcarse en ciertos nodos. Hemos definido dos tipos de caminos:

- **Caminos obligatorios**. Estos son los caminos que deben recorrerse para obtener una solución. Si una consulta tiene varios caminos obligatorios, solo se hallará una solución si ha sido posible recorrer todos ellos.
- **Caminos opcionales**. Estos caminos aparecen en bifurcaciones y representan diferentes maneras de obtener una solución. Son opcionales porque basta con que uno de los caminos alcance su último paso para que se obtenga un resultado. Sin embargo, si dos o más caminos opcionales pueden recorrerse completamente, se construyen tantas soluciones como caminos. Hay que tener en cuenta que, tras un conjunto de caminos opcionales, puede existir un camino obligatorio que reúna los últimos nodos de cada uno de ellos.

A continuación, mostramos cada una de las consultas epidemiológicas. En los caminos a recorrer aparece entre paréntesis el número de saltos (aristas atravesadas) que necesita el camino. Este número de saltos es tentativo, ya que en la implementación de las consultas en un lenguaje de consulta específico pueden variar tanto los caminos como su longitud.

- ***Consulta epidemiológica 1 (CE1)***
  - **Tarea clínica:** Detección de un brote en un *Servicio*.
  - **Descripción:** Dado un *Servicio*, un *Microorganismo* y un intervalo de búsqueda, buscamos todos los *Pacientes* que hayan tenido un *TestMicro* positivo para el *Microorganismo* dado, mientras estuvieron hospitalizados en dicho *Servicio*.
  - **Parámetros:**

        - Servicio ID
        - Microorganismo ID
        - Fechas de inicio y fin del intervalo de búsqueda
    - **Caminos a recorrer:**
        1. **Obligatorio** (3): Desde el *Microorganismo* dado hasta sus *Pacientes* conectados.
        2. **Obligatorio** (4): Desde el *Servicio* dado hasta sus *Pacientes* conectados.
    - **Resultado:** Un grafo sin aristas compuesto por los nodos de los *Pacientes* que cumplen con las condiciones de la consulta.

- ***Consulta epidemiológica 2 (CE2)***
    - **Tarea clínica:** Detección de un brote en una *Localización*.
    - **Descripción:** Dada una *Localización*, un número mínimo de pacientes infectados y un intervalo de búsqueda, buscamos todos los *Microorganismos* que hayan sido detectados al menos el número mínimo de veces en la *Localización* dada. La *Localización* puede ser de cualquier nivel de la jerarquía física. Además, el *TestMicro* debe haberse realizado mientras el paciente estaba en la *Localización*.
    - **Parámetros:**
        - Localización ID
        - Número mínimo de pacientes infectados
        - Fechas de inicio y fin del intervalo de búsqueda
    - **Caminos a recorrer:**
        1. **Obligatorio** (2): Desde los *Microorganismos* hasta sus *Episodios* conectados.
        2. **Obligatorio** (1): Desde los *Episodios* hasta sus *Pacientes*.
        3. **Obligatorio** (2): Desde los *Episodios* hasta sus *Eventos*, y desde estos hasta sus *Camas* a través de la arista *tieneLocalización*.
        4. **Obligatorio** (0-6): Desde las *Camas* hasta la *Localización* dada a través de aristas *situadoEn*.
    - **Resultado:** Un grafo que conecta los *Microorganismos* que cumplen con la condición de la consulta con sus respectivos *Pacientes* infectados.

- ***Consulta epidemiológica 3 (CE3)***
    - **Tarea clínica:** Estudio mediante análisis de contactos de la propagación de un brote desde un paciente.

- **Descripción:** Dado un *Paciente*, un *Microorganismo* y un intervalo de búsqueda, buscamos todos los *Pacientes* que estuvieron en contacto con el *Paciente* dado durante ese intervalo. Además, estos *Pacientes* deben haber tenido un *TestMicro* positivo para el *Microorganismo* dado dentro del período de búsqueda.
- **Parámetros:**
  - Paciente ID
  - Microorganismo ID
  - Fechas de inicio y fin del intervalo de búsqueda
- **Caminos a recorrer:**
  1. **Obligatorio** (2): Desde el *Paciente* dado hasta todos sus *Eventos*.
  2. **Obligatorio** (1): Desde los *Eventos* hasta las *Camas* en los que ocurrieron, a través de la arista *tieneLocalización*.
  3. **Opcional**: Desde las *Camas* hasta todas sus *Localizaciones* conectadas. Los caminos válidos son:
     a) (0-3) Desde las *Camas* hasta sus *Localizaciones* de nivel superior a través de las aristas *situadoEn* hasta *Área*.
     b) (2+4) Desde las *Camas* hasta sus *Habitaciones*, y de estas a otras *Habitaciones* contiguas que pertenezcan a diferentes *Áreas*.
     c) (5) Desde las *Camas* hasta sus *Áreas*, y de estas a otras Áreas que pertenezcan a la misma *ZL*.
  4. **Obligatorio** (0-3): Desde todas las **Localizaciones** del paso 3 hasta sus *Camas* conectados a través de la arista *situadoEn*.
  5. **Obligatorio** (3): Desde las *Camas* del paso 4 hasta sus *Pacientes* conectados.
  6. **Opcional** (1): Desde los *Eventos* del paso 1 hasta las *UH* conectadas mediante *tieneUnidadHosp*.
  7. **Opcional** (3): Desde las *UH* del paso 6 hasta sus *Pacientes* conectados.
  8. **Obligatorio** (3): Desde los *Pacientes* obtenidos en los pasos 5 y 7 hasta el *Microorganismo* dado.
- **Resultado:** Existen dos versiones de la consulta, según el tipo de resultado obtenido:
  1. Un grafo sin aristas compuesto por los nodos de los *Pacientes* resultado de la consulta.
  2. Un grafo en el que cada nodo de *Pacientes* resultado está conectado con los *Eventos* en los que el *Paciente* entró en contacto con el *Paciente* dado, así como los caminos de los pasos 2 a 7 que conectan a ambos *Pacientes*. Los nodos *Paciente* resultado también se conectan con los *TestMicro* con los que se encontró el *Microorganismo* dado.

- ***Consulta epidemiológica 4 (CE4)***
  - **Tarea clínica:** Estudio mediante análisis de contactos de la propagación de un brote desde un conjunto de pacientes.
  - **Descripción:** Dado un conjunto de *Pacientes*, buscamos todos los *Pacientes* que hayan estado en *contacto físico* con al menos uno de los *Pacientes* dados durante un intervalo de tiempo determinado. Por *contacto físico*, nos referimos a los contactos que ocurren a través de la parte física de la dimensión espacial.
  - **Parámetros:**
    - Conjunto de Paciente IDs
    - Fechas de inicio y fin del intervalo de búsqueda
  - **Caminos a recorrer:**
    1. **Obligatorio** (2): Desde los *Pacientes* dados hasta todos sus *Eventos*.
    2. **Obligatorio** (1): Desde los *Eventos* hasta las *Camas* en los que ocurrieron, a través de la arista *tieneLocalización*.
    3. **Opcional**: Desde las *Camas* hasta todas sus *Localizaciones* conectadas. Los caminos válidos son:
       a) (0-3) Desde las *Camas* hasta sus *Localizaciones* de nivel superior a través de las aristas *situadoEn* hasta *Área*.
       b) (2+4) Desde las *Camas* hasta sus *Habitaciones*, y de estas a otras *Habitaciones* contiguas que pertenezcan a diferentes *Áreas*.
       c) (5) Desde las *Camas* hasta sus *Áreas*, y de estas a otras Áreas que pertenezcan a la misma *ZL*.
    4. **Obligatorio** (0-3): Desde todas las **Localizaciones** del paso 3 hasta sus *Camas* conectados a través de la arista *situadoEn*.
    5. **Obligatorio** (3): Desde las *Camas* del paso 4 hasta sus *Pacientes* conectados.
  - **Resultado:** Un grafo en el que cada nodo *Paciente* resultado está conectado con:
    - Los *Eventos* en los que tuvo contacto físico con alguno de los *Pacientes* dados.
    - Los caminos de los pasos 2-5 que conectan ambos *Pacientes*.
- ***Consulta epidemiológica 5 (CE5)***
  - **Tarea clínica:** Investigación de las fuentes de contagio mediante análisis de contactos.
  - **Descripción:** Dado un conjunto de *Pacientes* y un intervalo de búsqueda, buscamos las *Localizaciones*, *Servicios*, *UH* y *Microorganismos* que han sido compartidos entre ellos. El objetivo es descubrir los contactos entre

los pacientes dados y determinar si fueron infectados por el mismo *Microorganismo*.

- **Parámetros:**
  - Conjunto de Paciente IDs
  - Fechas de inicio y fin del intervalo de búsqueda
- **Caminos a recorrer:**
  1. **Obligatorio** (2): Desde los *Pacientes* dados hasta todos sus *Eventos*.
  2. **Opcional**: Desde cada *Evento* del paso 1 hasta otros *Eventos* del paso 1 pertenecientes a un *Paciente* diferente. Los caminos válidos entre Eventos son:
     a) A través de sus *Localizaciones* compartidas, pasando por las *Camas* de cada Evento. Los caminos válidos son:
        i. (2-8) Desde las *Camas* hasta sus *Localizaciones* de nivel superior a través de las aristas *situadoEn* hasta *Área*.
        ii. (6+8) Desde las *Camas* hasta sus *Habitaciones*, y de estas a otras *Habitaciones* contiguas que pertenezcan a diferentes *Áreas*.
        iii. (10) Desde las *Camas* hasta sus *Áreas*, y de estas a otras Áreas que pertenezcan a la misma *ZL*.
     b) (2) A través de su *UH* compartida, a través de la arista *tieneUnidadHosp*.
     c) (4) A través de su *Servicio* compartido, a través de las aristas *tieneUnidadHosp* y *unidadHospDeServicio*.
     d) (2) A través del mismo *Microorganismo*, conectando los *Eventos* de la subclase *TestMicro* con otros *TestMicro* a través de la arista *haEncontrado*.
- **Resultado:** Un grafo con los nodos y aristas de los caminos que proporcionan una solución. Si un *Evento* obtenido en el paso 1 no está conectado a ningún otro *Evento* mediante alguno de los caminos del paso 2, el camino hacia ese *Evento* no aparecerá en la solución.

- ***Consulta epidemiológica 6 (CE6)***
  - **Tarea clínica:** Descubrimiento del *caso índice*.

    El caso índice es el primer paciente documentado en un estudio epidemiológico o la primera persona que llama la atención de los investigadores de salud, alertándolos sobre la posible aparición de un brote. Por tanto, el caso índice no debe confundirse con el *caso primario*, que es la primera persona que contrae la enfermedad y la introduce en la población. El caso primario no siempre coincide con el caso índice y, en ocasiones, nunca se llega a identificar [54].

- **Descripción:** Dado un *Microorganismo*, un conjunto de *Pacientes* y un intervalo de búsqueda, buscamos el primer paciente que tuvo un *TestMicro* positivo para el *Microorganismo* dado dentro del intervalo de búsqueda.
- **Parámetros:**
  - Conjunto de Paciente IDs
  - Microorganismo ID
  - Fechas de inicio y fin del intervalo de búsqueda
- **Caminos a recorrer:**
  1. **Obligatorio** (3): Desde los *Pacientes* dados hasta sus *TestMicro*, y desde estos hasta el *Microorganismo* dado.
- **Resultado:** Un grafo que conecta el nodo del *Paciente* índice con el *TestMicro* que lo identifica como caso índice. El nodo *TestMicro* aparece conectado con su propiedad *inicio*, indicando cuándo se realizó la prueba.

3.3 Selección de la tecnología para el almacenamiento y procesamiento de los datos

En la Sección 3.2.1 hemos propuesto un modelo de datos con el que poder analizar espacial y temporalmente la propagación de brotes infecciosos dentro de un hospital. Dadas la alta conectividad existente entre los conceptos que componen el modelo, hemos orientado su diseño a la creación de una red (también llamada grafo). Por tanto, nuestro objetivo no es el de crear redes que sirvan de estructura de soporte en la simulación de brotes, sino la de crear redes semánticamente complejas cuya información deba ser persistida y sirva de base para diversos tipos de análisis espacio-temporales o basados en técnicas de grafos.

El objetivo de esta sección es analizar las diferentes alternativas que se presentan para la implementación de redes con semántica. En concreto, tomamos como base nuestro modelo de datos y consultas epidemiológicas y evaluamos empíricamente el rendimiento de las bases de datos orientadas a grafos, centrándonos en los siguientes aspectos:

- Espacio de almacenamiento para la persistencia de los datos.
- Tiempo de ejecución de las consultas epidemiológicas.
- Memoria principal utilizada durante la ejecución de las consultas epidemiológicas.

Además, se también hemos tenido en cuenta la capacidad expresiva del lenguaje de consulta.

3.3.1. Alternativas

La aparición en los últimos años de la necesidad de almacenamiento y análisis de grandes conjuntos de datos, generalmente, sin esquemas rígidos ("*big data*") en todo tipo de dominios ha fomentado la aparición de nuevos tipos de bases de datos, englobadas bajo el término de bases de datos NoSQL [109]. Dentro del análisis de big data, ha aumentado el interés por el modelado de datos altamente interconectados como grafos debido a sus ventajas en el análisis de redes. Las bases de datos orientadas a grafos (BDOG) se presentan como un tipo de bases de datos NoSQL para el almacenamiento y gestión eficiente de este tipo de datos. Dentro de almacenamiento y procesamiento de datos en forma de grafo encontramos dos alternativas principales: los grafos de propiedades y los grafos de conocimiento (KG, del inglés, *knowledge graph*).

3.3.1.a. Grafos de propiedades

Los grafos de propiedades (GP) se definen en teoría de grafos como *multigrafos dirigidos etiquetados*. Un multigrafo es un grafo en el que está permitido definir varias aristas entre un mismo par de nodos. Al ser dirigido, estas aristas tienen un sentido definido, es decir, tienen un nodo de origen y un nodo de destino. Además, los nodos y aristas pueden tener asociadas una o varias etiquetas para definir sus propiedades. Estas etiquetas se definen como un par *clave-valor*, donde la *clave* es una cadena de texto *(string)* y el *valor* puede ser tanto un *string* como un valor numérico. Podemos establecer una analogía entre las BDOG dedicadas al almacenamiento de GP y la programación orientada a objetos (POO), dado que en ellas se suele distinguir entre dos tipos de etiquetas: las etiquetas de propiedades, que serían análogas a los atributos en POO, y las etiquetas de clase, con las que se definen las clases de nodos los y tipos de aristas.

En cuanto al almacenamiento de la información, las diferentes alternativas comerciales de bases de datos orientadas a grafos de propiedades (BDOGP) presentan sus propias estrategias para la representación interna de los grafos. Estas pueden estar orientadas al almacenamiento eficiente de los datos o a la consulta rápida de estos.

3.3.1.b. Grafos de conocimiento

La investigación en el desarrollo y uso de los grafos de conocimiento está muy relacionada con el campo de la web semántica, en la que se busca dotar de una definición formal al significado, propiedades y relaciones de los datos publicados, de manera que puedan ser interpretados por sistemas informáticos de procesamiento de datos. Los KG se utilizan como herramienta para la formalización de la información, de manera que se cree un modelo de datos autodescriptivo que con el uso de diferentes técnicas se pueda inferir y generar nuevo conocimiento [171]. No hay una definición clara y compartida del término "grafo de conocimiento" y, en muchas ocasiones, se utiliza indistintamente junto con el de "ontología" [53]. Podríamos definir un grafo de conocimiento como

un *grafo que se utiliza para describir entidades del mundo real (representadas como nodos) y las relaciones potencialmente diferentes entre estas entidades (representadas como aristas). Para ello, define en un esquema las posibles clases y relaciones de las entidades, permitiendo la interrelación potencial de entidades arbitrarias entre sí. No se limitan a entidades y relaciones abstractas, sino a todo tipo de elementos relevantes para un dominio específico, de manera que se cree un modelo de datos autodescriptivo* [39, 81].

Podemos destacar dos formatos para la representación de KG:

- **RDF.** Un modelo de datos estándar utilizado para la representación de KG es RDF (*Resource Description Framework*; en español, Marco de Descripción de Recursos) [47]. Este framework se basa en el concepto de *declaración* (en inglés, *statement*): una tripleta < *sujeto*, *predicado*, *objeto* >. El *sujeto* representa una entidad o concepto definido por un IRI (Internationalized Resource Identifier; en español, Identificador de Recursos Internacionalizado), y el *predicado* denota una propiedad o relación del sujeto. El *objeto* es el valor de la propiedad o el IRI de otro recurso. En una representación en forma de grafo, los sujetos y objetos serían los nodos, mientras que los predicados serían las aristas que los conectan. RDF define tres tipos de nodos: nodos *IRI*, utilizados para identificar las entidades y relaciones; *literales*, utilizados para representar los valores de las propiedades (se puede utilizar diferentes tipos de datos, como cadenas de texto, enteros, fechas, etc.); y *nodos en blanco* (en inglés, *blank nodes*), utilizados para indicar la existencia de una entidad. Nótese que en RDF no hay una distinción entre nodos IRIs de conceptos (clases y tipos de relaciones) y de entidades. Por ejemplo, la clase "*Habitación*" se representa como un nodo IRI que se conecta con cada una de las entidades concretas de habitaciones ("$Habitación/1$", "$Habitació/2$", etc.) mediante una relación de tipo "$rdf:type$" [29].

  Una diferencia de los KG en RDF con respecto a los GP es la forma en que se implementan las propiedades en las aristas. Mientras que, en los GP, las aristas pueden tener etiquetas con sus propiedades y valores de estas; en RDF, no es posible hacer un *statement* en el que una arista sea un sujeto. Por tanto, no hay manera de representar una arista con propiedades mediante un único *statement*. Existen cuatro enfoques para resolver este problema: reificación estándar, relaciones n-arias, propiedades aisladas (en inglés, *singleton properties*) y grafos con nombre (en inglés, *named graphs*) [13]. Para el análisis a realizar en esta sección, utilizamos la técnica de **reificación estándar** cuando debamos representar aristas con propiedades. Esta consiste en crear una nueva clase para representar el tipo relación. De este modo, se crean entidades concretas de esta nueva clase (nodos) sobre las que se puede definir propiedades y que se conectan a los nodos origen y destino originales. Es decir, se pueden crear tripletas en las que las aristas sean el sujeto y sus propiedades sean el objeto.

  Ha de indicarse que existe un lenguaje de consulta estandarizado para KG en formato RDF: **SPARQL** (acrónimo recursivo del inglés *SPARQL Protocol and*

RDF Query Language).

- **RDF*.** Motivada por la rigidez de la implementación mediante tripletas de RDF surge su extensión *RDF-STAR* (*RDF**) [13, 156]. Su principal novedad es que permite la creación de "*statements* sobre *statements*". De este modo un *statement* puede ser el sujeto o el objeto de otro *statement*, por lo que sería posible crear aristas con propiedades de una forma más compacta. Por ejemplo, se puede crear un *statement* $s1$ que indique que dos nodos están conectados, $< nodo1,\ relaciónA,\ nodo2 >$, y un *statement* $s2$ que tenga a $s1$ como sujeto y cuyo predicado y objeto sean el nombre y el valor de la propiedad de la arista, $< s1,\ propiedadA,\ valorPropiedadA >$.

  También se ha desarrollado una extensión de SPARQL para permitir la consulta sobre el nuevo tipo de *statements* surgidos con RDF*: SPARQL-STAR (SPARQL*) [13].

Las bases de datos orientadas a grafos de conocimiento implementan a este tipo de grafos íntegramente como un conjunto de tripletas RDF (o RDF*). De ahí que también se las llame ***motores de RDF*** (en inglés, *RDF engines*).

Para una mejor comprensión de las diferencias entre GP y KG, en la Figura 3.4 mostramos un grafo formado por dos nodos *Cama* y *Habitación* unidos por una arista *situadoEn*, la cual posee la propiedad *coste*. Este grafo se representa como un GP y como un KG en formato RDF y en formato RDF*. En el KG en formato RDF podemos observar la técnica de reificación estándar para representar la propiedad *coste* en la arista.

De aquí en adelante, para una unificación en la denominación de los componentes de los GP y KG, vamos a referirnos a los nodos de los GP y a los nodos que representan entidades en RDF y RDF* como "nodos de individuos", y a los nodos con el nombre de las clases y el valor de las propiedades como "nodos de literales".

3.3.1.c. Tecnologías de almacenamiento de grafos

En los últimos años, se han realizado diversas comparaciones entre distintas tecnologías concretas de BDOGP y motores de RDF. Estas comparaciones se centran en el rendimiento de las bases de datos, medido principalmente en función del tiempo de ejecución de un conjunto de consultas que abordan diferentes tareas basadas en grafos (adyacencia -pares de nodos conectados mediante una arista-, alcanzabilidad (del inglés, *reachability*) -capacidad de alcanzar un nodo del grafo desde otro nodo del grafo a través de las aristas definidas en este-, búsqueda de patrones, búsqueda del camino más corto, etc.) sobre conjuntos de datos reales o sintéticos de diversos tamaños. Los resultados de estos estudios dependen en gran medida de tres factores: el tipo de tarea, la estrategia de consulta interna de la BDOG y los conjuntos de datos. En [82] se analizó la equidad en los métodos más utilizados en la comparación de BDOG.

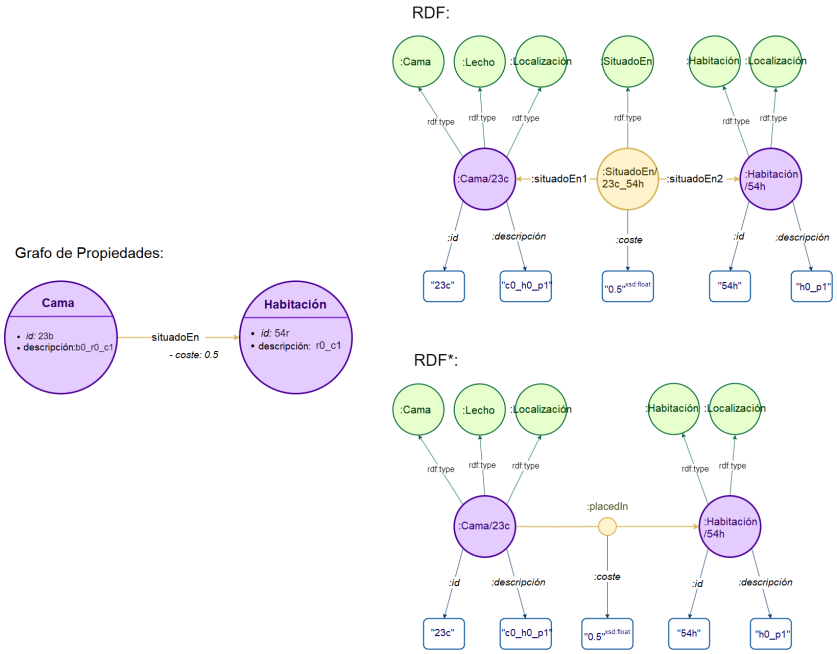


FIGURA 3.4. Representación de un grafo formado por dos nodos *Cama* y *Habitación* y la arista *situadoEn*. En la representación en RDF*, el pequeño nodo amarillo simboliza el *statement* de la unión entre los nodos morados y sobre el cual se crea un nuevo *statement* para añadirle la propiedad *coste*.

Los estudios comparativos de BDOG suelen centrarse en uno de los tipos de grafos: grafos de propiedades o grafos de conocimiento. En el caso de las **BDOGP**, *Neo4j* es la alternativa con más presencia en los estudios y la que presenta, en general, unos mejores resultados. En [214] se compararon tres BDOGP (*Neo4j*, *TigerGraph*, *TuGraph*) en redes sociales sintéticas. En [62] se compararon tres BDOGP (*ArangoDB*, *Neo4j*, *OrientDB*) y una base de datos relacional (*PostgreSQL*) utilizando un conjunto de datos basado en el Registro de Empresas de Italia. En [144] se analizaron el tiempo de ejecución, el uso de RAM y la carga de CPU en cuatro BDOGP (*JanusGraph*, *Nebula Graph*, *Neo4j*, *TigerGraph*) con datos del *LDBC* (Linked *Data Benchmark Council*, en español, *consejo de referencia de datos enlazados*) *Social Network Benchmark* [9].

En cuanto a los **motores de RDF**, *GraphDB* y *Stardog* se encuentran entre los motores de RDF que aparecen con mayor frecuencia y que ofrecen mejores resultados. En [14] se estudiaron la carga masiva, escalabilidad y tiempo de ejecución de consultas en cuatro motores de RDF (*GraphDB*, *Oracle 12c*, *Stardog*, *Virtuoso*) sobre dos conjuntos de datos reales de la Oficina de Publicaciones de la UE. En [24] se comparó el tiempo de ejecución de un conjunto de diez motores de RDF utilizando un conjunto de datos de referencia para ciudades inteligentes.

También hay estudios en los que se comparan BDOGP y motores de RDF. En [101] se compararon los tiempos de ejecución de dos BDOGP (*Neo4j*, *JanusGraph*) con un motor de RDF (*Blazegraph*) utilizando un conjunto de datos con información de Wikidata [211]. En [117] se analizó el tiempo de ejecución de cuatro motores de RDF

(*GraphDB*, *Virtuoso*, *RDF4j*, *Fuseki*) junto con una BDOGP (*Neo4j*) y un motor para el acceso a datos basados en ontologías (del inglés, Ontology-Based Data Access database) (*Ontop*) utilizando un conjunto de datos topológicos. En este trabajo, las consultas representan operaciones geométricas para el análisis espacial de puntos.

De entre las distintas opciones estudiadas en la literatura, hemos seleccionado las siguientes:

- **Neo4j**: Hemos seleccionado Neo4j como BDOGP. Se trata de una BDOGP basada en multigrafos dirigidos y etiquetados. Utiliza estructuras nativas para implementar los grafos, es decir, la estructura subyacente de la base de datos ha sido diseñada para almacenar datos en forma de nodos y aristas. En ella, los nodos y las aristas contienen su propia información en forma de propiedades. Neo4j presenta un modelo sin esquema, aunque sí que se vale de un mecanismo de etiquetas de clase para facilitar la definición de las consultas: los nodos y las aristas pueden pertenecer a una o varias clases que no restringen el número ni el tipo de datos de sus propiedades. Define su propio lenguaje de consulta, *Cypher*. Además, proporciona una API REST y controladores oficiales para varios lenguajes de programación.
- **GraphDB**: Para la selección de la base de datos para KG hemos valorado positivamente que se ofrezca soporte para el formato RDF*. GraphDB es totalmente compatible con RDF 1.1 y SPARQL 1.1, y es uno de los pocos motores de RDF comerciales que admite en su totalidad las extensiones RDF* y SPARQL*. GraphDB se implementa siguiendo el framework RDF4J (un framework de código abierto para el almacenamiento, consulta y análisis de datos en RDF que soporta SPARQL) con el objetivo de garantizar la compatibilidad con otros conjuntos de datos en y facilitar la integración en aplicaciones. GraphDB ofrece una API REST oficial para tareas de administración, aunque no para la consulta de la base de datos.

Cabe destacar que ambas bases de datos están implementadas en Java.

3.3.2. Pruebas comparativas (Benchmarks)

3.3.2.a. Consultas

Hemos evaluado el rendimiento de las BDOG en base a tres de las consultas epidemiológicas definidas en la Sección 3.2.2. Estas consultas pueden definirse como tareas de grafos de alcanzabilidad (del inglés, *reachability*) y búsqueda de patrones basadas en grafos, donde las rutas están condicionadas temporal y semánticamente. Hemos seleccionado consultas que presenten distintos niveles de complejidad (número de caminos obligatorios y opcionales, longitud de los caminos, número y tipo de variables a recuperar). La Figura 3.5 muestra una representación esquemática de cada una de las consultas seleccionadas.

Concretamente, hemos seleccionado las siguientes CE:

- **CE1**: Detección de un brote en un *Servicio*.
- **CE3**: Estudio mediante análisis de contactos de la propagación de un brote desde un paciente. Para esta consulta analizamos tanto la versión en la que la consulta devuelve los nodos *Paciente* conectados al pasado como parámetro, como la versión en la que se devuelve un subgrafo con los caminos recorridos.
- **CE5**: Investigación de las fuentes de contagio mediante análisis de contactos.

Cada una de las tres CE ha sido implementada en Cypher, SPARQL y SPARQL*.

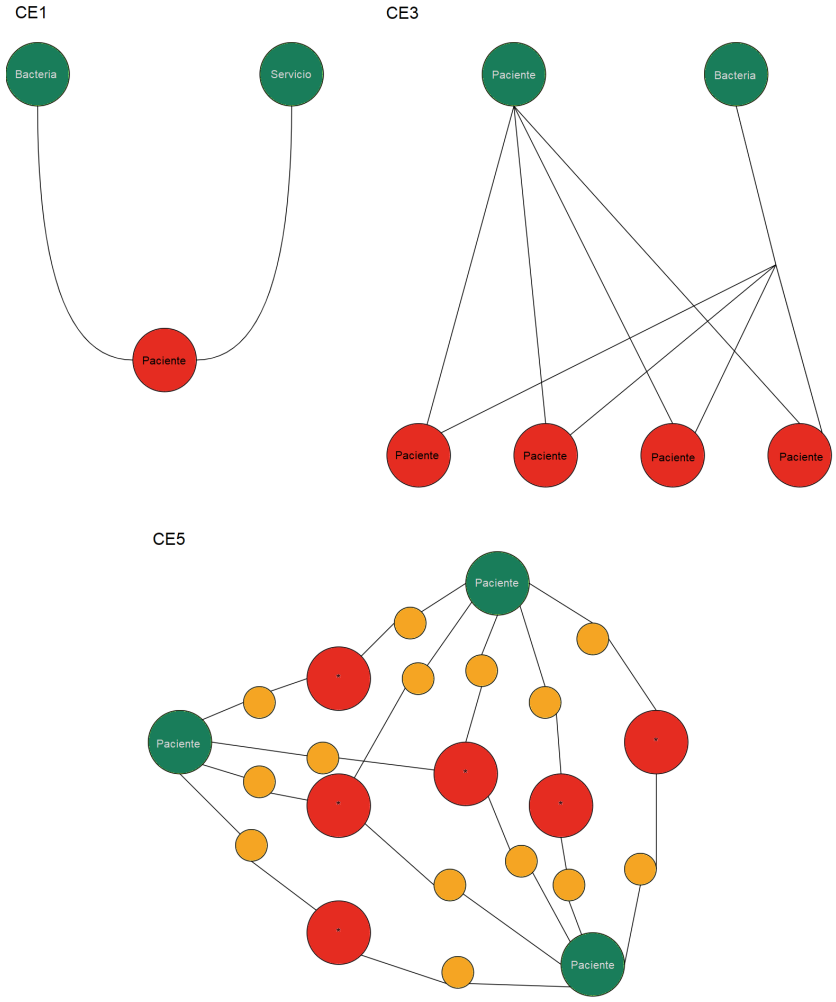


FIGURA 3.5. **Representación esquemática de las consultas a analizar**. Los **nodos verdes** representan los nodos iniciales proporcionados como parámetro, y los **nodos rojos** representan los nodos resultado tras recorrer los caminos obligatorios y opcionales, representados por las **líneas negras**. En el caso de ***CE5***, los **nodos rojos** tienen un **asterisco** porque pueden pertenecer a cualquier clase. Los **nodos amarillos** representan de manera simplificada a todos los nodos de los caminos recorridos, los cuales se devuelven junto con los **nodos rojos**.

3.3.2.b. Conjunto de datos

Como conjunto de datos de entrada para el análisis de las BDOG hemos utilizado la base de datos de libre acceso **MIMIC-III** [91], en la que se recopila información sobre pacientes ingresados en unidades de cuidados críticos de un gran hospital de atención terciaria. MIMIC-III incluye una gran variedad de tipos de datos como medicación, signos vitales, pruebas de laboratorio, códigos de diagnóstico o códigos de procedimiento, entre otros. Además, tiene información relativa a la duración, habitación y servicio de las hospitalizaciones. Esta última información puede ser adaptada a nuestro modelo de datos.

Hemos utilizado las tablas *admissions* (en español, *ingresos*) y *transfers* (en español, *transferencias*) para obtener las hospitalizaciones y los movimientos de los pacientes entre habitaciones, y la tabla *services* para identificar los servicios responsables de los pacientes. Hemos utilizado la tabla *microbiologyevents* (en español, *eventos microbiológicos*) para extraer las pruebas microbiológicas con resultados positivos.

Dado que la información proporcionada por MIMIC-III solo cubre parcialmente las clases y relaciones de nuestro modelo espacio-temporal, la hemos ampliado creando un hospital ficticio con una distribución sencilla.

En la **dimensión espacial física**, MIMIC-III trabaja con salas de hospitalización (*wards*). Para cada sala, hemos creado un conjunto de *Habitaciones* con un máximo de diez *Camas*. Es necesario que haya suficientes *Camas* en cada *Habitación* para abarcar el número máximo de pacientes hospitalizados simultáneamente si compactáramos todas sus hospitalizaciones en un solo año. Estas *Camas* se han distribuido homogéneamente entre las *Habitaciones*. Hemos organizado las *Habitaciones* en cuatro grupos (quirúrgicas, médicas, mixtas y para neonatos) según el tipo de servicio que atiende la mayoría de las hospitalizaciones en sus respectivas *Habitaciones*. Todas las *Habitaciones* están ubicadas en una única *Planta* y distribuidas en dos pasillos principales paralelos. Estos pasillos están divididos en varios *Pasillos* más cortos, cada uno con un máximo de veinte *Habitaciones*. Siguiendo este esquema, nuestro hospital ficticio cuenta con:

- 156 *Habitaciones quirúrgicas* con 1.413 *Camas*.
- 23 *Habitaciones médicas* con 179 *Camas*.
- 56 *Habitaciones mixtas* con 510 *Camas*.
- 4 *Habitaciones para neonatos* con 13 *Camas*.

La Figura 3.6 muestra una representación esquemática de la distribución del hospital, donde podemos ver los cuatro grupos de *Habitaciones* y su organización.

Para la **dimensión espacial lógica**, hemos asignado a cada *Servicio* tantas *UH* como fueran necesarias para garantizar que ninguna *UH* atiende a más de 1.500 *Hospitalizaciones*.

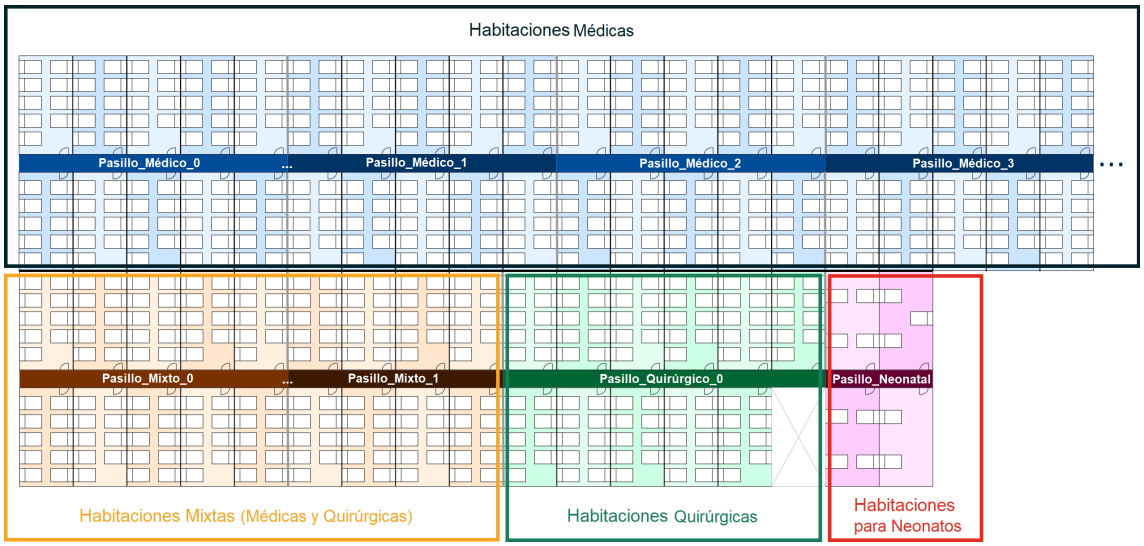


FIGURA 3.6. **Representación esquemática de la distribución de las habitaciones en el hospital ficticio para MIMIC-III.** Para simplificar la figura, cada pasillo (excepto el *Pasillo_Neonatal*) muestra la mitad del número real de *Habitaciones* asignadas. Además, solo se representa una porción del total de *Pasillo_Médico*.

TABLA 3.1
Clases con un alto impacto en la ejecución de las consultas.

Conjunto de Datos	# Pacientes	# Eventos	# Eventos con Localización	# TestMicro	# Microorganismos
G1	1.903	25.050	8.054	16.722	135
G2	5.702	75.039	24.612	49.560	197
G3	9.028	125.200	39.458	84.351	219
G4	14.027	173.307	61.862	109.379	227
G5	14.114	225.214	63.139	159.633	257
G6	15.471	275.625	71.090	201.674	288
G7	17.244	324.776	77.864	243.646	298
G8	20.032	375.211	98.471	281.996	303
G9	27.276	425.141	122.393	298.21	304
G10	38.066	474.922	169.435	299.870	304

Para evaluar la escalabilidad es necesario generar grafos de distinto tamaño. Tomando como base los datos de MIMIC-III y nuestro modelo de datos, hemos generado diez conjuntos de datos de distinto tamaño. Para crear los conjuntos de datos hemos modificado las fechas de inicio y fin de todas las hospitalizaciones y movimientos de pacientes de MIMIC-III (excepto los referidos a nuevos nacimientos, que no tienen asociados pruebas microbiológicas) para ocurran dentro de un mismo año, pero manteniendo su mismo orden que en MIMIC-III. Para cada grafo, hemos seleccionado un conjunto de eventos (junto con sus episodios, pacientes y microorganismos vinculados) de modo que el tamaño varíe desde 25.000 hasta 475.000 *Eventos*, aproximadamente. La distribución física y lógica del hospital es la misma en todos los conjuntos de datos, siendo la densidad de *Eventos* por *Localización* lo que varía. Hemos nombrado a los conjuntos de datos generados desde ***G1*** hasta ***G10***, donde *G1* es el más pequeño y *G10* el más grande. Hemos implementado cada conjunto de datos en un GP en Neo4j y en dos KG, uno formato RDF y otro en RDF*.

Debe mencionarse que se podría tomar de referencia un gran hospital en España el Complejo Hospitalario Universitario de La Coruña en Galicia, con aproximadamente 1.300 camas. Según el Instituto Nacional de Estadística de España, este hospital registra alrededor de 37.000 ingresos anuales, una cifra cercana a los 38.000 pacientes de *G10*.

La Tabla 3.1 muestra, para cada conjunto de datos, el número de entidades de las clases que tienen un impacto más significativo en la ejecución ya sea porque son las más pobladas o porque sus nodos son el inicio, fin o puntos de conexión de caminos en las consultas. La Tabla 3.2 presenta las principales características generales de los grafos generados a partir de cada conjunto de datos.

3.3.2.c. Espacio de almacenamiento

En cuanto al espacio en disco requerido para almacenar los grafos, hemos observado una diferencia significativa entre Neo4j y GraphDB, lo que podría ser un factor crítico para la escalabilidad de grafos de gran tamaño.

Para optimizar el almacenamiento de datos, Neo4j utiliza listas enlazadas de registros de tamaño fijo con las que se almacena cada tipo de dato (nodos, aristas y propiedades). Por su parte, GraphDB combina una colección de entidades (archivos en los que IRIs, literales y tripletas RDF* se almacenan como identificadores internos de 32 o 40 bits) con dos índices principales: el índice sujeto-predicado y el índice predicado-objeto.

La Figura 3.7 muestra el espacio en disco necesario para almacenar cada grafo en cada BDOG. En ella podemos comprobar que los grafos en RDF* ocupan un poco menos de espacio que en RDF. Esta diferencia varía entre un 6,2 % y un 14,4 %, aumentando conforme crece el tamaño del grafo.

Tabla 3.2
Características generales de los grafos. Los números están representados en miles..

Conjunto de Datos	Formato de Grafo	# Nodos de Individuos	# Nodos de Literales	# Aristas	Conjunto de Datos	Formato de Grafo	# Nodos de Individuos	# Nodos de Literales	# Aristas
G1	GP Neo4j	31	0	66	*G6*	GP Neo4j	312	0	643
	RDF	54	49	230		RDF	519	431	2.286
	RDF*	31	49	208		RDF*	312	431	2.079
G2	GP Neo4j	90	0	186	*G7*	GP Neo4j	365	0	750
	RDF	145	134	648		RDF	614	495	2.689
	RDF*	90	134	593		RDF*	365	495	2.440
G3	GP Neo4j	148	0	305	*G8*	GP Neo4j	422	0	866
	RDF	238	215	1.064		RDF	710	566	2.107
	RDF*	148	215	974		RDF*	422	566	2.819
G4	GP Neo4j	208	0	430	*G9*	GP Neo4j	490	0	1.088
	RDF	323	303	1.473		RDF	793	663	3.544
	RDF*	208	303	1.358		RDF*	490	663	3.240
G5	GP Neo4j	260	0	534	*G10*	GP Neo4j	565	0	1.168
	RDF	425	362	1.885		RDF	870	774	3.997
	RDF*	260	362	1.720		RDF*	565	774	3.692

Podemos destacar dos hechos:

- En RDF y RDF*, cada incremento en el tamaño del grafo representa un aumento aproximado de 55 MB.
- En los GP en Neo4j, este aumento oscila entre 5 MB en los grafos más grandes y 12 MB en los más pequeños.

Otro hallazgo que destacar es que, aunque el número de nodos y aristas se ha multiplicado por 17 de G1 a G10, el tamaño necesario para almacenar los grafos ha aumentado solo 7 veces en RDF y 6,5 veces en RDF*. Por el contrario, en Neo4j, G10 es 20 veces más grande que G1. Estos resultados sugieren que futuros estudios deberían analizar en qué punto los grafos almacenados en Neo4j superan en tamaño a los almacenados en GraphDB.

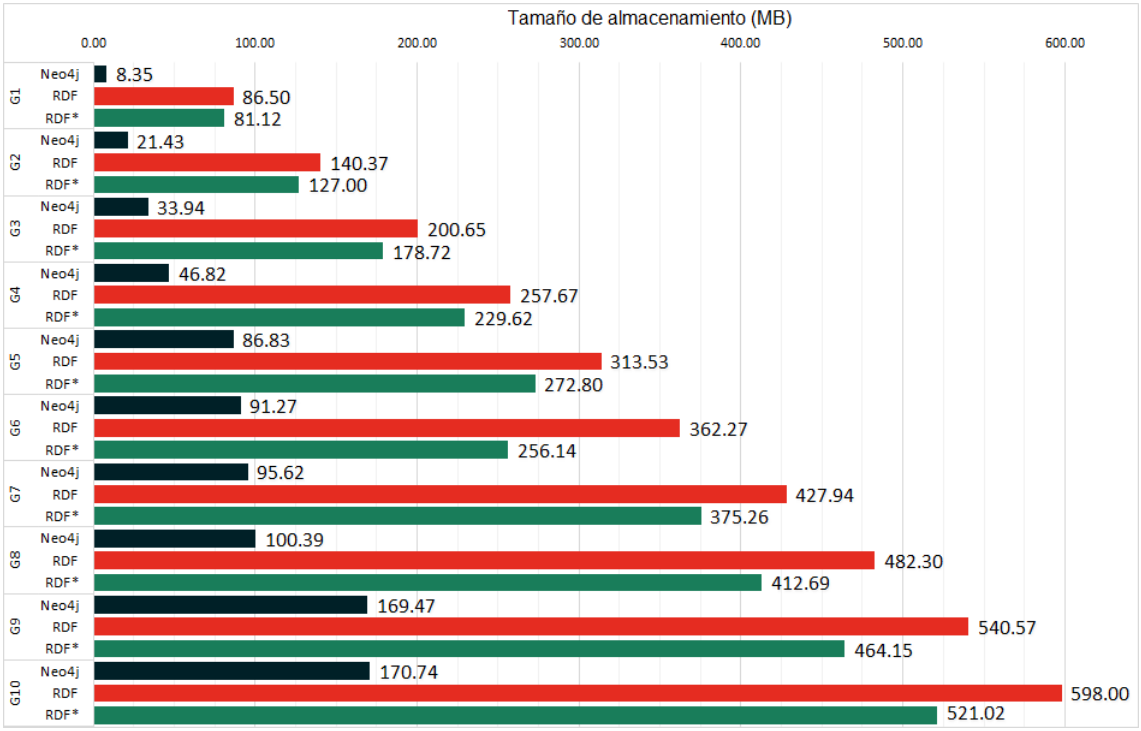


FIGURA 3.7. Espacio de almacenamiento utilizado por cada grafo en cada BDOG.

3.3.2.d. Características de las pruebas

Hemos establecido dos medidas representativas para analizar el rendimiento de la ejecución de consultas:

- **Tiempo de ejecución**: Representa el tiempo transcurrido entre el envío de la solicitud de consulta a la API de la BDOG y la recepción de la respuesta.
- **Memoria principal máxima**: Es la cantidad máxima de memoria principal requerida para ejecutar la consulta. Obtenemos este valor mediante un proceso en segundo plano que mide periódicamente el consumo de memoria principal de la BDOG. Tras cada ejecución de una consulta, borramos la memoria caché para garantizar que ésta no se conserva entre diferentes procesos, lo que falsearía los resultados registrados.

Vamos a referirnos al tiempo de ejecución como "**tiempo**" y a la memoria principal máxima como "**memoria**".

Hemos configurado tres pruebas comparativas (*benchmarks*) para estudiar el rendimiento de Neo4j y GraphDB. Cada benchmark evalúa una de las consultas (EC1, EC3, EC5) en los diez conjuntos de datos.

Para cada conjunto de datos, hemos definido 12 subconjuntos de datos, cada uno representando un mes del año. Es decir, las fechas pasadas como parámetro para definir

el período de búsqueda son el primer y último día de cada mes. Hemos ejecutado cada consulta tres veces para cada subconjunto. Los valores de tiempo y memoria para cada consulta son la media de los valores en todas las ejecuciones.

Para cada consulta, hemos seleccionado como parámetro de entrada un *Paciente*, *Microorganismo* o *Servicio* cuyo número de conexiones en el grafo es superior al tercer cuartil de los nodos de su clase. En las consultas que requieren un conjunto de pacientes, siempre utilizamos 15 pacientes. Consideramos que 15 es un número suficientemente alto para representar un brote significativo.

Hemos utilizado las siguientes versiones: **Neo4j CE 4.4.18** y **GraphDB Free 10.2.0**. Las solicitudes de consulta se han realizado localmente. Para la comunicación con las bases de datos, hemos utilizado: Neo4j Python Driver [152] para Neo4j y SPARQLWrapper [173] para GraphDB. Hemos realizado los experimentos en un PC con un procesador Intel i9-12900K, 16 GB de memoria RAM, una unidad de almacenamiento M2 de 1 TB. Hemos utilizado Linux Ubuntu 22.04 (64 bits) como sistema operativo.

3.3.2.e. B1: Consulta epidemiológica 1

CE1 consiste en la intersección de dos caminos obligatorios con una longitud fija.

En las Figuras 3.8.a y 3.8.b podemos ver que Neo4j presenta un mejor rendimiento en tiempo y memoria en todos los Gx. En cuanto a los KG, en la mayoría de los Gx, para RDF* se requiere menos tiempo y memoria que para RDF.

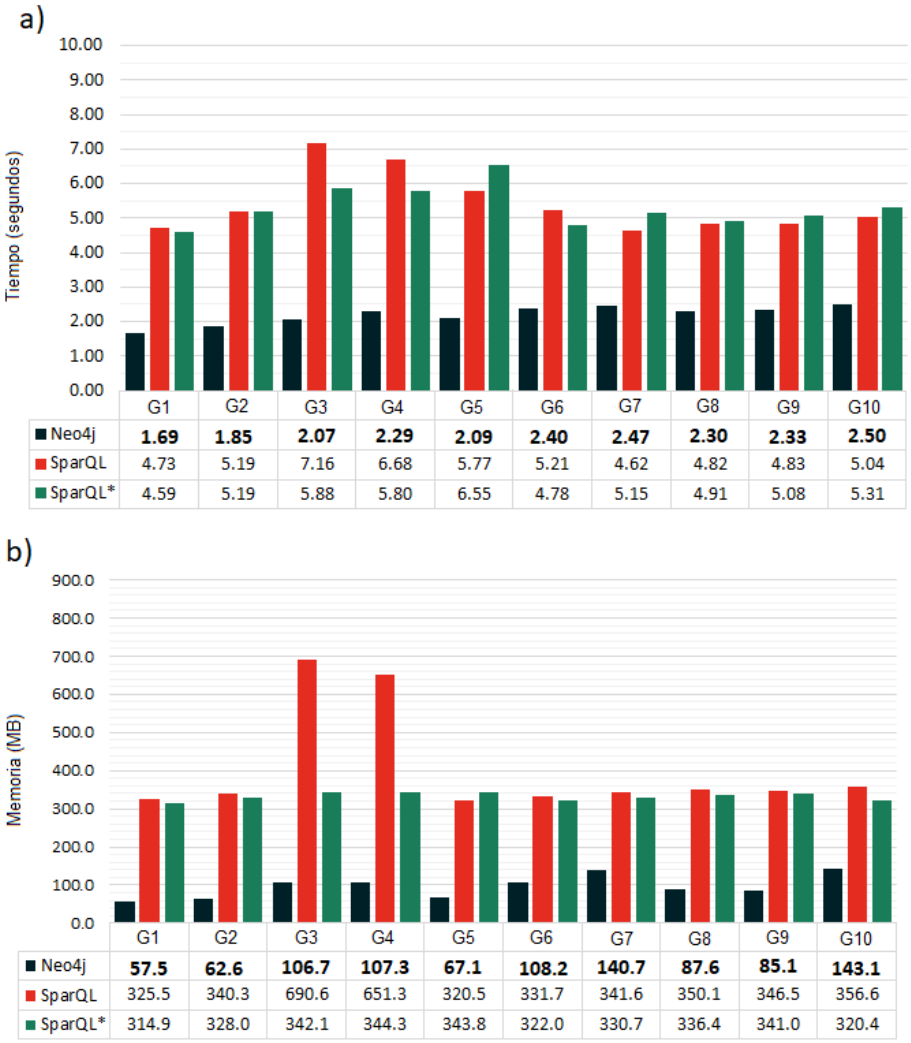
Podemos observar que, en general, tanto el tiempo como la memoria se mantienen estables en todos los Gx, a pesar de que las proporciones de *TestMicro* por *Microorganismo* y de *Eventos* por *Servicio* aumentan significativamente. La única excepción se encuentra en G3 y G4, donde la memoria en RDF se duplica respecto al resto de los Gx. Esta anomalía está presente en el resto de las consultas y podría estar relacionada con un cambio en el planificador de consultas.

En la Tabla 3.3 se muestra tanto el número promedio de resultados de la consulta en conjunto de datos Gx como el tiempo promedio para obtener un resultado. En ella podemos ver que el número de resultados va en incremento hasta G6, donde se mantiene estable hasta G10. Por tanto, el número de resultados no aumenta con el tamaño del grafo. De hecho, en estos últimos conjuntos de datos el tiempo promedio para obtener un resultado es muy similar y, además, menor que en los primeros Gx. Esto puede considerarse como una prueba del cambio en el planificador de consultas. Nótese que la cantidad de *Microorganismos* y *TestMicro* desde G6 a G10 es similar (ver Tabla 3.1). Esto nos da una pista de que en GraphDB no solamente el tamaño del grafo influye en el tiempo de las consultas, sino también los nodos y aristas a recorrer. Por su parte, en Neo4j el tiempo promedio por resultado se mantiene estable y menor que en GraphDB (entre un tercio y la mitad) para todos los conjuntos de datos.

TABLA 3.3

Comparativa entre Neo4j, GraphDB para RDF y GraphDB para RDF* en la ejecución de CE1. *Se muestra el tiempo medio de ejecución para los conjuntos de datos G1 a G10, así como el tiempo medio en obtener un resultado. Se han marcado en amarillo y verde los resultados similares. Además, el tiempo máximo relativo está resaltado en naranja.*

		G1	*G2*	*G3*	*G4*	*G5*	*G6*	*G7*	*G8*	*G9*	*G10*
	#Resultados	14	25	26	31	34	38	36	36	37	38
Neo4j	*Tiempo (s)*	**1,69**	**1,85**	**2,07**	**2,29**	**2,09**	**2,40**	**2,47**	**2,30**	**2,33**	**2,50**
	Tiempo/ #Resultados	0,12	0,07	0,08	0,07	0,06	0,06	0,07	0,06	0,06	0,07
GraphDB RDF	*Tiempo (s)*	4,73	5,19	7,16	6,68	5,77	5,21	4,62	4,82	4,83	5,04
	Tiempo/ #Resultados	0,33	0,21	0,28	0,22	0,17	0,14	0,13	0,13	0,13	0,13
GraphDB RDF*	*Tiempo (s)*	4,59	5,19	5,88	5,80	6,55	4,78	5,15	4,91	5,08	5,31
	Tiempo/ #Resultados	0,32	0,21	0,23	0,19	0,19	0,13	0,14	0,14	0,14	0,14



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Neo4j	**1.69**	**1.85**	**2.07**	**2.29**	**2.09**	**2.40**	**2.47**	**2.30**	**2.33**	**2.50**
SparQL	4.73	5.19	7.16	6.68	5.77	5.21	4.62	4.82	4.83	5.04
SparQL*	4.59	5.19	5.88	5.80	6.55	4.78	5.15	4.91	5.08	5.31

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Neo4j	**57.5**	**62.6**	**106.7**	**107.3**	**67.1**	**108.2**	**140.7**	**87.6**	**85.1**	**143.1**
SparQL	325.5	340.3	690.6	651.3	320.5	331.7	341.6	350.1	346.5	356.6
SparQL*	314.9	328.0	342.1	344.3	343.8	322.0	330.7	336.4	341.0	320.4

FIGURA 3.8. ***B1. CE1.*** **Comparativa de entre Neo4j, GraphDB para RDF y GraphDB para RDF* en la ejecución de CE1 en los grafos G1 a G10.** **a)** Comparativa del tiempo en segundos. **b)** Comparativa de la memoria en MB. Hemos marcado en negrita el valor más pequeño para cada conjunto de datos.

3.3.2.f. B2: Consulta epidemiológica 3. Resultado con sólo nodos *Paciente*

CE3 es una consulta que recorre un árbol completo desde un nodo *Paciente* hasta todos los nodos *Paciente* conectados a él a través de caminos opcionales que recorren todo el modelo de dominio. Cabe destacar que la proporción de *Eventos* por *Cama* y de *Eventos* por *UH* aumentan en cada grafo, lo que implica una "explosión" en los caminos a recorrer. Por lo tanto, CE3 es una consulta adecuada para comprender mejor cómo se comportan las BDOG seleccionadas al recorrer subgrafos de gran tamaño.

Las Figuras 3.9.a y 3.9.b muestran el tiempo y la memoria medios en la ejecución de CE3 por cada Cx. Excepto en los conjuntos más pequeños, GraphDB (tanto para RDF como para RDF*) presenta un tiempo menor que Neo4j, aumentando la diferencia con

el tamaño de los grafos. Si observamos el tiempo en Neo4j, hay dos subidas distintivas: la primera entre G1 y G5, y la segundo comienza en G7 y no parece alcanzar un máximo cercano. Además, podemos observar un comportamiento similar en la memoria.

Como en CE1, hemos calculado el tiempo promedio por resultado (nodos *Paciente* distintos conectados al nodo *Paciente* de entrada), el cual es mostrado en la Tabla 3.4. En ella podemos ver que, en Neo4j, el tiempo por resultado describe las mismas curvas que el tiempo total de ejecución. Sin embargo, podemos descubrir dos comportamientos distintos en cada subida: en la primera, el tiempo por resultado aumenta con el número de resultados; y en la segunda, aunque el número de resultados se mantiene en torno a 40, el tiempo por resultado presenta un crecimiento constante.

En el caso de GraphDB, el rendimiento para CE3 es muy similar al de CE1: crecimiento del tiempo y la memoria hasta G5, y bajada y estabilidad en los últimos grafos. Este hecho refuerza la hipótesis del cambio en la estrategia del planificador de consultas: el cambio comienza en G4 (grafo con alrededor de 500.000 nodos y 1.300.000 aristas) y se hace efectivo a partir de G6 (alrededor de 1.000.000 nodos y 2.000.000 aristas).

3.3.2.g. B2: Consulta epidemiológica 3. Resultado con subgrafo de caminos

El hallazgo más distintivo de esta versión de CE3 en la que se devuelve un subgrafo con todos los caminos recorridos por completo hasta los nodos *Paciente* resultado es el incremento significativo del tiempo y la memoria en todos los conjuntos de datos. Estos pueden verse en las Figuras 3.9.c y 3.9.d. Neo4j presenta las mismas curvas que en la versión anterior de la consulta, pero con un incremento aproximado del 30 % en el tiempo y del 20 % en la memoria.

En GraphDB el comportamiento del tiempo y la memoria es muy diferente. La memoria se mantiene similar, con un incremento medio ligero del 5 %. En cuanto al tiempo, el incremento es del 25 % hasta G5, y del 55 % en los restantes.

Los comportamientos de cada consulta pueden explicarse en base a la implementación de esta versión de CE3. En Neo4j ha sido necesario añadir instrucciones explícitas para guardar los caminos recorridos en forma de listas. Esto explicaría el aumento significativo en memoria. En GraphDB no ha sido necesario utilizar instrucciones explícitas para guardar los caminos recorridos, pero sí la creación de una consulta conectada a la original mediante una sentencia UNION para no perder los caminos que unen los nodos *Localización*. Esta nueva consulta no debería implicar un coste relevante en memoria, ya que los nodos y aristas atravesados deben estar cacheados. Sin embargo, la unión de los resultados de ambas consultas debe suponer un alto coste en tiempo de ejecución.

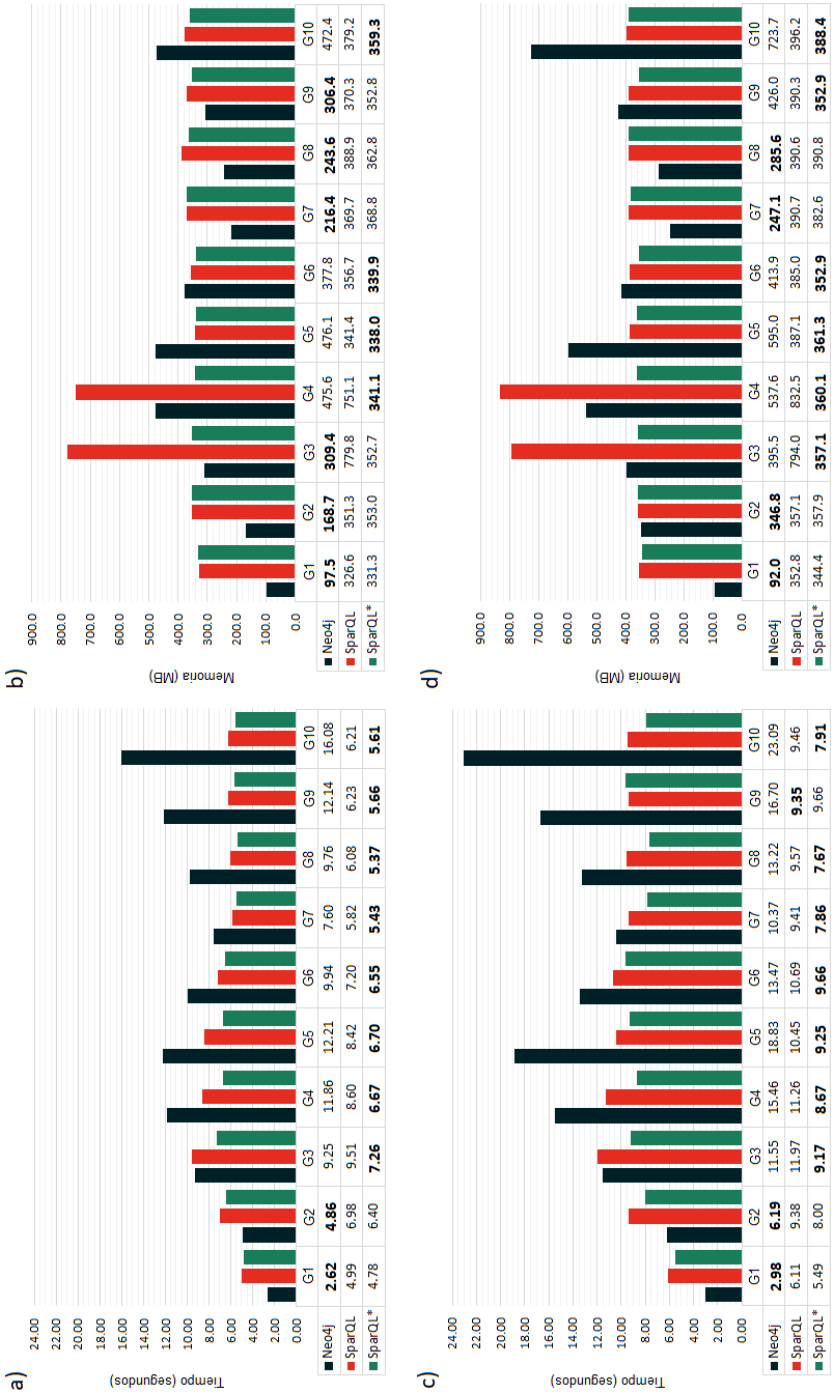


FIGURA 3.9. ***B2. CE3.*** **Comparativa de entre Neo4j, GraphDB para RDF y GraphDB para RDF* en la ejecución de CE3 en los grafos G1 a G10**. **a)** Comparativa del tiempo en segundos para la versión que solo devuelve nodos *Paciente*. **b)** Comparativa de la memoria en MB para la versión que solo devuelve nodos *Paciente*. **c)** Comparativa del tiempo en segundos para la versión que devuelve un subgrafo con los caminos recorridos. **d)** Comparativa de la memoria en MB para la versión que devuelve un subgrafo con los caminos recorridos. Hemos marcado en negrita el valor más pequeño para cada conjunto de datos.

TABLA 3.4

Comparativa entre Neo4j, GraphDB para RDF y GraphDB para RDF* en la ejecución de CE3. *Se muestra el tiempo medio de ejecución para los conjuntos de datos G1 a G10, así como el tiempo medio en obtener un resultado. Se han marcado en amarillo y verde los resultados similares. Además, el tiempo máximo relativo está resaltado en naranja.*

		G1	*G2*	*G3*	*G4*	*G5*	*G6*	*G7*	*G8*	*G9*	*G10*
	#Resultados	13	26	29	33	34	43	40	40	36	41
Neo4j	*Tiempo (s)*	**2,62**	**4,86**	9,25	11,86	12,21	9,94	7,60	9,76	12,14	16,08
	Tiempo/ #Resultados	0,20	0,19	0,31	0,36	0,36	0,23	0,19	0,25	0,34	0,39
GraphDB RDF	*Tiempo (s)*	4,99	6,98	9,51	8,60	8,42	7,20	5,82	6,08	6,23	6,21
	Tiempo/ #Resultados	0,37	0,27	0,32	0,26	0,25	0,17	0,14	0,15	0,17	0,15
GraphDB RDF*	*Tiempo (s)*	4,78	6,40	**7,26**	**6,67**	**6,70**	**6,55**	**5,43**	**5,37**	**5,66**	**5,61**
	Tiempo/ #Resultados	0,36	0,25	0,25	0,21	0,20	0,15	0,13	0,14	0,16	0,14

3.3.2.h. B3: Consulta epidemiológica 5

CE5 recorre los árboles desde 15 nodos *Paciente* dados hasta sus nodos *Evento*. Luego, se definen caminos opcionales para buscar la unión entre los *Eventos* de diferentes *Pacientes*. Por lo tanto, **es una consulta donde la exploración del grafo no es tan crucial como el paso repetitivo sobre un conjunto de nodos y aristas**. Por tanto, el rendimiento en tiempo y memoria (ver Figuras 3.10.a y 3.10.b) debería diferir de las otras dos consultas.

Este es el caso de Neo4j. En memoria presenta una memoria de aproximadamente 120MB (con algunas excepciones en G3 y G4 que podrían estar relacionadas con la estrategia del planificador de consultas), un valor bajo en comparación con las dos versiones de CE3, en las que se consume entre 90MB y 600MB. Sin embargo, el tiempo se mantiene elevado, en torno a los 8 segundos. Esta diferencia entre tiempo y memoria podría explicarse porque los nodos y aristas entre los *Eventos* quedan cacheados en memoria la primera vez que se acceden y permanecen en caché para el resto de los accesos.

En el caso de GraphDB, la memoria presenta valores similares a los de CE1 y CE3, mientras que el tiempo es mucho mayor que en cualquier otra consulta.

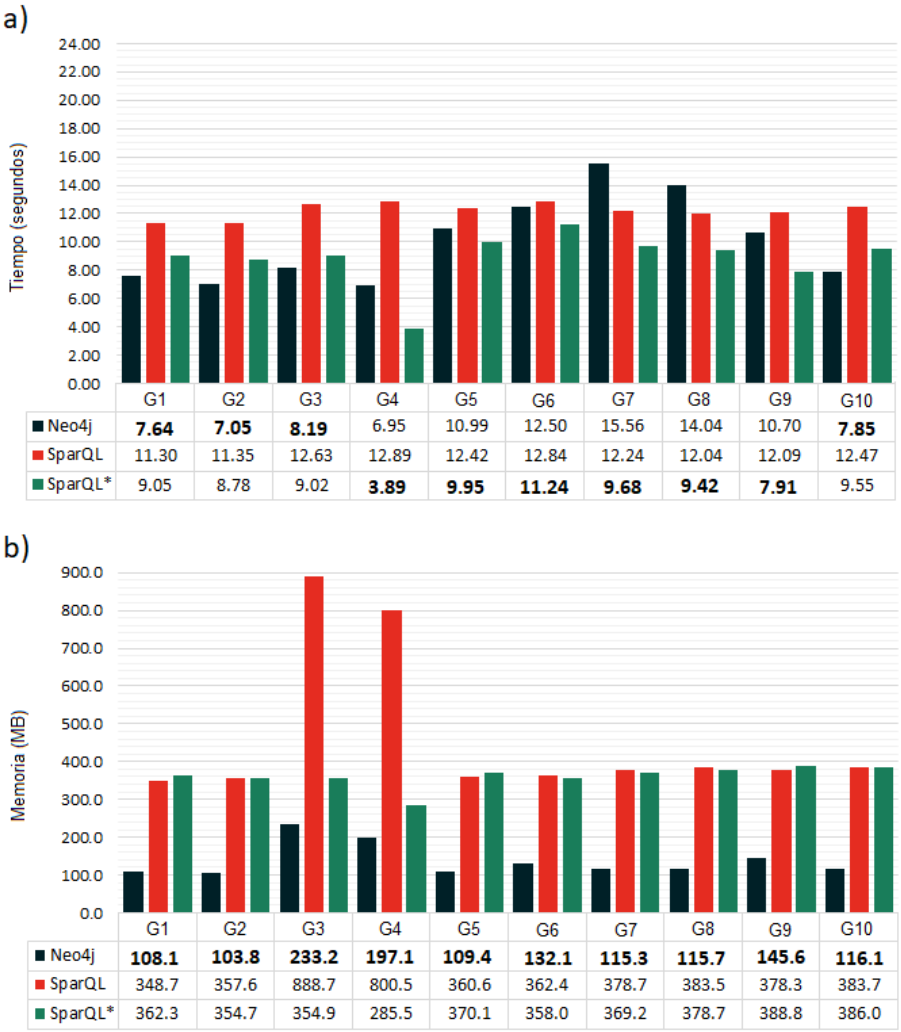
La Tabla 3.5 muestra la cantidad media de resultados (cantidad de rutas distintas que conectan un *Paciente* con otro) de CE5 en cada Gx. Cabe señalar que, aunque para los experimentos hemos seleccionado los *Pacientes* con el mayor número de conexiones, esto no significa que necesariamente existan conexiones entre los *Pacientes*. Además, un grafo más grande no implica una mayor conexión entre los de *Pacientes* entrada. De ahí que podamos ver en la Tabla 3.5 que en número de resultados no aumente con el tamaño de los grafos. De hecho, el máximo número de resultados se encuentra en G7 y el mínimo en G4 (marcados en negrita).

En Neo4j, el tiempo sigue la misma curva que el número de resultados: cuando el número de resultados aumenta, el tiempo también lo hace. Curiosamente, en GraphDB, el tiempo en las consultas sobre los grafos en RDF* presenta una curva más similar a la de Neo4j que la de los grafos en RDF. De hecho, las diferencias de tiempo entre RDF y RDF* son más notorias en esta consulta que en el resto: en RDF el tiempo de ejecución de la consulta es entre un 115 % y un 153 % mayor que en RDF*. Esta diferencia podría deberse a que la mayor parte del cómputo de CE5 lo conforma el recorrido de aristas con propiedades. Recuérdese que, en RDF*, estas aristas se representan con una arista (un *statement* sobre otro *statement*), mientras que en RDF se necesita crear un "nodo de individuo" y dos aristas (ver Figura 3.4).

Por tanto, podemos concluir que los KG en formato RDF* permiten un recorrido más rápido de las aristas con propiedades. Bajo esta suposición, podemos notar que en CE1, donde las aristas con propiedades son minoría, la diferencia de tiempo entre ambos tipos de KG no es tan pronunciada (incluso en algunos Gx, la consulta es más rápida sobre RDF).

TABLA 3.5
Número medio de resultados en la ejecución de CE5 para los conjuntos de datos G1 a G10.

Conjunto de Datos	Nº Resultados	Conjunto de Datos	Nº Resultados
G1	1683	*G6*	3030
G2	1340	*G7*	**4138**
G3	1598	*G8*	3715
G4	**1060**	*G9*	2595
G5	2453	*G10*	1517



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Neo4j	**7.64**	**7.05**	**8.19**	6.95	10.99	12.50	15.56	14.04	10.70	**7.85**
SparQL	11.30	11.35	12.63	12.89	12.42	12.84	12.24	12.04	12.09	12.47
SparQL*	9.05	8.78	9.02	**3.89**	**9.95**	**11.24**	**9.68**	**9.42**	**7.91**	9.55

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Neo4j	**108.1**	**103.8**	**233.2**	**197.1**	**109.4**	**132.1**	**115.3**	**115.7**	**145.6**	**116.1**
SparQL	348.7	357.6	888.7	800.5	360.6	362.4	378.7	383.5	378.3	383.7
SparQL*	362.3	354.7	354.9	285.5	370.1	358.0	369.2	378.7	388.8	386.0

FIGURA 3.10. ***B3. CE5.*** **Comparativa de entre Neo4j, GraphDB para RDF y GraphDB para RDF* en la ejecución de CE5 en los grafos G1 a G10.** **a)** Comparativa del tiempo en segundos. **b)** Comparativa de la memoria en MB. Hemos marcado en negrita el valor más pequeño para cada conjunto de datos.

3.3.3. Discusión y selección de tecnología

Hemos analizado el rendimiento (tiempo y memoria) en la ejecución de las consultas en dos BDOG: Neo4j, para los GP, y GraphDB, para los KG en formato RDF y RDF*.

En cuanto a la memoria máxima consumida en la ejecución de las consultas podemos destacar los siguientes hechos:

- En Neo4j, el uso de memoria no depende tanto del tamaño total del grafo, sino del subgrafo a recorrer: a medida que se recorren nuevas aristas y nodos, estos se almacenan en memoria y no se eliminan hasta el final de la consulta para evitar el coste de recuperarlos de la memoria secundaria. Para consultas simples, como CE1, el consumo máximo de memoria está entre 60MB y 150MB; mientras que para consultas que requieren mayor exploración del grafo, se han utilizado hasta 500MB. El hecho de tener que almacenar nodos y aristas para devolverlos como resultado supone un gran coste en términos de memoria.
- En GraphDB, el consumo máximo de memoria es similar tanto para RDF como para RDF*. A diferencia de Neo4j, este no fluctúa, sino que se mantiene constante entorno a 350MB y 400MB. La variación en memoria depende tanto de la complejidad de la consulta y datos a almacenar durante su ejecución como del tamaño del grafo. Por ejemplo, en CE3 y CE5, la memoria aumenta gradualmente de G1 a G10, aunque los subgrafos recorridos no son necesariamente más grandes por ser un grafo de mayor tamaño.

En relación con el tiempo de ejecución, también hemos observado una clara diferencia entre el rendimiento de Neo4j y de GraphDB. Mientras que en Neo4j el tiempo varía dependiendo del tipo de consulta y el tamaño del subgrafo a recorrer, en GraphDB las oscilaciones son más leves. En CE1 y CE3, el tiempo en los grafos más pequeños es elevado (alrededor de 5 segundos, en comparación con los 2 segundos de Neo4j), pero en los grafos de mayor tamaño este no llega a ser triplicarse (alrededor de 15 segundos). En contraste, en Neo4j, los tiempos máximos pueden multiplicar el mínimo por ocho.

En cuanto al almacenamiento de los datos, Neo4j ha requerido entre 4 y 10 veces menos de espacio que GraphDB, pero su proporción de crecimiento es mucho mayor respecto al número de nodos.

Podemos resumir que, para consultas simples, Neo4j presenta un mejor rendimiento que GraphDB, con un consumo de tiempo y memoria que puede ser hasta la mitad. Sin embargo, aunque GraphDB tiene un consumo inicial más alto, su crecimiento es más estable, lo que le otorga una mejor capacidad de escalabilidad en consultas que recorren grandes subgrafos. Por consiguiente, **hemos seleccionado GraphDB como BDOG más prometedora para nuestra investigación**.

Cabe destacar que RDF y RDF* son estándares de W3C (*Consorcio WWW*; en inglés, *World Wide Web Consortium*) y no formatos propietarios, como lo es Neo4j. Además, su lenguaje de consulta también es un estándar de libre uso. Esto supone

la ventaja de tener una mayor flexibilidad en dos aspectos: variedad de herramientas adicionales para trabajar con KG (extensión del lenguaje de consulta con módulos propios, de la comunidad u oficiales de la BDOG, APIs y drivers para la conexión con la BDOG, librerías y paquetes en otros lenguajes de programación para el manejo de los datos obtenidos de la BDOG) y solidez a la hora de elegir entre diferentes herramientas adicionales para trabajar con estas tecnologías, posibilidad de desarrollo de un software de almacenamiento y gestión de tripletas RDF propio, e integración sencilla con otros KG ya existentes.

3.3.3.a. RDF vs RDF*

Al comparar los grafos en formato RDF y los que están en RDF*, encontramos que su implementación más sencilla de las aristas con propiedades conlleva una menor necesidad de espacio de almacenamiento, así como un menor consumo y memoria en la ejecución de las consultas. Por consiguiente, **hemos implementado el modelo como un grafo de conocimiento en formato RDF*** y **el lenguaje de consulta para la implementación de las consultas epidemiológicas ha sido SPARQL***.

Existen diferentes formatos de archivo para la serialización RDF (Turtle [22], TriG [28], N-Triples [21], N-Quads [34], JSON-LD [197]), todos ellos basados en la creación de *statements* en forma de tripletas. Muchos de estos formatos también tienen su versión *STAR*, que permanece igual, con la única diferencia de agregar una sintaxis adicional para crear *statements* sobre *statements*. En nuestro caso, el modelo utilizado el formato de archivo *Turtle* (*.ttl*). En este formato, una tripleta se define de la siguiente forma: $< sujeto > < predicado > < objeto >$. La sintaxis extra para la creación de un *statement* sobre un *statement* es muy sencilla: un *statement* previamente definido se sitúa como sujeto u objeto de otro *statement*, siendo envuelto por $\ll$ y $\gg$. Por ejemplo: $\ll < sujeto > < predicado > < objeto > \gg < pred2 > < obj2 >$. Esta representación es significativamente más simple que las diversas técnicas aplicadas en RDF, como la reificación estándar, en la que es necesario agregar nuevas clases para representar las aristas con propiedades.

SPARQL es un lenguaje de consulta con una sintaxis sencilla, también basada en la creación de tripletas para describir los caminos a recorrer y la información a recuperar. Además, cuenta con varios operadores para filtrar los caminos y permite el uso de expresiones regulares. SPARQL*, la extensión de SPARQL para RDF*, permite consultar *statements* sobre *statements* con el mismo mecanismo con el que se definen en RDF*: envolviendo el primer *statement* entre $\ll$ y $\gg$. Sin embargo, SPARQL y SPARQL* no ofrecen operadores nativos para operaciones aritméticas sobre duraciones, fechas y horas. Varios motores RDF, como GraphDB, proporcionan extensiones para estas operaciones o permiten crear funciones definidas por el usuario.

3.4 Validación

Vamos a demostrar la idoneidad para la investigación epidemiológica del modelo espacio-temporal y las CE presentados a través de dos experimentos en los que analizamos dos brotes de *Clostridium difficile* (*C. diff*). Nuestro objetivo es representar dos casos de uso en los que se busca descubrir el origen del brote infeccioso mediante el uso secuencial de varias de las CE presentadas.

3.4.1. Conjunto de datos y herramientas

Hemos diseñado dos experimentos en los que varios brotes de *C. diff* ocurren en un hospital durante un período de seis meses.

Dada la dificultad de obtener datos reales que tengan el nivel de detalle similar al descrito por nuestro modelo espacio-temporal, hemos optado por utilizar un modelo de simulación realista de infecciones hospitalarias para la generación de datos sintéticos [98]. Este modelo de simulación genera un flujo de pacientes a través de las camas y servicios de un hospital. Este modelo presenta un tiempo discreto con pasos de 8 horas, representando tres turnos de trabajo: mañana (8:00-15:59), tarde (16:00-23:59) y noche (00:00-7:59). Se trata de un modelo de simulación basado en agentes en el que cada agente representa a un paciente. En cada paso de la simulación, un paciente puede permanecer en su cama o trasladarse a otra, creando así un nuevo *Evento*.

Además, durante la simulación, varios brotes ocurren basados en el modelo epidemiológico compartimental SEIRD (ver Sección 2.3.2). Cada agente almacena en qué estado definido por SEIRD se encuentra el paciente. El contagio de la infección puede darse a través del aire o por el contacto con superficies contaminadas. En cada paso, basándose en un conjunto de probabilidades, se calcula si cada paciente cambia de estado o no. Los *Eventos TestMicro* se han creado utilizando la fecha y hora en la que el estado del paciente pasa a estado *Infectado*.

Siguiendo los resultados de la sección 3.3, hemos implementado nuestro modelo epidemiológico como un grafo de conocimiento en RDF*. También hemos adaptado la salida del modelo de simulación hospitalario a nuestro modelo de datos de manera que no queden conceptos de la dimensión espacial vacíos y con ella hemos poblado el grafo de conocimiento.

Para generar los conjuntos de datos hemos utilizado los mismos valores que en el experimento de [98] para los parámetros que describen el comportamiento epidemiológico de la infección. En ambos experimentos hemos utilizado el mismo conjunto de datos, el cual cuenta con **7.883 pacientes hospitalizados entre el 1 de enero y el 30 de junio**. Sus *Eventos* se clasifican en: *Hospitalizaciones*, *Cirugías*, *Radiografías*, *TestMicro* y *Muertes*.

En cuanto al tamaño y estructura del hospital, hemos creado un hospital con **1.087 *Camas*** distribuidas en **5 *Plantas*** de la siguiente manera:

- La *Planta* baja alberga todas las *Habitaciones* dedicadas a los *Servicios* y *UH* de urgencias, Unidad de Cuidados Intensivos (UCI), así como los quirófanos y salas de radiografía. Esta *Planta* se divide en 3 *Áreas* (1 *Unidad* y 3 *Bloques*). Cada *Habitación* de UCI y urgencias tiene 10 *Camas* (dos filas enfrentadas con cinco *Camas* cada una). Cada quirófano y sala de radiografía tiene 1 *Cama*.

- Las cuatro plantas superiores sólo tienen *Habitaciones* para *Hospitalizaciones*. Cada *Planta* tiene 12 *Áreas* organizadas en 3 *Unidades* y 4 *Bloques*. Cada *Habitación* tiene 2 *Camas* contiguas.

En el hospital hay 12 *Servicios* que se organizan de la siguiente manera:

- UCI, Urgencias, Cirugía y Radiología tienen un único *Servicio* con una única *UH* cada uno.

- Hay 8 *Servicios* para el resto de hospitalizaciones, con 3 *UH* cada *Servicio*.

Además, hemos definido 3 *ZL*: una en la planta baja, otra en la segunda planta y otra en la cuarta planta.

La Tabla 3.6 muestra el número de nodos y aristas del grafo, así como el número de nodos de algunas clases significativas en la ejecución de las consultas.

Tabla 3.6
Número de nodos y aristas del grafo utilizado para los experimentos.

Características generales	**# Nodos**	104.993
	# Nodos de individuos	66.939
	# Nodos de literales	38.054
	# Aristas	184.802
Características específicas	**# Pacientes**	7.883
	# Eventos	16.055
	# Eventos con Localizaciones	14.479
	# Hospitalizaciones	13.574
	# Cirugías	364
	# Radiografías	541
	# TestMicro	1.023
	# Localizaciones	1.750
	# Camas	1.087
	# Habitaciones	509
	# Pasillos	94
	# Areas	51
	# Plantas	5
	# Servicios	12
	# UH	28

Hemos utilizado GraphDB 10.4.0 para almacenar el grafo y ejecutar las consultas en los experimentos. Las figuras de los experimentos han sido creadas con la herramienta 'visual graph' de GraphDB Workbench.

3.4.2. Experimentos

Para cada experimento mostramos su objetivo, las consultas epidemiológicas utilizadas y los resultados obtenidos con cada una de ellas, los cuales sirven de entrada para la siguiente consulta a ejecutar.

3.4.2.a. Experimento 1: Búsqueda del origen de un brote

En este experimento, consideramos que existe una tarea programada que verifica la presencia de brotes diariamente. Si en un mismo día hay más de dos pruebas microbiológicas positivas para el mismo microorganismo en una misma *Planta*, se genera una alerta de posible brote.

Objetivo: El objetivo principal de este experimento es determinar si había un brote en una *Localización* y, en caso afirmativo, identificar su posible origen.

Consultas epidemiológicas: Hemos utilizado **4 CE** en el siguiente orden:

1. **CE2**: Para detectar un posible brote en una *Planta*.
2. **CE4**: Para identificar los *Pacientes* que tuvieron contacto físico con los pacientes detectados en CE2 y determinar dónde ocurrieron estos contactos. Esta información permitiría formular una hipótesis inicial sobre el origen del brote.
3. **CE6**: Para identificar el caso índice entre los *Pacientes* obtenidos en CE4.
4. **CE3**: Para buscar a los *Pacientes* relacionados con el caso índice y determinar qué elementos compartieron (*Localizaciones*, *UH*, etc.). La combinación de esta información con la obtenida en CE4 permitiría definir con mayor precisión el origen del brote.

Resultados: A continuación, se presentan los detalles de cada consulta, junto con una figura que muestra sus resultados y parámetros.

1. **CE2**: El Experimento 1 comienza con una alerta el 8 de junio de 2023 porque tres pacientes en la cuarta planta, *Planta/1651*, obtuvieron un diagnóstico positivo de *C. diff*. La Figura 3.11 muestra un subgrafo con los nodos de los pacientes resultado de CE2 (*Paciente/6863*, *Paciente/6912*, *Paciente/6922*) y cómo se conectan con *Planta/1651* y el microorganismo *C. diff*. El resultado de esta consulta es una señal de que podría haber un brote en *Planta/1651*.

2. **CE4**: Para descubrir el origen del brote, primero exploramos los pacientes que estuvieron cerca de los tres pacientes obtenidos en CE2 mediante CE4. *C. diff* tiene un tiempo medio de incubación de entre 2 y 3 días [134]. Dado que el contagio no es inmediato y que existe una probabilidad de que los pacientes no se contagien con el primer contacto, hemos añadido un margen adicional hacia atrás para los parámetros de entrada de CE4. Por lo tanto, buscamos contactos espaciales durante cinco días, del 4 al 8 de junio.

   Como muestran los diferentes subgrafos en la Figura 3.12, estos pacientes no compartieron ninguna *Localización* en *Planta/1651* ni tuvieron contacto con el mismo paciente. Las Figuras 3.12.a y 3.12.b muestran los grafos con los pacientes conectados a *Paciente/6912* y *Paciente/6922*, respectivamente. Las Figuras 3.12.c y 3.12.d muestran que *Paciente/6863* tuvo *Eventos* en tres Localizaciones diferentes: *Cama/1583* en la cuarta planta, *Cama/981* en la cuarta planta y *Cama/30* en la planta baja. En las dos últimas Localizaciones, este paciente estuvo cerca de *Paciente/6865*. Por lo tanto, es posible inferir que el origen del brote no está en *Planta/1651* y que *Paciente/6863*, quien tuvo más contactos en más *Localizaciones*, estuvo involucrado en la propagación del brote en *Planta/1651*.

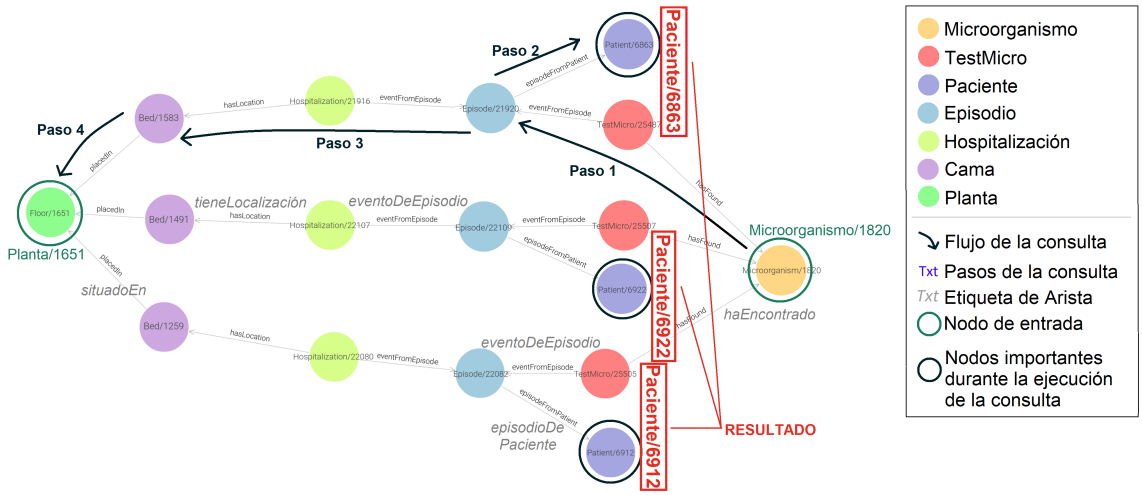


FIGURA 3.11. **Experimento 1, paso 1**. Resultado de la ejecución de CE2 con los siguientes parámetros: *08/06/23* como fecha de inicio y fin, y *1651* como ID de la *Localización* donde se obtuvieron los TestMicro positivos.

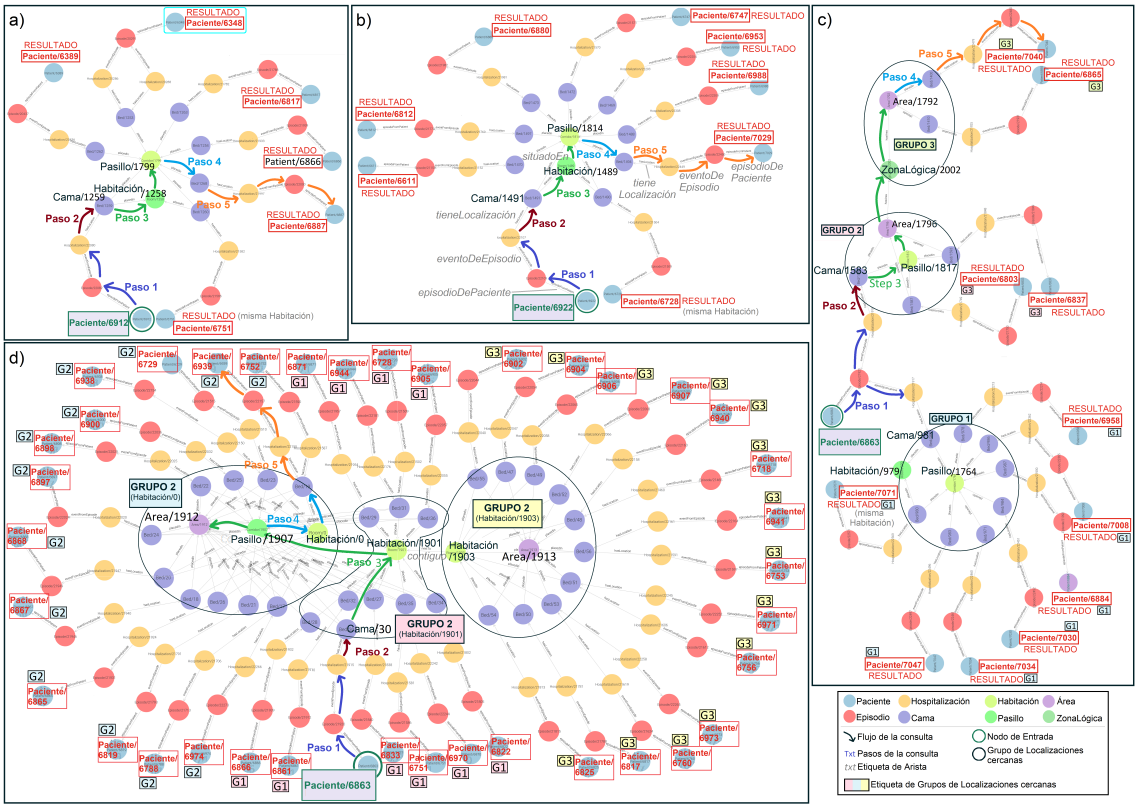


FIGURA 3.12. **Experimento 1, paso 2**. Resultado de la ejecución de EC4 con los siguientes parámetros: *04/06/23* como fecha de inicio, *08/06/23* como fecha de fin y *[6863, 6912, 6922]* como ID de los pacientes. **(a)** Pacientes relacionados con *Paciente/6912*. **(b)** Pacientes relacionados con *Paciente/6922*. **(c)** Pacientes relacionados con *Paciente/6863* en la cuarta planta. **(d)** Pacientes relacionados con *Paciente/6863* en la planta baja.

3. **CE6**: Para verificar si esta hipótesis era correcta, buscamos el caso índice entre todos los pacientes obtenidos en CE4. Ejecutamos CE6 desde el 25 de mayo hasta el 8 de junio (dos semanas). El resultado, que se muestra en Figura 3.13, revela que el *Paciente/6348* es el primer paciente con un test positivo el 26 de mayo. Por lo tanto, el brote se estuvo propagando sin ser detectado durante mucho tiempo. Como se muestra en la Figura 3.12.a , este paciente (marcado en azul) tuvo contacto con el *Paciente/6912* a través del *Pasillo/1799* en la *Planta/1651*, lo que permite descartar nuestra hipótesis inicial.

4. **CE3**: Utilizamos CE3 para buscar a los pacientes con un diagnóstico positivo de *C. diff* que estuvieron en contacto con el *Paciente/6348* después del 26 de mayo. La Figura 3.14 muestra el resultado de CE3.

   Cabe destacar que el *Paciente/6348* permaneció en la *Cama/1253* en la *Planta/1651* y fue atendido por la *UH/1627* desde el 25 de mayo, con solo dos excepciones: dos eventos de *Radiografía* el 30 de mayo y el 5 de junio en el *Pasillo/1919* en la planta baja.

El *Paciente/6348* tuvo contacto con diez pacientes a través de la *UH/1627*. Seis de ellos fueron hospitalizados en habitaciones del *Pasillo/1799*. Tres pacientes tuvieron contacto con el *Paciente/6348* a través de eventos de *Radiografía*, uno de ellos (*Paciente/6337*) en la misma cama (*Cama/14*) el mismo día. Además, el *Paciente/6339* tuvo un contacto el 30 de mayo a través del área de radiología y tuvo otros dos mediante la *UH/1627* entre el 25 de mayo y el 6 de junio.

En conclusión, sospechamos que hay una fuente de *C. diff* en el *Pasillo/1799* y que se ha propagado a través de la *UH/1627* a otros pacientes.

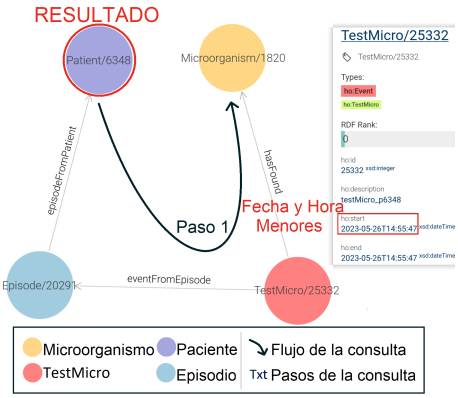


FIGURA 3.13. **Experimento 1, paso 3**. Resultado de la ejecución de CE6 con los siguientes parámetros: *25/05/23* como fecha de inicio, *08/06/23* como fecha de fin, *1820* como ID del *Microorganismo* y los pacientes obtenidos en CE4.

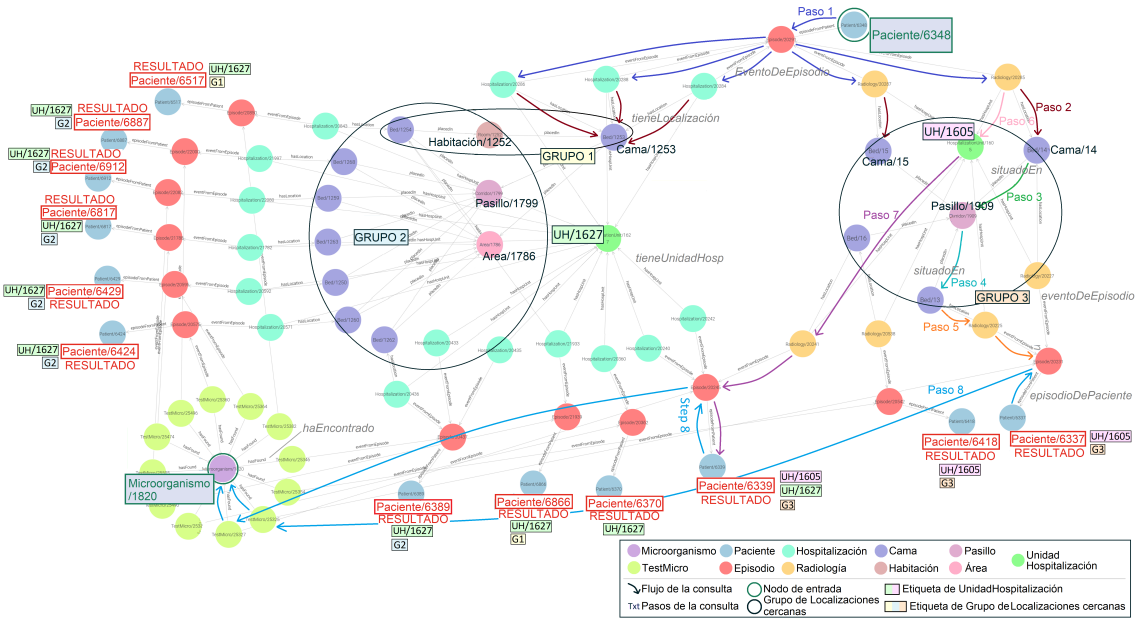


FIGURA 3.14. **Experimento 1, paso 4**. Resultado de la ejecución de CE3 con los siguientes parámetros: *26/05/23* como fecha de inicio, *08/06/23* como fecha de fin, *6348* como ID del *Paciente* y *1820* como ID del *Microorganismo*.

3.4.2.b. Experimento 2: Reconstrucción de un brote

Objetivo: Este experimento tiene como objetivo principal reconstruir cómo se propagó un brote y descubrir su origen utilizando el análisis de contactos entre pacientes infectados.

Consultas epidemiológicas: Hemos utilizado **2 CE** en el siguiente orden:

1. **CE1**: para detectar un posible brote en un *Servicio*.
2. **CE5**: para buscar qué compartieron los pacientes infectados de CE1. Esta información nos puede permitir determinar posibles focos de contagio.

Resultados: A continuación, se presentan los detalles de cada consulta, junto con una figura que muestra sus resultados y parámetros.

1. **CE1**: Primero ejecutamos Q1 para determinar cuántos pacientes atendidos por *Servicio/8* tuvieron un diagnóstico positivo de *C. diff* entre el 8 y el 14 de enero (ocho días).

   Como se muestra en la Figura 3.15, hubo once pacientes infectados. *UH/1629* atendió a 6 de ellos, mientras que *UH/1627* y *UH/1628* atendieron a 3 y 2 pacientes, respectivamente.

   Al verificar las fechas de las pruebas, podemos observar que los primeros casos positivos fueron el 8 de enero, correspondientes a *Paciente/590* y *Paciente/351*, quienes fueron atendidos por *UH/1629* y *UH/1627*, respectivamente.

   En este punto, sospechamos que debió haber contacto entre estos dos pacientes y que la *Localización* compartida estaba relacionada con el origen del brote.

2. **CE5**: Hemos utilizado CE5 para determinar qué *Localizaciones* y *Servicios* fueron compartidos por los once pacientes obtenidos en CE1. La Figura 3.16 muestra el resultado de CE5. Los pacientes que compartieron la misma *UH* están rodeados por el mismo color.

   Hemos identificado dos hechos significativos estrechamente relacionados con la propagación del brote:

   - La mayoría de los pacientes fueron atendidos por varias *UH* del *Servicio/8*.
   - Seis pacientes estuvieron hospitalizados en habitaciones de *UH/1603*. Esta *UH* pertenece al *Servicio* de UCI. Además, las habitaciones obtenidas en el resultado son contiguas o están en el mismo *Pasillo*.

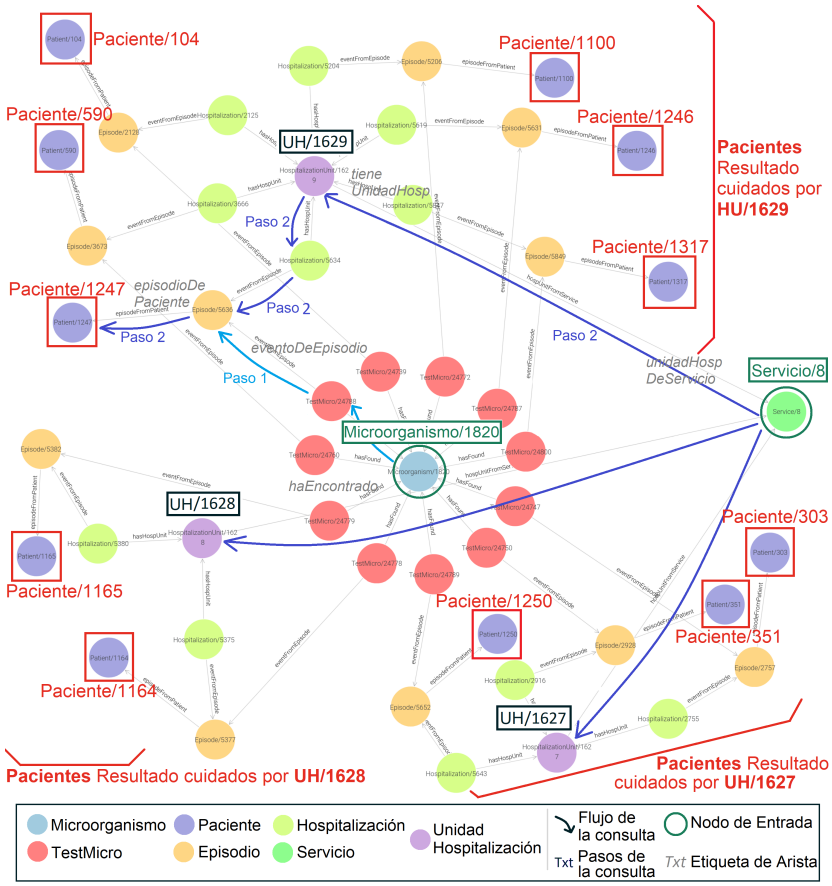


FIGURA 3.15. **Experimento 2, paso 1**. Resultado de la ejecución de EC1 con los siguientes parámetros: *07/01/23* como fecha de inicio, *14/01/23* como fecha de fin, *8* como ID del *Servicio* y *1820* como ID del *Microorganismo*.

La Tabla 3.7 muestra los eventos en orden según su fecha de inicio. La última columna de la tabla indica si el paciente tuvo un *TestMicro* positivo durante el *Evento*. En la columna '*Localización Descripción*' marcamos del mismo color aquellas *Localizaciones* cercanas entre sí. Cuando un *Evento* ocurre en una *Localización* que no es compartida ni cercana a otro *Evento* de la tabla, aparece un '-'. También hemos resaltado cada *UH* de un color. Los colores utilizados para las *UH* y las *Localizaciones* son los mismos que los grupos resaltados en la Figura 3.16. Esta tabla permite correlacionar ciertos hechos:

- *Paciente/590* y *Paciente/351* estuvieron en la misma habitación de UCI, *Habitación/1906*, del 9 al 12 de enero. Esto permite sospechar que esta habitación fue el origen del brote. Sin embargo, los diagnósticos positivos de *C. diff* de estos pacientes fueron el 8 de enero, cuando el *Paciente/590* estaba en la *UH/1629* y el *Paciente/351* en la *UH/1627*.
- Las hospitalizaciones en *UH/1603* y *Habitación/1906* de *Paciente/590* y *Paciente/351* fueron un foco de contagio del brote. *Paciente/1246* dio positivo el 14 de enero, tras haber estado en la *Habitación/1906* desde

el 11 de enero. *Paciente/303* y *Paciente/1164* dieron positivo el 11 y 12 de enero, después de haber estado en la *Habitación/1904*, vecina de la *Habitación/1906*. El *Paciente/1165* dio positivo el 12 de enero, tras haber estado en una habitación del mismo pasillo que la *Habitación/1906* desde el 9 de enero. Las fechas de todos estos *TestMicro* coinciden con el tiempo medio de incubación de *C. diff* tras el contacto con *Paciente/590* y *Paciente/351*.

- *Paciente/1100* estuvo en la misma habitación que *Paciente/351* hasta el 9 de enero. Al día siguiente, *Paciente/1100* tuvo un diagnóstico positivo de *C. diff* y fue trasladado a la *Habitación/1453* (*Área/1793*), siendo atendido por *UH/1629*. En la *Habitación/1453*, *Paciente/1100* estuvo con *Paciente/1247*, quien dio positivo el 13 de enero. Ese mismo día, el *Paciente/1317* también dio positivo, tras haber estado desde el 10 de enero en una cama de la misma *Área* que la *Habitación/1453* y ser atendido por *UH/1629*.

Nuestro modelo y consultas nos permiten concluir que hubo dos focos de contagio: *UH/1603* y *Área/1793*.

Tabla 3.7
Pacientes obtenidos en CE3 con sus eventos.

Paciente ID	**Fecha de Inicio - Fecha de Fin**	**UH ID**	**Descripción de *Localización***	**Fecha de *TestMicro***
104	01/01 - 07/01	1629	-	-
1100	04/01 - 09/01	1627	Habitación/1270	-
351	06/01 - 09/01	1627	Habitación/1270	08/01
590	06/01 - 09/01	1629	Habitación/1390	08/01
1164	06/01 - 09/01	1629	Cama/1458	-
303	06/01 - 10/01	1603	Habitación/1904	-
1250	08/01 - 11/01	1629	-	-
104	08/01 - 11/01	1629	-	10/01
1247	08/01 - 12/01	1629	Habitación/1453	-
1100	09/01 - 10/01	1629	Habitación/1453	10/01
351	09/01 - 12/01	1603	Habitación/1906	-
1164	09/01 - 12/01	1603	Habitación/1904	-
1165	09/01 - 12/01	1603	Pasillo/1910	-
590	09/01 - 13/01	1603	Habitación/1906	-
1317	10/01 - 13/01	1629	Cama/1430	13/01
303	10/01 - 13/01	1627	Área/1788	11/01
1246	11/01 - 14/01	1603	Habitación/1906	-
1250	11/01 - 15/01	1627	Habitación/1390	12/01
1164	12/01 - 14/01	1628	-	12/01
1165	12/01 - 14/01	1628	Área/1788	12/01
1247	13/01 - 13/01	1629	Habitación/1453	13/01
1246	14/01 - 18/01	1629	-	14/01

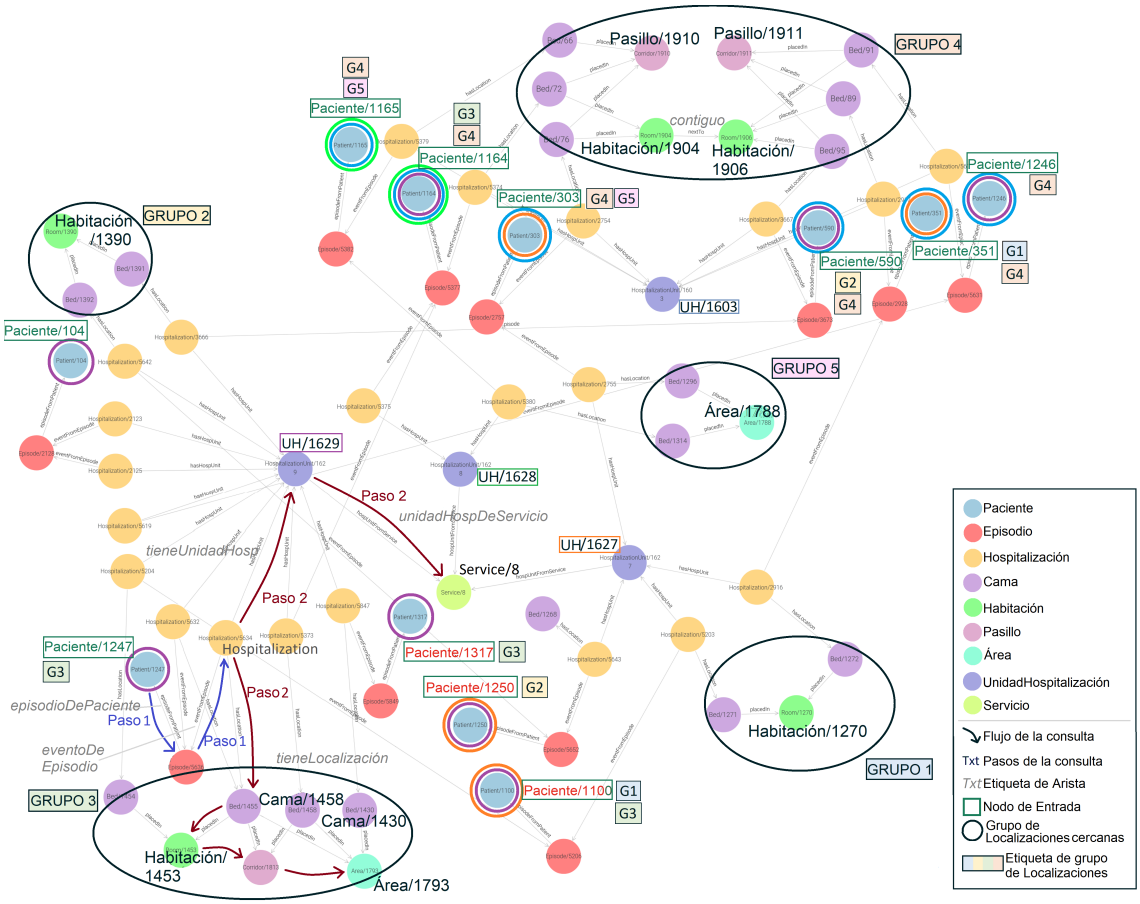


FIGURA 3.16. **Experimento 2, paso 2**. Resultado de la ejecución de CE5 con los siguientes parámetros: *07/01/23* como fecha de inicio, *14/01/23* como fecha de fin y los pacientes obtenidos en CE1. Los pacientes atendidos por *UH/1627* están rodeados en naranja, mientras que los pacientes cuidados por *UH/1628* están en verde, los pacientes cuidados por *UH/1629* están en lila y los pacientes cuidados por *UH/1603* están en azul. No hemos incluido los TestMicro en la figura por una mayor claridad.

3.5 Discusión

3.5.1. Cuantificación de la proximidad

Con las consultas epidemiológicas propuestas y la validación de este capítulo nos hemos centrado en analizar la propagación de infecciones en base a los contactos entre pacientes. Nuestra definición de contacto no se basa en una medida cuantificada, sino en la definición de un conjunto de caminos que recorren la dimensión espacial del modelo epidemiológico. Con los experimentos de la Sección 3.4 hemos demostrados que el análisis de contactos basado en este enfoque puede contribuir a descubrir cómo se propaga una enfermedad.

En este capítulo no hemos definido una forma de cuantificar la proximidad entre pacientes, aunque puede lograrse utilizando la propiedad *coste* en algunas relaciones de dimensión espacial. Recuérdese que el objetivo de esta propiedad es la de proporcionar una distancia semántica entre dos clases, la cual sería cualitativa y dependería de las clases que relaciona: cuanto más cercanas estén dos *Localizaciones*, mayor será la probabilidad de contagio y, por tanto, menor debería ser su distancia. Por ejemplo, cualquier relación *situadoEn* entre una *Cama* y una *Habitación* tendría el mismo valor y sería menor que cualquier relación *situadoEn* entre una *Habitación* y un *Pasillo*. Así, la distancia semántica entre dos *Localizaciones* podría definirse como el coste del camino más corto entre ellas al recorrer las aristas de la dimensión espacial.

3.5.2. Flexibilidad en la representación del espacio y el tiempo

Hemos diseñado nuestro modelo espacio-temporal epidemiológico con el objetivo de que fuera general y adaptable a cualquier distribución hospitalaria y fuente de datos. Pero también queremos que sea lo suficientemente detallado para describir, por ejemplo, cualquier aspecto espacial que influya en la propagación de un brote.

3.5.2.a. Dimensión temporal

En cuanto a la **dimensión temporal**, cabe destacar que conocer **la localización exacta de los pacientes y sanitarios dentro de un edificio es una tarea compleja**. Para conseguir los resultados más precisos posible, deberían llevar algún tipo de dispositivo que registrase su posición (por ejemplo, sus coordenadas relativas dentro hospital). Sin embargo, esta solución sería difícil de aplicar pues se necesitaría un sistema capaz de registrar cada pocos segundos o minutos grandes cantidades de datos, así como administrarlos y permitir su consulta eficientemente. Además, los dispositivos pueden fallar (dejar de funcionar o registrar datos inexactos), por lo que el sistema debería ser capaz de gestionar estos errores. Por otro lado, se requeriría del permiso expreso y legal de todos los pacientes y trabajadores, el cual, por temas de privacidad, no se podría garantizar.

Como solución, proponemos utilizar los SIH y HCE como principal fuente de datos para el modelo, pues de ellos se puede extraer información como el lugar y la fecha en que un paciente fue hospitalizado o se sometió a ciertas acciones, así como el clínico o la unidad que lo atendió.

También queremos señalar que, en nuestro modelo, la clase ***Episodio*** **representa un "episodio de atención"**, es decir, los servicios de atención médica prestados durante una estancia hospitalaria en respuesta a una solicitud específica [83]. Este abarca desde la admisión hasta el alta, traslado a otro hospital o fallecimiento del paciente. Un *Episodio* se diferencia de una *Hospitalización* en que esta última se refiere al momento en que un paciente es ubicado en una *Cama* para recibir atención médica. Por tanto, un *Episodio* puede contener varias *Hospitalizaciones*. Por ejemplo, un *Episodio* puede comenzar con una *Hospitalización* en Emergencias y seguir con una *Hospitalización* en planta.

Hemos decidido no definir los *Episodios* como *Eventos* para facilitar la descripción de los caminos a recorrer en las consultas. Además, son útiles para filtrar temporalmente los *Eventos*: solo los *Eventos* de aquellos *Episodios* que coincidan con el intervalo de búsqueda se consideran adecuados para la consulta.

3.5.2.b. Dimensión espacial

Un aspecto relevante de la **dimensión espacial** es su **flexibilidad para agrupar semánticamente las *Localizaciones*** de los niveles inferiores de la jerarquía: mediante las clases *Área* y *ZL*, así como las relaciones *contiguo* y *opuesto*.

Las *Áreas* dividen los *Plantas* en cuadrículas. Están pensadas principalmente para plantas de grandes dimensiones, donde sería necesario añadir un nivel intermedio entre *Pasillo* y *Planta* para representar que algunas *Localizaciones* están más cerca que otras. Una opción para definir las *Áreas* podría ser en base a los puestos de enfermería, que suelen abarcar varios pasillos en una misma planta. También se podría dividir el espacio según los *Servicios* o *UH* que operan en la planta, permitiendo una conexión más fuerte entre las partes física y lógica de la dimensión espacial.

La Figura 3.17 muestra una representación esquemática de una planta que ha sido dividida en cuatro *Áreas*. Estas nos permiten describir que, por ejemplo, los pacientes en el *Pasillo p0* están más cerca de los pacientes en *p1* que de aquellos en *p6*. Cuando un *Pasillo* atraviesa varias *Áreas*, este es dividido en segmentos, como ocurre con el pasillo principal en la Figura 3.17.

En caso de que el hospital presente una planta de dimensiones reducidas, puede no ser necesario dividirla en una cuadrícula, sino que haya un *Área* que abarca toda la planta. En nuestros ejemplos, el hospital está configurado como una cuadrícula. Sin embargo, el modelo no está limitado a una estructura basada en rectángulos, sino que la principal limitación es que sea posible identificar las *Áreas* mediante un sistema de coordenadas bidimensional.

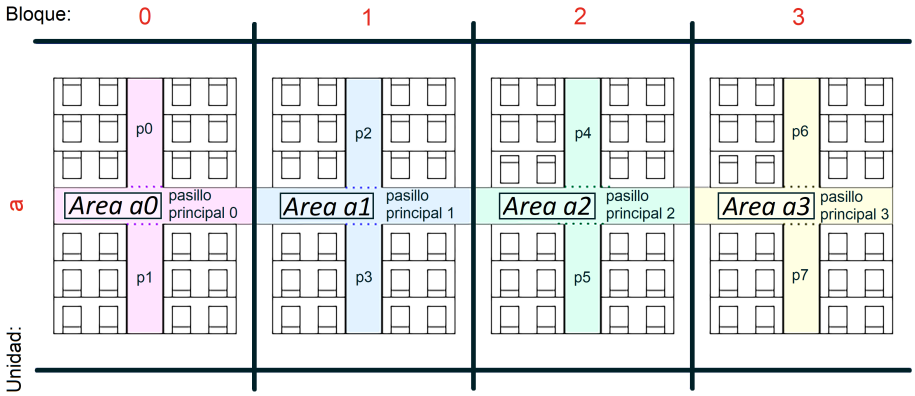


FIGURA 3.17. Representación de una *Planta* dividida en cuatro *Áreas*. Cada *Área* cuenta con dos *Pasillos* con *Habitaciones* y una sección del pasillo principal.

Hemos aprovechado **las relaciones *contiguo* y *opuesto* para dotar de más precisión a la ubicación de las *Localizaciones*** de los niveles inferiores de la jerarquía espacial, así como permitir relaciones de cercanía que no serían posibles utilizando únicamente las relaciones *tieneLocalización* con las que definimos los niveles de la jerarquía espacial. Por ejemplo, es posible representar que dos *Habitaciones* contiguas de diferentes *Pasillos* están más cerca entre sí que dos *Habitaciones* del mismo *Pasillo* que no son adyacentes.

La transmisión de infecciones entre pacientes en *Localizaciones* conectadas mediante estas relaciones es más probable por diversas razones. Por ejemplo, los sanitarios atienden a los pacientes siguiendo el orden de habitaciones, es decir, la siguiente *Habitación* a visitar será la *opuesta* o *contigua* a la *Habitación* de su paciente anterior.

El concepto *ZL* se ha incluido para representar que pueden existir conexiones de proximidad entre algunas *Áreas* que no estén relacionadas con la distribución espacial del hospital: puede haber *Áreas* no contiguas que hayan estado estrechamente relacionadas con la transmisión de antiguos brotes. Por ejemplo, hay estudios que sugieren que un brote puede propagarse a través de tuberías o del sistema de ventilación, afectando zonas amplias de una misma planta o zonas superpuestas en diferentes plantas [32, 69]. Mediante las *ZL* es posible modelar que la transmisión de infecciones entre pacientes hospitalizados en las *Áreas* que abarca es más probable.

Con respecto a la **parte lógica de la dimensión espacial**, conocer la ubicación exacta de los sanitarios en todo momento puede ser un desafío. Además de dispositivos de ubicación en tiempo real, los sanitarios podrían llevar un registro manual de las habitaciones visitadas a lo largo de su jornada laboral. Sin embargo, esta solución también sería difícil de implementar en un escenario real. Aun así, nos interesa modelar posibles rutas de transmisión en las que los servicios puedan tener cierta influencia.

Además del concepto *Servicio*, hemos añadido el de *UnidadHospitalización* para representar situaciones más concretas de proximidad entre pacientes como, por ejemplo, que un grupo de sanitarios solo atiende ciertas *Camas* o *Habitaciones*, o describir un cirujano con su equipo de cirugía.

Nuestro modelo ha sido diseñado para que pueda adaptarse a hospitales de diverso tamaño y organización espacial. Además, las consultas epidemiológicas han sido diseñadas e implementadas en SPARQL* en base a dos objetivos:

- Permitir su funcionamiento incluso en bases de datos que no hayan implementado al completo el modelo de datos propuesto. Este es el caso del conjunto de datos MIMIC-III, que solo incluye *Habitaciones* y *Servicios*.
- Facilitar su extensión, por ejemplo, añadiendo nuevos caminos o filtros a los ya definidos.

3.5.3. Comparativa con modelos similares

En los estudios que buscan analizar la transmisión espacio-temporal de infecciones nosocomiales lo más común es desarrollar un modelo ad-hoc para un problema específico. Estos trabajos suelen centrarse en determinadas áreas del hospital y utilizando los SIH y HCE para determinar la ubicación de los pacientes. Posteriormente, agrupan a los pacientes en clústeres para representar entre cuáles es más probable la transmisión de una infección. Aunque no es lo más común, en algunos trabajos se ha desarrollado un modelo detallado para la representación espacial. A continuación, vamos a mostrar y comparar con nuestro modelo algunos de los más destacados:

- En [159], los autores complementaron los datos de HCE con un grafo para modelar la distribución de la planta del hospital. En este grafo, los nodos representan habitaciones y pasillos, y las aristas representan la distancia (medida en metros) entre habitaciones. Utilizan esta estructura para medir la distancia entre pacientes y sanitarios a lo largo del tiempo, aplicando un algoritmo de camino más corto sobre el grafo del hospital.
- En [170], se ha utilizado un Sistema de Información Geográfica (SIG) para implementar la distribución del hospital. En él, los movimientos de los pacientes durante su estancia son georreferenciados con las coordenadas de la habitación en la que estuvieron $(X, Y, planta)$. Para determinar los contactos entre pacientes, se utiliza una matriz de distancia entre habitaciones. Además, cada habitación guarda información referente a la fecha en la que se han realizado pruebas microbiológicas con resultado positivo para el microorganismo del estudio (*Klebsiella pneumoniae*), así como el fenotipo encontrado.
- En [48] se ha utilizado un grafo para modelar las áreas de un hospital. Este grafo se asemeja a las redes de metapoblaciones estudiadas en la Sección 2.3.4: los nodos

representan las áreas y tienen propiedades como el número de pruebas positivas de SARS-CoV-2 registradas en el área y el número acumulado de días que los pacientes han permanecido en ella; y sus aristas, que son dirigidas, representan los movimientos de pacientes entre áreas en cada sentido, teniendo como propiedad la cantidad de movimientos. Por tanto, una diferencia principal de su modelo con el nuestro es la granularidad del enfoque: pacientes individuales frente a pacientes como conjunto.

Estos estudios no trabajan con modelos con un elevado detalle en la descripción de la dimensión espacial ni temporal, lo que los hace difíciles de adaptar a otros estudios en los que se quiera analizar otros aspectos de la transmisión. En nuestro caso, hemos diseñado un modelo que pueda ser reutilizado en otros estudios, pudiendo necesitar pequeños cambios que no alteren su significado general, como la adición de nuevas propiedades, relaciones o clases. Por ejemplo, podríamos adaptar [159] modificando el valor de la propiedad *coste* en las relaciones entre *Habitaciones* y *Pasillos*. En el caso de [170], bastaría con añadir dos propiedades, x e y, a las *Habitaciones* para representar sus coordenadas, ya que la planta y la fecha de las pruebas positivas pueden obtenerse directamente de nuestro modelo. Finalmente, consideramos que nuestro modelo es lo suficientemente expresivo y flexible como para consultar y recuperar la misma información que en [48].

3.6 Conclusiones

El objetivo principal de este capítulo es el de proporcionar un modelo de datos que sirva de base para el estudio espacio-temporal de la transmisión de infecciones en entornos hospitalarios. Este modelo ha sido diseñado considerando que sea flexible y genérico (capacidad de adaptación a diversas distribuciones hospitalarias y distintas fuentes de datos), así como detallado. En este sentido, hemos diseñado el modelo mediante la definición por separado, y posterior unión, de la dimensión espacial y la dimensión temporal. La dimensión espacial representa la estructura arquitectónica del hospital mediante una jerarquía de *Localizaciones*. Además, hemos añadido clases y relaciones adicionales, como *Área*, *ZL*, *contiguo* y *opuesto*, para lograr una descripción más precisa del espacio. También incluímos una representación simplificada del personal sanitario. La dimensión temporal describe todo lo que ocurre a los pacientes durante su estancia en forma de *Eventos*, los cuales se conectan con la dimensión espacial, permitiendo conocer la ubicación de los pacientes en todo momento.

Para validar la utilidad del modelo en el contexto de la investigación epidemiológica hospitalaria, hemos definido seis consultas epidemiológicas que representan diferentes tareas representativas de esta investigación: detección de brotes, análisis de propagación de infecciones e identificación del paciente índice. Estas consultas toman como base los movimientos de los pacientes dentro del hospital, con el objetivo de detectar posibles rutas de transmisión en base a sus contactos. Para ello, hemos propuesto una definición

de contacto en la que se trata de minimizar los nodos que no conducen a una posible ruta de transmisión.

Además, hemos analizado las principales tecnologías de bases de datos orientadas a grafos (BDOG) desde un punto de vista técnico: espacio de almacenamiento requerido y tiempo de ejecución de las consultas y máxima memoria principal consumida para la ejecución. Como resultado de presentar una mayor estabilidad de rendimiento en estos tres aspectos y sencillez en la implementación del modelo y consultas, hemos decidido implementar el modelo de datos como un grafo de conocimiento en formato RDF* y las consultas en su lenguaje de consulta SPARQL*. Hemos utilizado GraphDB como BDOG.

Finalmente, hemos realizado dos experimentos en los que se simulan dos brotes de *Clostridioides difficile*. Los resultados muestran que el modelo permitiría a los epidemiólogos analizar brotes de forma eficaz. En ambos experimentos, los análisis de los brotes representan de forma fidedigna la información hospitalaria. Trabajar con varios niveles de detalle nos permite resaltar ciertos pacientes o eventos sin excluir otras posibles rutas de transmisión.

Destacamos que, en nuestra investigación de la detección y análisis de brotes nosocomiales, sería interesante planear como próximo paso a realizar el estudio de cuantificación de la proximidad semántica entre *Localizaciones*, así como el desarrollo de diversas técnicas de razonamiento espacial y temporal basadas en esta que permitan la automatización de la investigación epidemiológica.

3.7 *Open-science*

Para el desarrollo de este capítulo hemos implementado diferentes herramientas de software y conjuntos de datos, cuyo acceso es público a través de repositorios en GitHub.

- Implementación de las seis consultas epidemiológicas en SPARQL* y conjunto de datos utilizado en los experimentos de la Sección 3.4. El conjunto de datos se presenta como archivos en formato *N-Triples* (`.nt`) y *Turtle* (`.ttl`) para la creación de un grafo de conocimiento en formato RDF*. `https://github.com/LorenaPujante/HospitalKG`
- Software para adaptar la salida del modelo de simulación hospitalario [98] a un grafo de conocimiento en formato RDF y RDF* que siga nuestro modelo de datos: `https://github.com/LorenaPujante/HospitalGeneratorRDF`
- Software para adaptar el conjunto de datos de libre acceso MIMIC-III a nuestro modelo de datos: `https://github.com/LorenaPujante/MimicToRDF`
- Software con el código utilizado en las pruebas comparativas entre las BDOG: `https://github.com/LorenaPujante/MimicIII_BDOG_Benchmark`

Capítulo 4

Similitud epidemiológica espacio-temporal basada en trayectorias de pacientes

En este capítulo, presentamos el método StESPT con el fin de servir de soporte a los profesionales clínicos en la investigación de las relaciones epidemiológicas espacio-temporales entre pacientes infectados, facilitando así la detección de brotes y la identificación de posibles rutas de transmisión. Para ello, StESPT recupera los movimientos dentro del hospital de los pacientes infectados y los transforma en trayectorias. A continuación, empleamos un algoritmo para medir distancias entre trayectorias (TDMA) con el que cuantificar la similitud espacio-temporal entre las trayectorias de cada par de pacientes. Finalmente, aplicamos el algoritmo de *clustering* k-means sobre las similitudes entre cada par de pacientes para distribuirlos en grupos de pacientes "cercanos" que puedan ayudar a comprender la propagación de la infección. Comparamos la idoneidad de tres TDMA de uso común en la literatura científica (*Dynamic Time Warping*, *Spatiotemporal Linear Combine similarity* y *Spatiotemporal Longest Common Subsequence*), que son adaptados para evaluar la similitud entre trayectorias, en lugar de distancia. Para cada TDMA, también proponemos una versión que se adapta mejor a la semántica de nuestro problema.

4.1 Introducción

Como se indicó con anterioridad, dentro de un hospital, las infecciones pueden transmitirse tanto por contacto directo como indirecto. El **contacto directo** se define como el contacto físico o el contacto dentro de la distancia umbral para la transmisión por gotas o fómites, por ejemplo, a través de la actividad respiratoria. El **contacto indirecto** puede ocurrir a través de una cadena de personas (por ejemplo, el personal sanitario) o un entorno contaminado (las gotas y los fómites pueden permanecer suspendidos o en superficies como ropa, muebles o instrumentos médicos) [232].

El análisis de contactos desempeña un papel fundamental en el control de infecciones. Sin embargo, incluso en los hospitales, este análisis es realizado de forma

manual, lo cual puede retrasar la detección de la infección. El análisis de contactos también puede ayudar a descubrir las rutas de transmisión de la infección, pero para ello solamente se basa en los contactos directos. Por tanto, los contactos indirectos pueden no ser detectados [37].

En general, la densidad de población y los patrones de movilidad tienen una influencia significativa en la propagación de un brote [241]. En el caso de los entornos hospitalarios, la estructura espacial del hospital, la distribución física de los pacientes y sus movimientos por las diversas estancias y servicios dentro del hospital (sus trayectorias) son factores determinantes en la transmisión de las infecciones nosocomiales.

Cuando se registra un número elevado de infecciones en un hospital durante un determinado período, los epidemiólogos deben determinar si se trata de una situación que puede ser caracterizada como un brote. Para ello, deben realizarse dos tareas principales.

- En primer lugar, deben **investigar las relaciones epidemiológicas entre los pacientes infectados**. Es usual que se emita una alerta por posible brote cuando el número acumulado de casos de una bacteria (u otro patógeno) alcanza un umbral en una ubicación espacial específica (servicio, unidad de hospitalización o planta) durante un período determinado (generalmente una semana) [57, 217]. El brote se confirma cuando se demuestra que existe alguna relación epidemiológica espacio-temporal entre los pacientes infectados.
- En segundo lugar, deben **detectar las posibles rutas de transmisión** con el fin de controlar la infección. Para ello, podrían identificar grupos de pacientes con trayectorias similares.

En Capítulo 3, propusimos un modelo de datos que nos permite analizar todos los movimientos de los pacientes durante su hospitalización. También propusimos un conjunto de consultas epidemiológicas para ayudar a los epidemiólogos en tareas relacionadas con la detección de brotes nosocomiales y el análisis de contactos entre pacientes. Además, mostramos dos experimentos donde ejemplificamos cómo con el uso combinado de varias consultas epidemiológicas es posible inferir cuáles podrían ser los principales focos de transmisión de la infección. Sin embargo, se trata de un proceso manual que puede requerir de un gran conocimiento sobre el funcionamiento del hospital y de esfuerzo para detectar en el grafo las relaciones entre los pacientes.

Además, en este capítulo quedaba como tarea pendiente la definición de una distancia entre las clases que forman la dimensión espacial del modelo, por ejemplo, a través de la propiedad *coste* entre sus relaciones.

En este capítulo, presentamos el método ***Spatio-temporal Epidemiological Similarity for Patient Trajectories*** (**StESPT**) (en español, *Similitud Epidemiológica Espacio-temporal basada en las Trayectorias de Pacientes*) para que sirva de soporte a los epidemiólogos en las dos tareas descritas a continuación:

- Para la primera tarea, proponemos **cuantificar el grado de conexión espacio-temporal entre los pacientes infectados a partir de sus trayectorias**. La hipótesis base de StESPT es que los pacientes con un contacto más cercano o prolongado (ya sea directo o indirecto) con un paciente infectado deberían, en teoría, tener un mayor riesgo de infección [241]. Esta cercanía entre dos pacientes puede interpretarse como el hecho de que compartan trayectorias similares, lo cual podemos calcular utilizando un *algoritmo para medir distancias entre trayectorias* (del inglés, *Trajectory Distance Measure Algorithm* (**TDMA**)). Basándonos en nuestro modelo, podemos transformar los *Eventos* registrados para cada *Paciente* en trayectorias a comparar.

- Para la segunda tarea, **la similitud entre las trayectorias** calculada en la primera tarea **puede utilizarse de entrada en un algoritmo de *clustering*, con el fin de identificar los pacientes más cercanos** y, por tanto, las rutas de transmisión más plausibles.

- Además, tanto las trayectorias de los pacientes como la similitud entre ellas pueden representarse de forma gráfica, sirviendo de apoyo al epidemiólogo para la interpretación de los resultados obtenidos de ambas tareas.

El método StESPT toma como base el modelo epidemiológico espacio-temporal y las consultas epidemiológicas definidas en el Capítulo 3. Además, tal y como propusimos en dicho capítulo, los datos han sido implementados como un grafo de conocimiento en formato RDF y almacenados en el motor RDF *GraphDB*.

Dentro del método StESPT, comparamos tres TDMA de uso común en la literatura científica y proponemos una extensión para cada uno. Nuestro análisis trata de evaluar cómo cada una de las seis opciones propuestas refleja las relaciones espaciales y temporales entre los pacientes.

Hemos probado el método StESPT mediante un experimento donde basándonos en unos datos sintéticos con los que se simula una infección de *C. diff* en un hospital, aplicamos cada uno de los pasos que componen el método StESPT para analizar las relaciones espacio-temporales entre los pacientes infectados durante un período de tiempo corto. En el experimento, demostramos la relación epidemiológica entre los casos identificados y la existencia del brote e identificar los grupos de pacientes entre los cuales la transmisión era más probable.

El capítulo queda organizado de la siguiente manera: la Sección 4.2 proporciona información relacionada con el estado del arte (Sección 4.2.1), las características deseables para los TDMA a comparar (Sección 4.2.2) y algunas definiciones preliminares sobre conceptos que utilizamos a lo largo del capítulo (Sección 4.2.3). En la Sección 4.3 se describen en detalle los pasos que conforman el método StESPT. Son de especial interés las Secciones 4.3.3 (definición de la ecuación para calcular la similitud epidemiológica espacio-temporal entre puntos) y 4.3.4 (definición de los TDMA a analizar). La Sección 4.4 presenta una evaluación experimental del método

StESPT, cuyos resultados se discuten en la Sección 4.5. Las conclusiones se resumen en la Sección 4.6.

4.2 Contexto

4.2.1. Estado del arte

Los avances tecnológicos relacionados con los sensores para registrar datos de distinta naturaleza, así como con los dispositivos GPS (*Global Positioning System*; en español, Sistema de Posicionamiento Global) y otras tecnologías para rastrear la posición han llevado a la generación de grandes volúmenes de datos espacio-temporales que permiten describir el movimiento de *objetos móviles*, que pueden ser tanto seres inanimados como animados (personas, animales). El conjunto de datos con los que se describe el movimiento de un *objeto móvil* recibe el nombre de **trayectorias**. El interés en el análisis de trayectorias ha aumentado en los últimos años en diversos sectores, como el transporte, la predicción del tráfico, el estudio del comportamiento, la migración animal o la identificación de puntos de interés [198]. El análisis de trayectorias abarca varias tareas, como la creación de índices, clasificación, agrupamiento, consulta y minería de datos. En todas ellas, el cálculo de la distancia o similitud entre trayectorias es un componente fundamental, por lo que debe definirse con la precisión necesaria para que capture adecuadamente la semántica de los datos y sus relaciones subyacentes [215]. En consecuencia, los **algoritmos para medir distancias entre trayectorias** (del inglés, *Trajectory Distance Measure Algorithm* (**TDMA**)) han sido ampliamente estudiados durante décadas.

En los últimos años, con la aparición del COVID-19, también ha crecido el interés por el análisis de trayectorias en epidemiología. Predominan los estudios donde se utilizan datos de posición obtenidos mediante GPS [93, 241], Bluetooth y operadores de telefonía móvil [44]. Estos datos se utilizan para diversas tareas, como el análisis de contactos y la detección de patrones en la movilidad de las personas en ciudades o regiones. En entornos hospitalarios, para recrear los movimientos de los pacientes se utilizan tanto datos reales recogidos de los Sistemas de Información Hospitalaria (SIH) o de la Historia Clínica Electrónica (HCE) [55, 149, 150, 170, 212], como datos sintéticos generados con simuladores [159]. Algunos de estos estudios se caracterizan por la creación de redes de contacto entre pacientes, donde se considera que hay un contacto si dos pacientes están en la misma habitación [48, 159], pabellón o unidad hospitalaria (entendida como un conjunto de habitaciones de un mismo servicio -nótese la similitud del concepto con el definido en la Sección 3.2.1.a-) [148, 149, 170, 212] en algún momento del día. Es decir, en estos estudios se establece una granularidad de *un día*, no incluyendo la hora del contacto. Sobre estas redes se han aplicado diferentes técnicas estadísticas y de análisis de redes para detectar agrupaciones de pacientes. Consideramos que un período de 24 horas puede ser demasiado largo para un análisis detallado de los movimientos de un paciente, ya que durante este tiempo el

paciente puede haber estado en varias habitaciones. Además, como unidad espacial utilizan un conjunto de habitaciones sin definir niveles adicionales de subdivisión del espacio para el análisis. Para un análisis más preciso, esta unidad puede ser demasiado amplia y altamente compartimentada, lo que podría implicar que todos los pacientes en esa unidad sean considerados igualmente cercanos, sin distinguir diferencias significativas en las distancias espaciales entre cada ubicación. En caso de que se utilizasen habitaciones individuales como unidad espacial, incluir información sobre la disposición del hospital o la distancia física entre habitaciones podría permitir un análisis más completo.

Algunos estudios han analizado las trayectorias de los pacientes con un mayor nivel de detalle [52, 150]. En [150], se utilizó una ecuación de *proximidad espacio-temporal* para cuantificar la probabilidad de transmisión de una infección entre pacientes hospitalizados. En este trabajo, la proximidad se expresó como el inverso de la distancia, y se creó una trayectoria para cada paciente, registrando la unidad hospitalaria en la que estuvo cada día. Sin embargo, no se utilizó ningún TDMA para calcular la proximidad total entre dos trayectorias, sino que se sumó la proximidad de cada punto de una trayectoria con cada punto de la otra. La distancia espacial entre unidades hospitalarias se basó en la proporción de pacientes que se trasladaron de una unidad a otra durante un año [30]. Finalmente, se utilizó el método *K-vecinos más cercanos continuo* (en inglés, *continuous K-nearest neighbors*)[25] para filtrar a los pacientes más cercanos. En [52], se realizó un seguimiento de la posición horaria de algunos pacientes de una residencia mediante los sensores GPS de sus teléfonos inteligentes para crear sus trayectorias. Posteriormente, se aplicaron dos TDMA (*Dynamic Time Warping* y *Longest Common Subsequence*) para obtener una matriz de similitud entre los pacientes, que luego se utilizó como entrada para un método de *clustering* jerárquico con el objetivo de detectar patrones en sus rutinas de desplazamiento.

4.2.2. Algoritmos para medir distancias entre trayectorias (TDMA)

En este capítulo, proponemos un análisis detallado basado en la similitud entre las trayectorias de los pacientes, calculada mediante un TDMA. Los estudios [85, 198, 215] son tres revisiones en las que se caracterizan, clasifican y comparan algunos de los TDMA más utilizados. Seleccionamos los algoritmos que cumplen los siguientes criterios:

- Su medida de distancia considera (o puede considerar) **información espacial y temporal**, ya que ambas dimensiones influyen en la transmisión por aire y por superficies contaminadas.
- **Definen el tiempo de forma discreta**. Solo tienen en cuenta los puntos de muestreo que definen la trayectoria y no el movimiento intermedio, ya que los pacientes no están desplazándose constantemente.

- **Utilizan medidas elásticas**. Permiten la comparación *uno-a-varios* o *uno-a-ninguno* entre puntos, y no solo la comparación del i-ésimo punto de una trayectoria con el i-ésimo punto de otra (medidas *paso a paso* (del inglés, *lock-step*)). Nuestro objetivo es determinar qué pacientes tienen mayor probabilidad de transmitir la infección, por lo que un algoritmo restrictivo no detectaría situaciones como el contacto indirecto o el caso en que dos pacientes han estado cerca, pero no exactamente dentro del mismo intervalo de tiempo.

- **Utilizan medidas basadas en heurísticas** o, también llamadas, "no basadas en aprendizaje". Su estrategia consiste en identificar los emparejamientos de puntos óptimos mediante la comparación de pares de puntos de dos trayectorias y calcular la distancia total mediante la agregación de los valores promedio entre todos estos puntos [238]. Estas medidas pueden adaptarse fácilmente a diferentes dominios. Por el contrario, las *medidas basadas en aprendizaje* dependen de técnicas de *machine learning* con las que se reconstruyen los datos de entrada de alta dimensionalidad (tienen un alto número de variables) en una nueva representación de baja dimensionalidad [85]. Normalmente, están desarrolladas para un dominio específico, siendo las redes de carreteras el más común.

De los TDMA que cumplen estas características, elegimos los siguientes tres para la comparación:

- ***Dynamic Time Warping*** (**DTW**) (en español, *Deformación dinámica del tiempo*)

- ***Spatiotemporal Linear Combine*** (**STLC**) (en español, *Combinación lineal espacio-temporal*)

- ***Spatiotemporal Longest Common Subsequence*** (**ST-LCSS**) (en español, *Subsecuencia común más larga espacio-temporalmente*)

Otros TDMA comúnmente utilizados, como la *Distancia de Edición*, la *Distancia Hausdorff* o la *Distancia Fréchet*, no cumplen todos los requisitos mostrados, por lo que no han sido seleccionados para este capítulo.

4.2.3. Conceptos preliminares

En esta subsección introducimos algunos conceptos preliminares relacionados con las trayectorias a los que recurriremos a lo largo del capítulo.

Definición 1: Trayectoria. Una trayectoria T_T es una secuencia ordenada de puntos de muestreo que representa el camino seguido por un objeto móvil. En este capítulo, la trayectoria T_n describe los movimientos dentro del hospital del paciente n.

$$T_n = \{p_1,\ p_2,\ \ldots,\ p_m\} \qquad donde \quad p_i(1 \leq i \leq m) \tag{4.1}$$

$m = |T_n|$ representa la longitud de la trayectoria, es decir, el número de puntos de muestreo. Usamos T_n^i para referirnos a i-ésimo punto de T_n.

Definición 2: Punto de muestreo. El punto de muestreo p_i (para simplificar, punto) representa la ubicación espacial de un paciente en un instante de tiempo específico. Es decir, es un par formado con la identificación de la dimensión espacial y la marca de tiempo (en inglés, *timestamp*) de la dimensión temporal. En este capítulo, tomamos como referencia el modelo espacio-temporal presentado en la Sección 3.2.1 del Capítulo 3. Formalizamos la ubicación espacial como un par con los identificadores de dos clases de nodos: la *Cama* donde estuvo el paciente y la *UnidadHospitalización* que lo atendió.

$$p_i = ((CamaID,\ UnidadHospitalizaciónID),\quad timestamp) \tag{4.2}$$

Definición 3: Cabeza(T). Es una función que devuelve el primer punto de muestreo de la trayectoria T.

$$Cabeza(T) = p_1 \tag{4.3}$$

Definición 4: Resto(T). Es una función que devuelve la cola de la trayectoria T, es decir, todos los puntos de muestreo excepto el primero.

$$Resto(T) = [p_2,\ p_3,\ \ldots,\ p_m] \tag{4.4}$$

Definición 5: Algoritmos para medir distancias entre trayectorias (TDMA). Un TDMA es una función $d(T_1, T_2) \in \mathbb{R}_0^{+\infty}$ que evalúa la distancia entre dos trayectorias T_1 y T_2 en base a la distancia entre sus puntos de muestreo [198].

4.3 Propuesta: el método StESPT

En este capítulo proponemos el método StESPT, cuyos pasos son presentados en esta sección. Primero, buscamos aquellos pacientes que han sido infectados por una determinada bacteria en un período de tiempo especificado. A continuación, creamos una trayectoria para cada paciente a partir de sus *Eventos* que tienen una *Localización* asociada a través de la relación *tieneLocalización*. Los siguientes pasos consisten en definir la similitud epidemiológica espacio-temporal entre dos puntos de muestreo y utilizar un TDMA para calcular la similitud entre las trayectorias de cada par de pacientes. Con este resultado, creamos una *matriz de similitud*, que es representada como un mapa de calor. Finalmente, sobre las similitudes entre cada par de pacientes aplicamos el algoritmo de *clustering k-means* (en español, *k-medias*) para detectar grupos de pacientes que hayan estado espacio-temporalmente cerca de forma que se pueda obtener una relación epidemiológica entre los casos.

4.3.1. Paso 1: Obtener los pacientes infectados

El primer paso de StESPT es recuperar de la BDOG a los pacientes que pueden formar parte de un posible brote. Para esta tarea, hemos utilizado la consulta epidemiológica "**CE2 - Detección de un brote en una *Localización***" definida en la Sección 3.2.2. Esta consulta nos permite recuperar aquellos *Pacientes* que, estando en una determinada *Planta* durante un período de tiempo específico, se les realizó una prueba microbiológica con un resultado positivo (*TestMicro*) de una bacteria específica durante un período determinado, independientemente de si esta prueba ocurrió mientras el paciente estaba en esa *Planta*.

Cabe señalar que el resultado positivo de una prueba microbiológica indica que sabemos que, a partir del momento en que obtenemos el resultado, el paciente está infectado. Sin embargo, este momento no tiene que coincidir con el momento en que el paciente se infectó realmente. Por ejemplo, existen infecciones con un período de incubación de varios días, lo que puede provocar que un paciente se infecte en una ubicación del hospital y tenga un resultado positivo en otra.

4.3.2. Paso 2: Convertir las estancias de los pacientes en trayectorias

El segundo paso consiste en obtener las trayectorias de los pacientes identificados en el primer paso. Proponemos definir un **Tiempo de Búsqueda para las Trayectorias de Pacientes** (**TBTP**), que sería distinto al tiempo de búsqueda de los pacientes infectados. El *tiempo final del TBTP* sería la fecha del último *TestMicro* positivo pues, después de este momento, todos los pacientes ya están infectados. Por tanto, el posterior contacto entre ellos no afectaría a la transmisión entre ellos. El *tiempo inicial del TBTP* debería ser anterior al tiempo inicial de búsqueda de pacientes infectados.

Dado que un *TestMicro* positivo no representa el momento exacto de la infección, para definir el tiempo inicial del TBTP, hemos añadimos un margen de varios días previos a la fecha del primer *TestMicro* positivo. Por ejemplo, podemos añadir el tiempo medio de incubación de la bacteria según la literatura científica.

Una vez definido el TBTP, buscamos todos los *Eventos* que tengan asociada una *Localización* mediante la arista *tieneLocalización* que ocurran durante el TBTP. Vamos a utilizar una **Frecuencia de Muestreo** (**FM**) para transformar estos *Eventos* en una trayectoria de puntos de muestreo. Esta frecuencia se puede configurar para diferentes unidades temporales, en función del nivel de detalle del análisis. Como resultado de este muestreo, logramos que todos los puntos de todas las trayectorias estén alineados temporalmente. Aunque también es posible que los primeros y últimos puntos de cada trayectoria pueden no coincidir. Definimos cada punto de muestreo como en la Ecuación 4.2.

4.3.3. Paso 3: Similitud epidemiológica espacio-temporal entre puntos

En esta sección explicamos el proceso por el cual se llega a la definición de la **Similitud Epidemiológica eSpacio-Temporal**, sim_{ST}, entre un par de puntos de muestreo.

4.3.3.a. Distancia espacio-temporal

Partimos de una definición de la **distancia espacio-temporal**, d_{ST}, entre dos puntos de muestreo como una combinación lineal ponderada (parámetro α) de la distancia espacial d_{sp} y la distancia temporal d_{temp} [150, 186, 238].

$$d_{ST}(p_1, p_2) = \alpha \cdot d_{sp}(p_1, p_2) \ + \ (1 - \alpha) \cdot d_{temp}(p_1, p_2) \tag{4.5}$$

$\alpha \in [0, 1]$ puede modificarse en función de la infección y las características espaciales y temporales del brote. Es decir, α nos permite ajustar la importancia relativa del espacio y el tiempo en la distancia espacio-temporal.

Sin embargo, nuestro objetivo no es medir cuán distantes son las trayectorias, sino su cercanía: cuanto menor sea la distancia entre ellas, mayor será su similitud. Por lo tanto, la similitud puede interpretarse como el opuesto de la distancia, lo que nos lleva a definir la similitud espacio-temporal entre puntos de muestreo.

4.3.3.b. Distancia espacial

La dimensión espacial suele definirse utilizando coordenadas bidimensionales o tridimensionales, midiéndose la distancia mediante la distancia euclidiana. En este capítulo, proponemos utilizar una distancia basada en la dimensión espacial descrita

en el modelo propuesto en la Sección 3.2.1, utilizando tanto las clases y relaciones definidas en la jerarquía de *Localizaciones* como las complementarias (*ZonaLógica*, *opuesto*, *contiguo*, etc.)

Dado un par de *Camas*, definimos su distancia espacial, d_{sp}, como la suma de la propiedad *coste* definida en las aristas de la dimensión espacial física que forman el camino más corto entre ellos. Hemos creado un matriz para almacenar la distancia entre cada par de *Camas*. Con el objetivo de que las distancias espaciales y temporales sean comparables, normalizamos los valores de la matriz al **intervalo** $[0,\ 1]$. El valor máximo es la suma de *costes* entre dos *Camas* que solo tienen en común el *Hospital* (o el nodo de la clase de mayor nivel implementada). En caso de que se implementen todas las relaciones de la dimensión espacial, el valor menor sería el *coste* entre dos *Camas* unidos por una arista *contiguo* u *opuesto*, según cuál tenga el menor valor.

Los trabajadores sanitarios también pueden representar un medio de transmisión de la infección, pues los pacientes atendidos por el mismo sanitario tienen una mayor probabilidad de infectarse cuando hay un caso positivo en la *UH* o *Servicio*. Para representar esta situación, multiplicamos la distancia espacial entre dos puntos de muestreo que han sido atendidos por el mismo personal sanitario por un valor entre $(0, 1)$. En nuestros experimentos, hemos utilizado $0,5$ cuando los pacientes sean atendidos por la misma *UH* y $0,7$ cuando pertenezcan al mismo *Servicio*.

4.3.3.c. Distancia temporal

Hemos definido la distancia temporal entre dos puntos, p_1 y p_2, en base a dos aspectos.

El primer componente es el **valor absoluto de la diferencia entre sus marcas de tiempo** ($diff_{tmp}$). De manera similar a la distancia espacial, normalizamos esta diferencia al **intervalo** $[0,\ 1]$. El valor máximo es la diferencia entre los tiempos que definen el TBTP. El valor mínimo es el de dos marcas de tiempo iguales, es decir, 0.

El segundo componente es la **transmisibilidad** (o **velocidad de transmisión**) de la infección. Cuanto mayor sea la transmisibilidad, mayor será la probabilidad de que un contacto directo e indirecto dé lugar a un nuevo paciente infectado. Por lo tanto, se puede utilizar para determinar si dos puntos están temporalmente cerca. Son varios los parámetros epidemiológicos que se utilizan en la literatura científica para representar la transmisibilidad de la infección.

Entre estos parámetros, podemos destacar:

- **Número reproductivo básico** (R_0). Representa el número promedio de personas infectadas por una persona infecciosa. Es decir, estima la velocidad a la que una enfermedad infecciosa puede propagarse a través de una población. R_0 suele utilizarse para inferir el tamaño potencial de un brote, y existen diversos enfoques para estimar su valor basados en modelos de simulación. Su rango es $[-\infty, +\infty)$. [50]

- **Tasa de crecimiento exponencial** (r). Representa el cambio porcentual positivo de las personas infectadas en una población por unidad de tiempo (por ejemplo, un día). Solo puede utilizarse cuando el número de infectados sigue una curva exponencial con base e. Es decir, se utiliza para modelar crecimientos rápidos y sin una previsión inicial de ralentización. Sin embargo, su valor puede cambiar rápidamente a lo largo del progreso de la infección. Su rango es $[-\infty, +\infty)$. [126]
- **Tasa de transmisión** (β). Representa la probabilidad de que un individuo en estado *Susceptible* pase a estar *Infectado* (la infección es descrita con un modelo *SIR*) o *Expuesto* (modelo *SEIR*) tras un contacto con un individuo en estado *Infectado*. Su valor puede calcularse dividiendo la incidencia (nuevos casos positivos por unidad de tiempo) entre la prevalencia (total de casos de la infección). Su rango es $[0,\ 1]$. [145]

Dada la semántica de los tres parámetros, todos ellos podrían ser opciones válidas para representar la transmisibilidad de la infección en nuestra definición de distancia temporal. Sin embargo, β presenta varios aspectos ventajosos: un cálculo sencillo y un rango fijo. Este último facilita la normalización de los datos a un rango determinado.

Formalizamos la relación entre la diferencia temporal y la transmisibilidad mediante la multiplicación del primer componente, $diff_{tmp}$, por β.

$$diff_{tmp}(p_1, p_2) = |p_{1_{time}} - p_{2_{time}}| \tag{4.6}$$

$$d_{tmp}(p_1, p_2) = \beta \times diff_{tmp}(p_1, p_2) \tag{4.7}$$

4.3.3.d. Similitud epidemiológica espacio-temporal

Hemos definido las distancias espacial y temporal por separado, siendo posible calcular la distancia espacio-temporal entre dos puntos definida en la Ecuación 4.5.

Dado que la similitud puede interpretarse como el inverso de la distancia, para obtener la similitud epidemiológica espacio-temporal invertimos las ecuaciones para calcular las distancias espacial y temporal de la manera más lineal posible (ver Figura 4.1). Además, se debe cumplir que: *a)* cuando dos puntos tienen la misma *Localización*, la similitud espacial debe ser cero; y *b)* cuando dos puntos tienen la misma marca de tiempo, la similitud temporal debe ser cero.

En el caso de la distancia temporal, el parámetro β también debe ser invertido. En caso de no invertirse, los valores altos de β implicarían una mayor distancia temporal, lo que indicaría una menor probabilidad de infección.

Las definiciones de **similitud espacial** y **similitud temporal** entre dos puntos de muestreo quedan de la siguiente manera:

$$sim_{sp}(p_1, p_2) = 1 \; - \; d_{sp}(p_1, p_2) \tag{4.8}$$

$$sim_{tmp}(p_1, p_2) = e^{\ln(\beta) \; \cdot \; |p_{1_{time}} - p_{2_{time}}|} \tag{4.9}$$

Basándonos en la ecuación 4.5, definimos la **similitud epidemiológica espacio-temporal** entre dos puntos de muestreo como una combinación lineal ponderada de 4.8 y 4.9 con $\alpha \in [0,\ 1]$:

$$sim_{ST}(p_1, p_2) = \alpha \cdot sim_{sp}(p_1, p_2) \; + \; (1 - \alpha) \cdot sim_{tmp}(p_1, p_2) \tag{4.10}$$

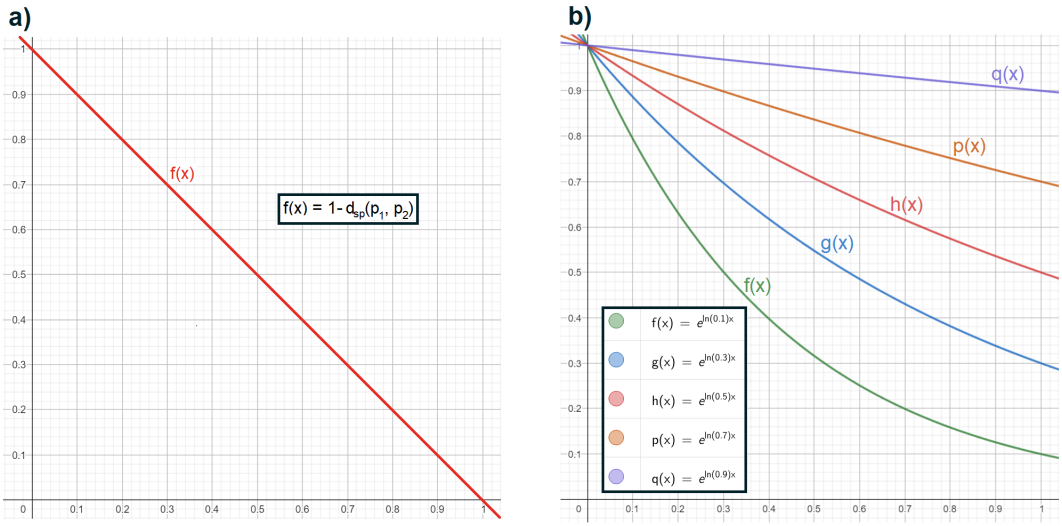


FIGURA 4.1. **Representación gráfica en el intervalo** $[0,\ 1]$ **de las ecuaciones con las que calcular las similitudes espacial y temporal. a)** Ecuación para de sim_{sp}. **b)** Ecuación de sim_{tmp}. En concreto, se representa la curva con distintos valores fijos para β, siendo $diff_{tmp}$ la variable (representada con x). Podemos observar que, a medida que los valores de β aumentan, también lo hace la d_{tmp}.

4.3.4. Paso 4: TDMA

Para evaluar la similitud entre las trayectorias de dos pacientes, hemos comparado los tres TDMA seleccionados en la Sección 4.2.2: **DTW**, **STLC** y **ST-LCSS**. Los hemos modificado ligeramente para que se adapten a la semántica de nuestro problema y, en los casos pertinentes, calculen *similitud* en lugar de *distancia*. Además, hemos diseñado una extensión de cada algoritmo para una adaptación más profunda a los requisitos semánticos del nuestro problema. En total, hemos comparado seis TDMA, cuya descripción aparece en los siguientes apartados.

Cabe señalar que, para calcular la similitud de cada par de trayectorias, seleccionamos únicamente los *puntos de intersección* entre ambas trayectorias. Es decir, solo evaluamos la similitud del período de tiempo en el que ambos pacientes han estado en el hospital, no el tiempo en que solo uno de ellos ha estado hospitalizado. Para simplificar, nos referimos a estos períodos de tiempo como **Trayectorias de Intersección** (**TTII**).

La Figura 4.2 muestra dos trayectorias que coinciden temporalmente desde el punto 3 al 7. Por tanto, al calcular la similitud entre sus dos pacientes, solo comparamos las sub-trayectorias entre estos puntos (delimitadas por líneas rojas en la figura).

FIGURA 4.2. Representación temporal de la trayectoria de intersección entre dos trayectorias.

4.3.4.a. DTW

Dynamic Time Warping (**DTW**) [227] es un algoritmo que surgió para la comparación de series temporales en las que puede encontrarse un desfase en el tiempo o variaciones en la velocidad. Se trata de un algoritmo ampliamente estudiado, quedando su versatilidad demostrada al haber sido aplicado en una gran variedad de campos, como el reconocimiento automático del habla [183], el reconocimiento del modo de andar [172], la clasificación de señales genómicas [193] o el reconocimiento de firmas manuscritas [167], entre otros. DTW busca encontrar el emparejamiento óptimo (en inglés, *optimal match*) entre los puntos de dos secuencias. Aunque se han propuesto diversas implementaciones y optimizaciones, su definición original es la de un algoritmo recursivo que busca entre todas las alineaciones posibles de puntos entre dos trayectorias, T_1 y T_2, de manera que la suma de las distancias entre puntos sea mínima y cumpla con los siguientes requisitos:

- T_1^1 debe emparejarse con T_2^1.
- T_1^m debe emparejarse con T_2^m.
- Cada punto intermedio de cada trayectoria debe emparejarse con uno o más puntos de la otra trayectoria, no estando permitido retroceder.

Por tanto, DTW busca la diferencia mínima entre dos trayectorias. Sin embargo, lo que nosotros buscamos es encontrar la similitud máxima. Cabe destacar que, al

haber definido la similitud como la inversa de la distancia, los cambios serían mínimos: calcular la similitud e ir seleccionando el camino con la mayor similitud total. Sin embargo, para evitar ciertos errores, nuestra versión de DTW calcula tanto la distancia como la similitud entre puntos, utilizando la distancia, como en su definición original, para determinar el camino óptimo. Además, hemos añadido otro cambio relacionado con la similitud entre trayectorias sin elementos: la similitud es cero cuando la longitud de una o ambas trayectorias sea cero. Si no es posible hacer ningún emparejamiento, no debería aumentar la similitud total.

Cabe destacar que, basándonos en su definición original, DTW solo evalúa la dimensión espacial. Es decir, solo calcula d_{sp} y sim_{sp}.

4.3.4.b. ST-DTW

Proponemos una extensión de DTW en la que, además de la dimensión espacial, también se evalúe el tiempo y la transmisibilidad de la infección. Hemos llamado a este nuevo algoritmo **eSpacio-Temporal DTW** (**ST-DTW**). El principal cambio de ST-DTW con respecto a DTW consiste en que calcula la distancia y similitud espacio-temporal definidas mediante las Ecuaciones 4.5 y 4.10, respectivamente. La Figura 4.3 muestra la definición de ST-DTW.

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

  donde

  - $$ST\_DTW(T_1, T_2) = \begin{cases} (0,0), & \text{si } |T_1| = 0 \ y \ |T_2| = 0 \\ (0, \infty) & \text{si } |T_1| = 0 \ o \ |T_2| = 0 \\ \begin{pmatrix} \boxed{sim_{ST}(Cabeza(T_1), Cabeza(T_2))} + A^1, \\ \boxed{d_{ST}(Cabeza(T_1), Cabeza(T_2))} + A^2 \end{pmatrix} & \textit{de lo contrario} \end{cases}$$

    donde

    - $A = \textit{seleccionar dupla de } B \textit{ donde } B^2 \textit{ es mínima},$
    - $B = \{ST\_DTW(T_1, Resto(T_2)) \cup ST\_DTW(Resto(T_1), T_2) \cup ST\_DTW(Resto(T_1), Resto(T_2))\}$

FIGURA 4.3. **Definición de ST-DTW**. Dadas dos trayectorias, T_1 y T_2, ST-DTW devuelve un par ordenado (dupla), donde el primer elemento es la *similitud* espacio-temporal y el segundo elemento es la *distancia*. Los superíndices representan la posición de un elemento dentro de una dupla. Cabe mencionar que B es un conjunto de duplas.
En rojo, destacamos las funciones para calcular la similitud y distancia espacio-temporales entre dos puntos. En **DTW**, estas funciones se sustituyen por sim_{sp} y d_{sp}.

4.3.4.c. STLC

Spatiotemporal Longest Common Subsequence (**STLC**) [187, 198] propone una combinación lineal de las similitudes espacial y temporal entre dos trayectorias para obtener su similitud espacio-temporal, calculando la similitud espacial y la similitud temporal por separado.

Para una explicación más sencilla del algoritmo de STLC, nos centramos en la dimensión espacial, dado que la similitud temporal se calcula siguiendo el mismo proceso. En primer lugar, calculamos la distancia espacial *punto-trayectoria* para cada punto de cada trayectoria, es decir, la mínima distancia espacial entre ese punto y un punto de la otra trayectoria. A continuación, estas distancias son transformadas en similitudes, y calculamos la similitud espacial punto-trayectoria promedio para cada trayectoria. Luego, sumamos ambas similitudes espaciales. Finalmente, combinamos linealmente las similitudes espacial y temporal utilizando el parámetro $\alpha \in [0,\ 1]$. Este parámetro cumple la misma función que el definido en las Secciones 4.3.3.a y 4.3.3.d. La Figura 4.4.a muestra la definición original de STLC.

Hemos adaptado STLC para que utilice las definiciones de similitud espacial y temporal descritas con las Ecuaciones 4.8 y 4.9. La Figura 4.4.b muestra la definición de nuestra adaptación de STLC. Cabe destacar que seleccionar el punto con la mínima distancia y transformar dicha distancia en similitud es equivalente a seleccionar el punto con la máxima similitud.

La versión original de STLC calcula la similitud punto-trayectoria promedio para cada trayectoria y dimensión. Sin embargo, no tiene en cuenta la proporción de tiempo que los pacientes han estado en el hospital en comparación con todo el TBTP. En el caso más extremo, si dos pacientes han estado en la cama más cercana posible durante todo el tiempo abarcado por sus TTII, su similitud será máxima, independientemente de la duración de las TTII.

Para evitar esta situación, hemos modificado STLC para obtener una tasa de la similitud en la que el sumatorio de las similitudes punto-trayectoria de cada trayectoria y dimensión no se divide entre la longitud de la trayectoria, sino entre un parámetro λ, que representa la diferencia entre los puntos que definen el TBTP.

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

donde:

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

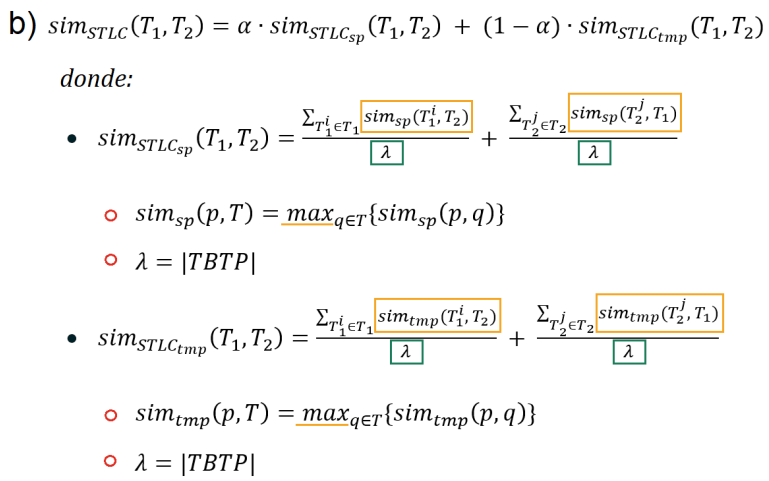


FIGURA 4.4. **Definición de STLC**. **a)** Definición original de STLC. Las ecuaciones utilizadas para transformar distancias en similitudes están marcadas en azul (espacial) y rojo (temporal). **b)** Nuestra adaptación de STLC. Los cambios introducidos están marcados en naranja y verde.

4.3.4.d. JSTLC

Proponemos una extensión de STLC a la que hemos llamado Joint STLC (en español, *Combinación lineal conJunta espacio-temporal*). Su cambio principal respecto a STLC es que en lugar de calcular por separado las similitudes punto-trayectoria espacial y temporal, calcula directamente la similitud punto-trayectoria espacio-temporal. Por tanto, utilizamos la Ecuación 4.10 para calcular sim_{ST}, en lugar de las ecuaciones para calcular sim_{sp} y sim_{tmp} por separado. La Figura 4.5 muestra la definición de JSTLC.

Con esta modificación, buscamos superar la limitación de que el punto más cercano en el espacio puede no ser el punto más cercano en el tiempo. Ahora buscamos el punto que esté más cerca en ambas direcciones según la proporción indicada por el parámetro α (ver Ecuación de la similitud epidemiológica espacio-temporal 4.10).

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

donde:

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

FIGURA 4.5. Definición de JSTLC.

4.3.4.e. ST-LCSS

Spatiotemporal Longest Common Subsequence (**ST-LCSS**) (en español, *subsecuencia común más larga espacio-temporalmente*) [198] es un algoritmo que devuelve la longitud de la subsecuencia común más larga (LCSS) entre dos trayectorias T_1 y T_2. Esta subsecuencia incluye los emparejamientos entre los puntos de muestreo de cada trayectoria siguiendo las siguientes condiciones:

- Un punto no puede coincidir con más de un punto.
- No es necesario que los siguientes puntos emparejados sean consecutivos del par anterior, pero deben respetar el orden de las trayectorias. Es decir, si hay un emparejamiento entre T_1^i y T_2^j, entonces no puede haber un emparejamiento entre T_1^{i+k} y T_2^{j-l} (o T_1^{i-k} y T_2^{j+l}).
- Puede haber puntos sin emparejar.

Una definición estricta de emparejamiento exigiría que los dos puntos de muestreo compartiesen exactamente los mismos valores en las dimensiones espacial y temporal. Sin embargo, encontrar dos puntos con las mismas *Localizaciones* y marca temporal es complicado. Además, para nuestro problema, sería semánticamente imposible que dos *Pacientes* hubiesen estado en el mismo *Cama* durante el mismo tiempo. Por ello, ST-LCSS introduce dos parámetros, ε y δ, que definen la distancia máxima en el espacio y en el tiempo a las que pueden estar dos puntos para considerar que puede haber un emparejamiento entre ellos.

ST-LCSS mide la similitud entre dos trayectorias, entendida como el número de emparejamientos que conforman la LCSS. Por tanto, nuestra versión del algoritmo implementa pocos cambios respecto a la original. Los más significativos son los siguientes:

- En lugar de la distancia máxima, ε y δ representan las similitudes espaciales y temporales mínimas que dos puntos deben tener para que puedan ser emparejados.
- Nuestra versión de ST-LCSS calcula y devuelve tanto la longitud de la LCSS como la suma de las similitudes espacio-temporales entre cada par de puntos coincidentes. Al igual que en la versión original, la solución de ST-LCSS es la solución con mayor número de emparejamientos, pero, en caso de empate, se selecciona aquella con la mayor similitud espacio-temporal.

La Figura 4.6 muestra la definición de nuestra versión de ST-LCSS.

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

donde

- $ST\_LCSS(T_1, T_2) = \begin{cases} (0,0), & \text{si } |T_1| = 0 \text{ or } |T_2| = 0 \\ \begin{pmatrix} ST\_LCSS\big(Resto(T_1), Resto(T_2)\big)^1 \\ + sim_{ST}\big(Cabeza(T_1), Cabeza(T_2)\big) \ , \\ ST\_LCSS\big(Resto(T_1), Resto(T_2)\big)^2 + 1 \end{pmatrix}, & \text{si } sim_{sp}\big(Cabeza(T_1), Cabeza(T_2)\big) \geq \varepsilon \\ & \text{y } sim_{temp}\big(Cabeza(T_1), Cabeza(T_2)\big) \geq \delta \\ A & \textit{de lo contrario} \end{cases}$

donde

- A = *seleccionar dupla de* B *donde* B^2 *es máxima,*
- $B = \{ST\_LCSS\big(T_1, Resto(T_2)\big) \cup ST\_LCSS(Resto(T_1), T_2)\}$

FIGURA 4.6. **Definición de nuestra versión de ST-LCSS**. Dadas dos trayectorias, T_1 y T_2, ST-LCSS devuelve un par ordenado (dupla), donde el primer elemento es la similitud espacio-temporal entre T_1 y T_2, y el segundo elemento es la longitud de la LCSS. Los superíndices representan la posición de un elemento dentro de una dupla. Cabe mencionar que B es un conjunto de duplas.

4.3.4.f. ST-LCSS-WTW

Hemos diseñado una extensión de ST-LCSS, denominada ***ST-LCSS With Time Window*** (**ST-LCSS-WTW**) (en español, *ST-LCSS con Ventana Temporal*), con la que pretendemos mejorar la precisión de ST-LCSS para evaluar la similitud entre dos trayectorias.

ST-LCSS sigue un esquema voraz (en inglés, *greedy*) en el que no se explora el posible espacio de soluciones al completo: cuando empareja dos puntos, no reconsidera sus elecciones en los siguientes pasos. Por tanto, es posible que no se encuentre la solución óptima. Por ejemplo, el emparejamiento entre dos puntos con la mayor similitud puede resultar posteriormente en un menor número de emparejamientos o una menor similitud total. Además, la similitud entre T_1^i y T_2^{j+1} (o entre T_1^{i+1} y T_2^j) podría ser mayor que la calculada para un emparejamiento entre T_1^i y T_2^j.

Nuestro objetivo es identificar tantas conexiones como sea posible entre cada par de pacientes, dado que, a mayor número de conexiones entre ellos, mayor es la probabilidad de que haya podido ocurrir un contagio entre ambos. Para ello, proponemos la adición de una ventana temporal ω que permita explorar varias soluciones en cada paso de la ejecución del algoritmo.

Dado un par de puntos de muestreo, T_1^i y T_2^j, se evalúan todos los posibles emparejamientos entre cada uno de estos puntos y cualquier otro punto de la otra trayectoria dentro del intervalo $[T_2^{j-\omega}, T_2^{j+\omega}]$ o $[T_1^{i-\omega}, T_1^{i+\omega}]$, respectivamente. Para cada emparejamiento, se llama recursivamente a ST-LCSS-WTW. Finalmente, y de manera similar a ST-LCSS, ST-LCSS-WTW devuelve aquella solución con el mayor número de emparejamientos.

El parámetro ω se define en términos de puntos de muestreo. Proponemos que el tiempo que represente sea menor que el valor de $diff_{tmp}$ utilizado para calcular δ. De lo contrario, habría que evaluar pares de puntos cuya similitud temporal no estaría dentro del rango definido.

La Figura 4.7 muestra la definición de ST-LCSS-WTW.

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

  donde

  - $ST\_LCSS\_WTW(T_1, T_2, i, j, U) = \begin{cases} (0,0), & \text{si } i \geq |T_1| \text{ o } j \geq |T_2| \\ \begin{pmatrix} sim_{ST}(T_1^i, T_2^j) + A^1, \\ 1 + A^2 \end{pmatrix}, & \text{si } |B| > 0 \\ C & \textit{de lo contrario} \end{cases}$

    donde

    - A = *seleccionar dupla de B donde* B^2 *es máxima,*
    - $B = \{ST\_LCSS\_WTW(T_1, T_2, i+1, j+1, U \cup \{T_1^i, T_2^j\}) \mid sim_{sp}(T_1^i, T_2^j) \geq \varepsilon, sim_{temp}(T_1^i, T_2^j) \geq \delta, T_1^i \notin U, T_2^j \notin U\}$
      $\cup \{ST\_LCSS\_WTW(T_1, T_2, i, j+1, U \cup \{T_1^{i+x}, T_2^j\}) \mid -\omega \leq x \leq \omega, i + x > 0, sim_{sp}(T_1^{i+x}, T_2^j) \geq \varepsilon, sim_{temp}(T_1^{i+x}, T_2^j) \geq \delta, T_1^{i+x} \notin U, T_2^j \notin U\}$
      $\cup \{ST\_LCSS\_WTW(T_1, T_2, i+1, j, U \cup \{T_1^i, T_2^{j+x}\}) \mid -\omega \leq x \leq \omega, j + x > 0, sim_{sp}(T_1^i, T_2^{j+x}) \geq \varepsilon, sim_{temp}(T_1^i, T_2^{j+x}) \geq \delta, T_1^i \notin U, T_2^{j+x} \notin U\}$
    - C = *seleccionar dupla de D donde* D^2 *es máxima,*
    - $D = \{ST\_LCSS\_WTW(T_1, T_2, i, j+1, U)\} \cup ST\_LCSS\_WTW(T_1, T_2, i+1, j, U)\}$

FIGURA 4.7. **Definición de ST-LCSS-WTW**. Dadas dos trayectorias, T_1 y T_2, ST-LCSS-WTW devuelve un par ordenado (dupla), donde el primer elemento es la similitud espacio-temporal entre T_1 y T_2, y el segundo elemento es la longitud de la LCSS. Los superíndices representan la posición de un elemento dentro de una dupla. Cabe mencionar que B y C son conjuntos de duplas. ω es la ventana temporal, expresada en puntos de muestreo. i y j son los índices para recorrer T_1 y T_2, respectivamente. U es el conjunto de puntos de muestreo que ya han sido emparejados. Para una mejorar legibilidad, algunas partes han sido marcadas en distintos colores.

La Figura 4.8 muestra un ejemplo con el que entender mejor la diferencia entre ST-LCSS y ST-LCSS-WTW. En ella aparecen dos trayectorias, T_1 y T_2, que han sido representadas como cadenas de caracteres. Para el ejemplo, consideramos que un emparejamiento es posible cuando los dos puntos tienen la misma letra y hay una diferencia máxima de una posición entre ellos. Además, para ST-LCSS-WTW, ω = 1. Podemos apreciar que T_1 y T_2 representan dos cadenas con las mismas letras, pero en diferente orden. Los resultados de la ejecución de ST-LCSS y ST-LCSS-WTW son distintos:

- Al ejecutar ST-LCSS, obtenemos 2 emparejamientos: cuando se emparejan las A (o las B), la primera B (o A) ya no puede formar parte de ningún otro emparejamiento, por lo que el único emparejamiento posible es entre las C.
- Al ejecutar ST-LCSS-WTW, obtenemos 3 emparejamientos: en este algoritmo es posible obtener dos emparejamientos entre las A y las B, que se suman al emparejamiento entre las C.

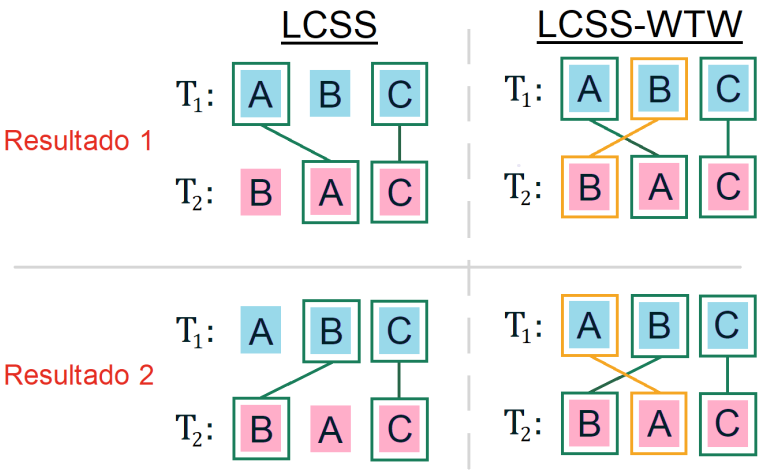


FIGURA 4.8. Ejemplo en el que se ejecuta ST-LCSS y ST-LCSS-WTW sobre un mismo par de trayectorias, T_1 y T_2.

4.3.4.g. Resultados de los TDMA en forma de matriz de similitud

Como resultado de la ejecución de cada TDMA para cada par de pacientes, obtenemos una matriz cuadrada, $M_{p\times p}$, donde p es el número de pacientes evaluados y $M_{i,j}$ representa la similitud epidemiológica espacio-temporal entre las trayectorias de intersección de los pacientes i y j. Hemos llamado a esta matriz como *matriz de similitud* (MS), la cual comprende las similitudes entre las TTII de los pacientes. Cada una de sus filas (o columnas) es un *vector de similitud* (VS) de tamaño p, que contiene las similitudes entre un paciente y los demás (incluido él mismo, con una similitud 0).

Cabe destacar que los resultados de DTW, ST-DTW, ST-LCSS y ST-LCSS-WTW se encuentran en el rango $[0, |TBTP| \times \gamma]$, donde $|TBTP|$ es el número de puntos del TBTP tras el muestreo y γ es la máxima similitud posible entre dos puntos de muestreo (dos camas vecinas en la misma marca de tiempo, o la misma cama en dos marcas de tiempo consecutivas). En cambio, los resultados de STLC y JSTLC se encuentran en el rango $[0, 2]$ cuando las distancias espaciales y temporales están normalizadas a $[0,\ 1]$, como es el caso de este capítulo.

Para comparar las MS generadas por los seis TDMA, hemos normalizado los resultados de todos los TDMA al rango $[0,\ 1]$. Estos valores normalizados pueden representarse gráficamente en forma de mapa de calor (en inglés, *heatmap*), facilitando la detección visual de grupos de pacientes. Por ejemplo, una fila (o columna) con varias celdas cálidas representa un paciente altamente interconectado, mientras que varias celdas cálidas que comparten el mismo conjunto de pacientes pueden indicar un grupo de pacientes cercanos.

En este punto queremos recordar que la similitud entre dos pacientes no es una probabilidad de transmisión, sino una representación numérica de su proximidad en el espacio y el tiempo en comparación con otros pacientes.

4.3.5. Paso 5: Método de *clustering*

Aunque los mapas de calor pueden ayudar a obtener una representación visual, no resultan intuitivos para identificar agrupar a los pacientes pues depende de la ordenación de la matriz. En consecuencia, hemos utilizado un algoritmo de *clustering* para agrupar los pacientes en función de su similitud, de manera que cada *clúster* (grupo) esté formado por pacientes cercanos. Estos grupos podrían representar diferentes brotes o pacientes cuyas *Localizaciones*, *Servicios* o *UH* compartidos deberían ser investigados como posibles focos de transmisión.

Hemos aplicado el algoritmo de *clustering* ***k-means*** para los resultados de cada TDMA, cuya entrada está compuesta por los *vectores de similitud* (*VS*) de cada paciente.

K-means es un algoritmo de clustering particional ampliamente utilizado que, mediante un proceso iterativo, descompone el conjunto de datos de entrada en k particiones disjuntas (*clústeres*), representadas por un centroide. El centroide de un clúster es un vector con las mismas variables que los datos de entrada, cuyo valor para cada variable es la media de los valores para dicha variable de todos los puntos asignados al clúster. El objetivo de k-means es minimizar la distancia (calculada como la distancia euclidiana) de cada punto a su centroide [70, 77]. Como resultado, los VS (y sus pacientes asociados) quedan agrupados en varios clústeres.

K-means requiere que se especifique como parámetro de entrada el valor de k, por lo que para cada TDMA buscamos su valor óptimo. Para ello, ejecutamos el **método *Silhouette*** [176]: para cada TDMA, aplicamos k-means con diferentes valores de k en el rango $[2, p]$ y seleccionamos aquel cuyo **coeficiente de *Silhouette*** (**CS**) sea el más alto.

SC es una métrica de validación de clústeres ampliamente utilizada [77, 160]. En ella se combinan la evaluación de dos propiedades deseables en un resultado de *clustering*: **cohesión**, que representa cuán cerca están los puntos de un mismo clúster; y **separación**, que representa cuán alejados están los puntos pertenecientes a distintos clústeres. En concreto, la separación se calcula en base a la distancia promedio entre los puntos de un clúster y los del clúster más cercano. El CS se define en el intervalo $[-1, 1]$, donde:

- Los valores positivos indican alta separación entre clústeres (los puntos están "bien agrupados").
- Los valores negativos indican que hay clústeres superpuestos.
- Los valores cercanos a 0 indican que los datos están distribuidos uniformemente, es decir, es difícil de determinar a qué clúster pertenece cada punto.

4.4 Validación

Hemos demostrado la idoneidad de nuestro método a través de un experimento en el que hemos utilizado StESPT para analizar un brote de *Clostridium difficile* (*C. diff*) en un hospital. Hemos analizado cómo cada uno de los seis TDMA propuestos han captado la semántica del problema y las relaciones espacio-temporales entre los pacientes infectados.

4.4.1. Conjunto de datos y herramientas

Hemos creado un conjunto de datos sintéticos para la evaluación de StESPT en este experimento, utilizando el mismo modelo de simulación [98] para la validación del modelo de datos y consultas epidemiológicas en la Sección 3.4. Al igual que en dicha sección, hemos generado nuestro conjunto de datos utilizando los mismos valores que los propuestos en [98] para los parámetros que describen el comportamiento epidemiológico de la infección.

Nuestro conjunto de datos abarca los **tres primeros meses del año 2024**. En ellos, observamos que, desde el **28 de febrero hasta el 6 de marzo**, la evolución de casos positivos de *C. diff* podría representar un brote. La Figura 4.9 muestra, para este período de tiempo, el número diario de pacientes infectados confirmados mediante prueba microbiológica.

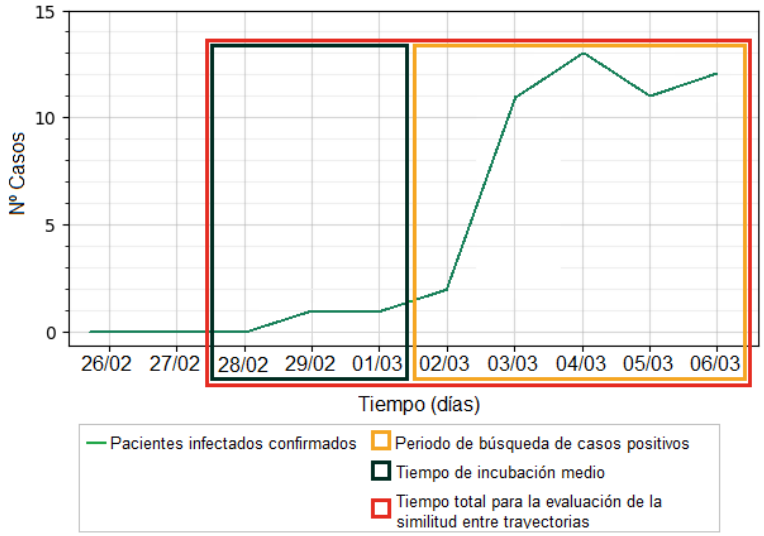


FIGURA 4.9. Ejemplo en el que se ejecuta ST-LCSS y ST-LCSS-WTW sobre un mismo par de trayectorias, T_1 y T_2.

Hemos diseñado una disposición hospitalaria basada en un hospital que, según los estándares españoles, podría considerarse de grandes dimensiones y que tendría un área de influencia de 250.000 personas.

Este hospital tiene **729** ***Camas*** distribuidas entre **4 Plantas**. La ***Planta 0*** (ver Figura 4.10) cuenta con:

- **4** ***Habitaciones*** de *Urgencias* (A&E) con **5** ***Camas*** cada una.
- **4** ***Habitaciones*** de *Cuidados Intensivos* (UCI) con **4** ***Camas*** cada una.
- **24** ***Habitaciones*** de *Radiología* con **1** ***Cama*** cada una.
- **27 quirófanos**. Es decir, 27 *Habitaciones* para cirugía con **1** ***Cama*** cada una.

Las otras **3 Plantas** se utilizan para ***Hospitalizaciones***. En cada una de las ***Plantas 1*** y ***2*** trabajan **7** ***Servicios*** con un total de **13** ***UH***: en la *Planta 1* están los *Servicios S0* a *S6*, y en la *Planta 2* están los *Servicios S7* a *S13*. En la ***Planta 3*** trabajan **14** ***UH*** de **3** ***Servicios*** diferentes (*S14* a *S16*). **Cada** ***UH*** **tiene entre 6 y 10** ***Habitaciones***, y **cada** ***Habitación*** **tiene 2** ***Camas***. En total, hay **17 Servicios para** ***Hospitalizaciones***, 1 para *A&E*, 1 para *UCI* y 1 para *Radiología*. Todos los *Servicios*, excepto *Radiología*, tienen una *UH* para cirugías.

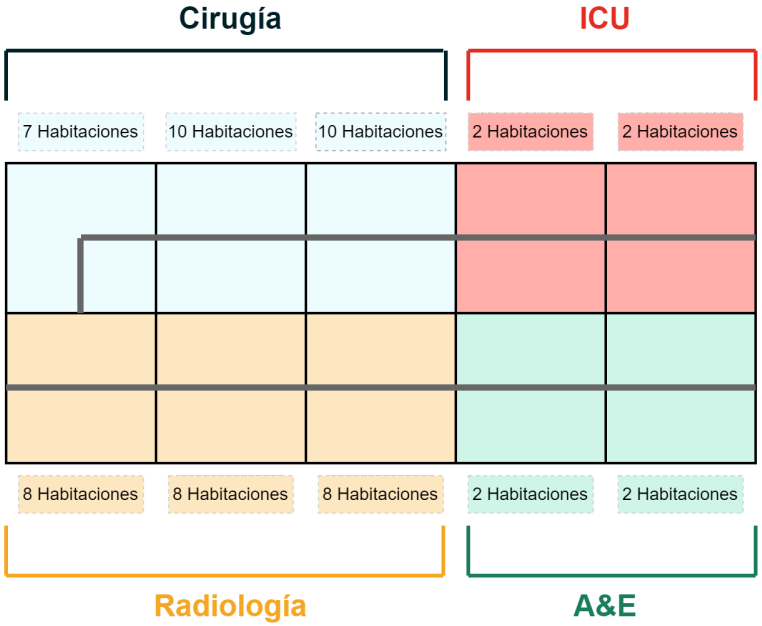


FIGURA 4.10. Representación esquemática de la *Planta 0* en la que se muestra cómo se distribuye el espacio y las *Habitaciones* entre *Servicios*, además de los *Pasillos* (líneas en gris).

En cuanto a la asignación de los valores de la propiedad *coste* en las relaciones de la dimensión espacial del modelo de datos, la Figura 4.11 muestra los valores para esta propiedad en función del tipo de relación y la clase de sus nodos de origen y destino. Estos valores han sido calculados en función de las dimensiones (en metros) de una *Cama* hospitalaria y una *Habitación* para hospitalizaciones con dos *Camas*, una al

lado de la otra (ver Apéndice A). Además, cuando dos pacientes hayan sido atendidos por el mismo *Servicio* o *UH*, la distancia espacial entre esos puntos es multiplicada por $0,7$ o $0,5$, respectivamente.

La salida del simulador es transformada en grafo de conocimiento en formato RDF*, que es almacenado en GraphDB 10.4.0. Hemos implementado StESPT íntegramente en Python 3.10.8. Utilizamos la librería *SPARQLWrapper* [173] para la conexión con la base de datos, *scikit-learn 1.5.2* [31] para la implementación del algoritmo de *clustering*, *Matplotlib 3.9.2* [87, 201] para la creación de imágenes y *NetworkX 3.2.1* [76] para los algoritmos de grafos.

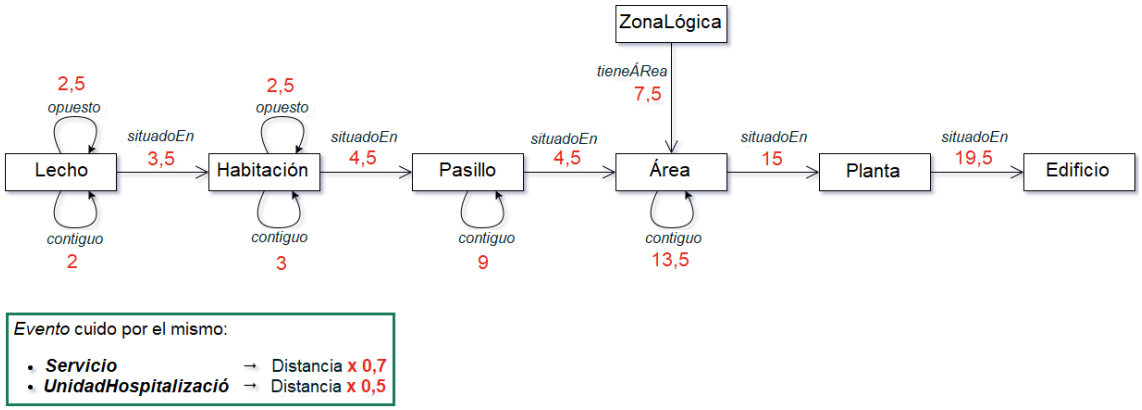


FIGURA 4.11. Asignación de los valores para la propiedad *coste* en función del tipo de relación y la clase de sus nodos de origen y destino.

4.4.2. Resultados

En las siguientes secciones, presentamos los resultados para cada uno de los pasos del método StESPT.

4.4.2.a. Paso 1: Obtener los pacientes infectados

Como podemos observar en la Figura 4.9, del 2 de marzo al 4 de marzo, hubo un aumento significativo de hasta 13 casos de *C. diff*, que se mantuvo estable durante los siguientes dos días. La mayoría de estos pacientes estuvieron en la *Planta 2* en algún momento durante estos días, por lo que podrían estar relacionados espacio-temporalmente. En consecuencia, buscamos a los *Pacientes* que tuvieron un *TestMicro* de *C. diff* entre el **2 de marzo y el 6 de marzo** y que pasaron algún tiempo en la ***Planta 2***. Como resultado, obtenemos un total de **17 *Pacientes***.

4.4.2.b. Paso 2: Obtener las trayectorias de los pacientes

El primer y último *TestMicro* de los 17 pacientes obtenidos en el paso 1 ocurrieron el *2 de marzo* y el *6 de marzo*, respectivamente. El período de incubación de *C. diff* suele ser de menos de una semana, habiendo estudios que indican que la infección puede ocurrir dentro de las primeras 48 a 72 horas tras la exposición [45, 192]. Para nuestro experimento, añadimos tres días antes del 2 de marzo para obtener los *Eventos* de los 17 pacientes. Así, el **TBTP abarca desde el 28 de febrero al 6 de marzo**.

Hemos aprovechado que el modelo de simulación tiene una precisión de 8 horas y hemos establecido dicho valor como nuestra frecuencia de muestreo para transformar los *Eventos* en trayectorias. Es decir, cada paso en la ejecución de la simulación se corresponde con un punto de muestreo. Dado que nuestro TBTP es de 8 días, hemos evaluado trayectorias de intersección (TTII) que tienen una longitud máxima de **24 puntos de muestreo**. La Figura 4.12 muestra las trayectorias de cada paciente, diferenciando entre las *Camas* donde fueron hospitalizados (Figura 4.12.a) y la *UH* que los atendió (Figura 4.12.b).

Podemos observar que los pacientes tienen diferentes fechas de inicio y fin de sus trayectorias, y que algunos pacientes no se encuentran en *Localizaciones* cercanas en momentos próximos. Este escenario nos lleva a evaluar TTII de diferentes longitudes (la más corta tiene 9 pasos, y la más larga tiene 24) y diversas circunstancias espacio-temporales. Sin embargo, casi todos los pacientes compartieron dos características:

- Sus trayectorias comienzan en la *Planta 0*. Concretamente, en las dos áreas adyacentes (*3B* y *4B*) dedicadas al *Servicio de A&E*.
- Todos los pacientes tuvieron, al menos, un punto de muestreo en la *Planta 2*. Varios de ellos coincidieron en el mismo *Área* o en Áreas adyacentes.

También podemos observar que algunos *Servicios* y *UH* han sido compartidos por varios pacientes, como el *Servicio 7* y la *UH 8B*.

4.4.2.c. Pasos 3 y 4: Similitud espacio-temporal entre trayectorias

En esta sección, analizamos los resultados de cada TDMA al ser aplicados sobre las TTII entre los 17 pacientes obtenidos en el paso 1. Presentamos cada MS con sus valores normalizados al rango $[0,\ 1]$ en forma de un mapa de calor, donde las similitudes bajas se representan en azul oscuro y las similitudes más altas aparecen en tonos naranjas y rojos.

En una comparación inicial de los resultados de **DTW** y **ST-DTW** (Figuras 4.13.a y 4.13.b, respectivamente), parecen ser completamente diferentes: en el mapa de calor de DTW, la mayoría de los valores están por debajo de 0.25, con algunos grupos de

a)

Paciente	28-feb			29-feb			01-mar			02-mar			03-mar			04-mar			05-mar			06-mar		
ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1222	450	450	1088	450	450	450	450	450	450	450	450	450	450	450	450	450	450	450	450	450	1077	450	450	450
1579	89	89	89	89	89	89	89	89	780	780	780	668	668	668	668	668	668	668	668	668	102	102	102	102
1583	84	84	84	84	84	84	84	84	1050	1050	1050	1050	1050	1050	647	647	647							
1585	72	72	72	72	72	392	392	392	704	704	704	704	704	704	704	704	704	704	704	704	704	704	704	44
1587	87	87	87	87	87	87	87	87	996	996	996	996	996	996	560	560	560	560	560	560	560	560	560	560
1598			86	86	86	86	86	86	516	516	516	645	645	645	645	645	645	645	645	645	645	645	645	645
1602			76	76	76	76	76	76	989	989	989	552	552	552	552	552	552	552	552	552	552	552	552	552
1604			78	78	78	78	78	78	78	78	78	440	440	440	183	183	183	183	183	183	183	183	183	183
1621						70	70	70	70	70	70	449	449	449	449	449	449	555	555					
1624						83	83	83	83	83	83	83	83	83	914	914	914	590	590	590	590	590	590	590
1626						75	75	75	75	75	75	75	75	75	498	498	498	498	498	498	498	498	498	818
1627						80	80	80	80	80	80	80	80	80	359	359	359	359	359	359	359	359	359	675
1629						88	88	88	88	88	88	88	88	88	831	831	831	609	609	609	609	609	609	609
1631						82	82	82	82	82	82	82	82	82	512	512	512	512	512	512	899	899	899	899
1632						73	73	73	73	73	73	73	73	73	840	840	840	840	840	840	608	608	608	608
1640									71	71	71	71	71	71	71	71	71	864	864	864	864	864	864	506
985	705	705	705	705	705	19	705	705	705	705	705	705	705	705	705	705	705	705	705	705	705	705	705	1073

Cama ID	Planta	Área
19	Planta0	0A
44		2A
1073		0B
1077		1B
1088		2B
102		4A

Bed ID	Floor	Área	Pasillo	Habitación
70	Planta0	3B	P0	H0
71				H1
72				
73				
75				
76				
78				
80		4B	P0	H0
82				
83				
84				H1
86				
87				
88				
89				

Cama ID	Planta	Área	Pasillo	Habitación
183	Planta1			
359				
392				

Cama ID	Planta	Área	Pasillo	Habitación
440	Planta2	0A	P1	
449			P2	H8
450		1A		
498				
506		2A	P0	
512				
516				
552		3A	P1	
555				
560				

Cama ID	Planta	Área	Pasillo	Habitación
590	Planta2	0B	P0	
608			P1	H9
609				
645		1B	P0	
647				
668		2B	P0	
675				
704		3B	P0	
705				

Cama ID	Planta	Área	Habitación
780	Planta3	1A	
818		2A	H0
831			
840		3A	
864			
899		0B	
914			
989		1B	H1
996			
1050		3B	

b)

Paciente	28-feb			29-feb			01-mar			02-mar			03-mar			04-mar			05-mar			06-mar		
ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1222	7B	7B	RAD 3	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	7B	RAD 1	7B	7B	7B
1579	EE	EE	EE	EE	EE	EE	EE	EE	14B	14B	14B	11A	11A	11A	11A	11A	11A	11A	11A	11A	UCI	UCI	UCI	UCI
1583	EE	EE	EE	EE	EE	EE	EE	EE	16C	16C	16C	16C	16C	16C	10A	10A	10A							
1585	EE	EE	EE	EE	EE	6A	6A	6A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12_CIR
1587	EE	EE	EE	EE	EE	EE	EE	EE	15E	15E	15E	15E	15E	15E	8B	8B	8B	8B	8B	8B	8B	8B	8B	8B
1598			EE	EE	EE	EE	EE	EE	7D	7D	7D	10A	10A	10A	10A	10A	10A	10A	10A	10A	10A	10A	10A	10A
1602			EE	EE	EE	EE	EE	EE	15E	15E	15E	8B	8B	8B	8B	8B	8B	8B	8B	8B	8B	8B	8B	8B
1604			EE	EE	EE	EE	EE	EE	EE	EE	EE	7A	7A	7A	2B	2B	2B	2B	2B	2B	2B	2B	2B	2B
1621						EE	EE	EE	EE	EE	EE	7B	7B	7B	7B	7B	7B	8B	8B					
1624						EE	EE	EE	EE	EE	EE	EE	EE	EE	15B	15B	15B	9A	9A	9A	9A	9A	9A	9A
1626						EE	EE	EE	EE	EE	EE	EE	EE	EE	7D	7D	7D	7D	7D	7D	7D	7D	7D	14D
1627						EE	EE	EE	EE	EE	EE	EE	EE	EE	5A	5A	5A	5A	5A	5A	5A	5A	5A	11A
1629						EE	EE	EE	EE	EE	EE	EE	EE	EE	14D	14D	14D	9B	9B	9B	9B	9B	9B	9B
1631						EE	EE	EE	EE	EE	EE	EE	EE	EE	7D	7D	7D	7D	7D	7D	15B	15B	15B	15B
1632						EE	EE	EE	EE	EE	EE	EE	EE	EE	14D	14D	14D	14D	14D	14D	9B	9B	9B	9B
1640									EE	EE	EE	EE	EE	EE	EE	EE	A&E	14F	14F	14F	14F	14F	14F	7D
985	12A	12A	12A	12A	12A	2 SURG	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	12A	RAD 0

Servicio	UH
EE	EE
UCI	UCI
RAD	RAD 0
	RAD 1
	RAD 3
2	2B
5	5A
6	6A

Servicio	UH
7	7A
	7B
	7D
8	8B
9	9A
	9B
10	10A
11	11A

Servicio	UH
12	12_CIR
	12A
14	14B
	14D
	14F
15	15B
	15E
16	16C

FIGURA 4.12. **Representación de las trayectorias de los 17 *Pacientes* infectados por *C. diff*. a)** Representación de la *parte física* de la dimensión espacial. Se muestra el ID de la *Cama* en la que cada *Paciente* ha estado en cada punto de muestreo. Las celdas de colores similares representan *Camas* que se encuentran en la misma *Habitación*, *Pasillo*, *Área* o *Planta*. Por ejemplo, las *Camas 71* y *72* están en el *Pasillo C0* del *Área 3B* de la *Planta 0*. **b)** Representación de la *parte lógica* de la dimensión espacial. Se muestra el nombre de la *UH* que atendió a cada *Paciente* en cada punto de muestreo. Las celdas de colores similares representan *UH* del mismo *Servicio*.

pacientes que tienen valores en el rango [0.45, 0.65]; mientras que en el mapa de calor de ST-DTW, hay un aumento general en la similitud en un rango aproximado de [0.15, 0.3], con la mayoría de los valores por encima de 0.4. Sin embargo, ambos mapas de calor exhiben patrones similares: los pacientes *1632* y *1579* tienen similitudes en un rango medio-alto (mayores de 0.45) con más de la mitad de los pacientes.

Basándonos en el mapa de calor de ST-DTW, los pacientes *1222*, *1626* y *1631* se destacan aún más como pacientes altamente interconectados. Los pacientes 1632 y 1629 también se presentan como el par que presenta la mayor similitud en DTW, estando presentes en ST-DTW como uno de los pares de mayor similitud junto con (*1222*, *1598*), (*1222*, *1604*), (*1579*, *1632*), (*1598*, *1626*), entre otros pares en los que también aparecen uno o dos de estos pacientes.

Para **ST-LCSS** y **ST-LCSS-WTW**, utilizamos diferentes valores para sus parámetros de entrada ε, δ y ω. En concreto, probamos que la diferencia temporal máxima entre dos puntos fuera de 2, 5 y 8 puntos de muestreo (es decir, dentro de 24, 48 y 72 horas, respectivamente), y que la distancia espacial máxima estuviera entre *Camas* de la misma *Área* o *Planta*. Para ST-LCSS-WTW, consideramos ventanas temporales de 2, 5 y 8 puntos de muestreo. Sin embargo, los resultados de todas las combinaciones probadas han sido muy similares. Los cambios más significativos (del orden de centésimas) se observan cuando permitimos emparejamientos en la misma *Planta* dentro de 48 horas (5 puntos) y, para ST-LCSS-WTW, una ventana temporal de 5 puntos. En esta sección solo presentamos los resultados para estos valores (Figuras 4.13.e y 4.13.f, respectivamente). Al usar valores más bajos, la mayor parte de las similitudes era 0. Por su parte, el uso de valores más altos requiere una cantidad de cálculos considerablemente mayor, mientras que los resultados son prácticamente idénticos a los que presentamos a continuación.

Los resultados de ST-LCSS y ST-LCSS-WTW han sido similares: la mayoría de las celdas tenían valores en el rango [0.2, 0.4], con dos grupos cuya similitud estaba en el rango [0.5, 0.65]. La principal diferencia entre ambos TDMA se encuentra en un aumento general de aproximadamente 0.1 puntos en ST-LCSS-WTW respecto a ST-LCSS. Cabe señalar que la mayoría de las TTII tienen una longitud de entre 12 y 15 puntos, mientras que el TBTP tiene 24 puntos. Si observamos los mapas de calor por filas (o columnas), podemos ver que, en general, la similitud entre cada paciente y el resto se encuentra en un rango estrecho. Sin embargo, en las filas de aquellos pacientes que presentan las similitudes más altas, podemos identificar con qué pacientes han estado más cerca y con cuáles no. Esto es diferente en los mapas de calor de DTW y ST-DTW, donde las filas (o columnas) de cada paciente presentan distintos rangos y distribución de los valores, lo que dificulta el análisis visual.

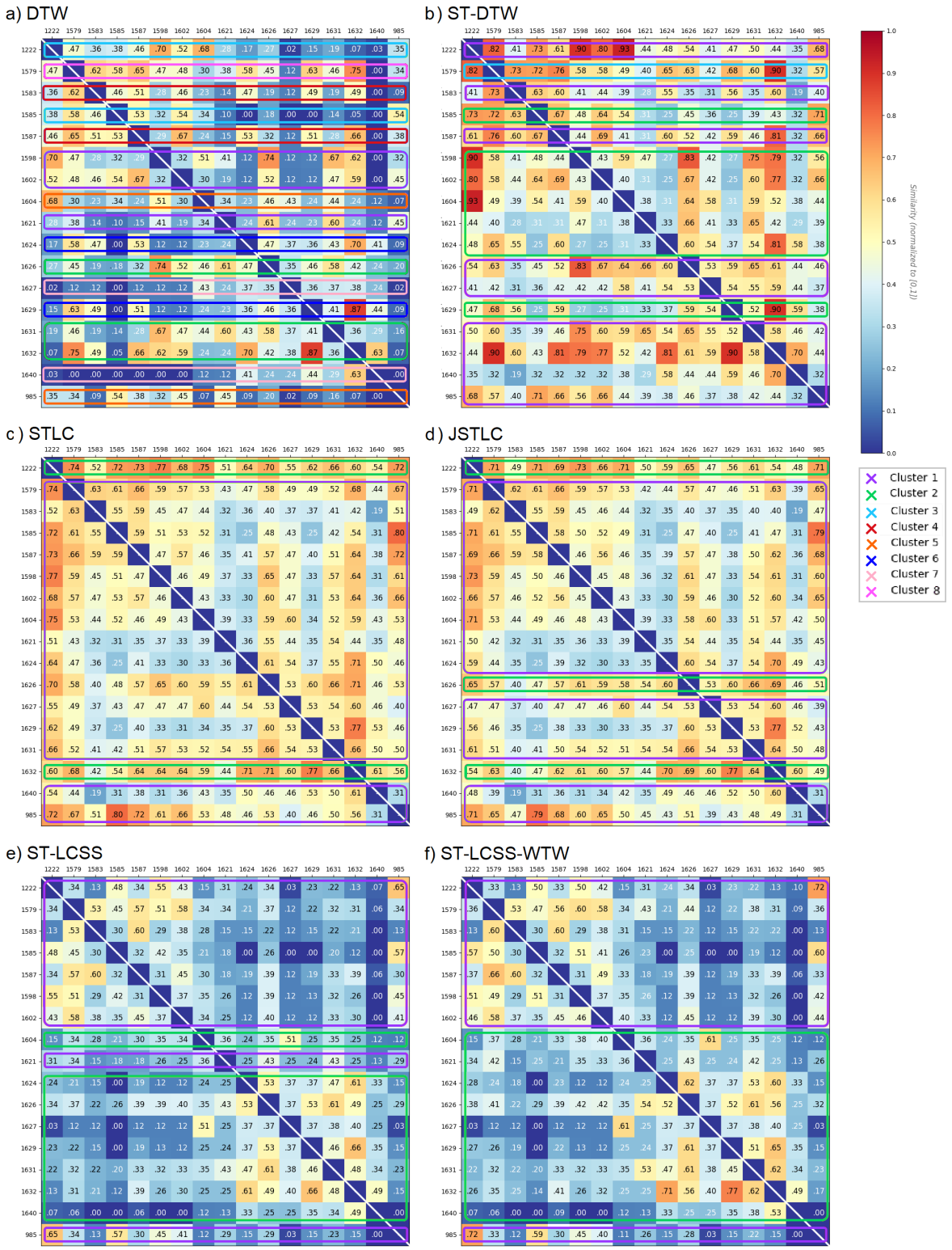


FIGURA 4.13. **Mapas de calor con los resultados de cada TDMA aplicado sobre las trayectorias de los 17 pacientes infectados por *C. diff*.** Los ID de los *Pacientes* se muestran como los índices de los mapas de calor, estando ordenados de forma alfabética. Hemos añadido una capa superpuesta para mostrar los clústeres identificados con el algoritmo k-means, de manera que las filas de los pacientes que pertenecen al mismo clúster están marcados con el mismo color.

STLC y **JSTLC** también presentan resultados similares entre sí (Figuras 4.13.c y 4.13.d, respectivamente), con valores ligeramente más bajos en JSTLC que en STLC. En ambos mapas de calor, la mayoría de los valores son medio-altos, aproximadamente en el rango $[0.4, 0.6]$, y no presentan cambios notables ni en el mapa de calor en su conjunto ni en las filas. Sin embargo, podemos identificar algunos pacientes con similitudes generales más altas (como los pacientes *985*, *1222* y *1632*) o más bajas (como los pacientes *1621*, *1624*, *1629* y *1640*). Algunos de estos pacientes son los que presentan las trayectorias más largas y cortas, respectivamente.

Una interpretación general inicial basada en los pacientes con las similitudes más altas en los seis mapas de calor sería la siguiente: Todos los pacientes comenzaron sus trayectorias en camas del *Servicio* de *A&E* en la *Planta 0*, la cual podría ser el origen de la infección. Posteriormente, pudieron surgir dos brotes que incluirían los siguientes pacientes:

- Los pacientes desde el *1222* hasta el *1602* más el paciente *985*. Este primer grupo incluye a los primeros pacientes que entraron al hospital, estando sus trayectorias solapadas en gran medida cuando estuvieron en la *Planta 0* y en algunas *Áreas* adyacentes en la *Planta 2*.
- Los pacientes desde el *1624* hasta el *1640*. Sospechamos que la *UH 7D* y *Servicio 14* están relacionados con la transmisión de la infección entre este segundo grupo de pacientes.

4.4.2.d. Paso 5: Método de *clustering*

Para agrupar a los pacientes hemos aplicado el algoritmo de *clustering* k-means sobre la MS de cada TDMA (para simplificar la lectura, nos referiremos a la MS de cada TDMA directamente con el nombre del TDMA). Para cada MS, probamos valores para k en el rango $[2,\ p]$, ejecutando el algoritmo 10 veces para cada k con diferentes inicializaciones para sus parámetros. Finalmente, para cada TDMA, hemos seleccionado el resultado con el mayor coeficiente de Silhouette. La Tabla 4.1 muestra el CS promedio de los clústeres para cada TDMA. Además, para cada clúster, muestra un conjunto de medidas cuantitativas: número de pacientes, similitud media y desviación estándar media de las similitudes de los VS en el clúster, la media de la distancia de un paciente en el clúster a cualquier otro paciente en el clúster (nos referiremos a esta medida como *distancia intra-clúster*), la media de la distancia de un paciente en el clúster a cualquier otro paciente en los clústeres más cercano y más lejano (*distancias inter-clúster mínima* y *máxima*), y el promedio del CS. Calculamos las distancias intra-clúster e inter-clúster como la distancia euclidiana entre los VS de dos pacientes.

La Figura 4.14 muestra el *clustering* de los pacientes para cada TDMA. En ella, los pacientes son representados mediante sus VS, que han sido reducidos a un espacio bidimensional mediante una técnica de *Descomposición en Valores Singulares* (en inglés, *Singular Value Decomposition* (*SVD*)) de los datos. En general, las figuras

TABLA 4.1
Clustering de los resultados de cada TDMA utilizando, para cada uno, el valor de k que proporciona el mayor Coeficiente de Silhouette (CS).

TDMA	**CS**	**Clústeres**							
		ID	**Nº Pacs**	**Similitud Media**	**Desviación Estándar de la Similitud**	***Distancia Intra-clúster* Media**	**Mínima *Distancia Inter-clúster* Media**	**Máxima *Distancia Inter-clúster* Media**	**CS Medio**
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
		Media		0.295	0.203	0.775	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
		Media		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	*Misma*	0.151
		2	2	0.596	0.144	0.870	1.190	*Misma*	0.268
		Media		0.531	0.163	0.914	1.049	*Misma*	0.210
JSTLC	0.130	1	14	0.448	0.160	0.986	1.082	*Misma*	0.089
		2	3	0.553	0.165	1.842	1.016	*Misma*	0.171
		Media		0.501	0.163	0.914	1.049	*Misma*	0.130
ST-LCSS	0.195	1	9	0.271	0.166	0.804	1.030	*Misma*	0.220
		2	8	0.268	0.168	0.856	1.030	*Misma*	0.170
		Media		0.269	0.167	0.830	1.030	*Misma*	0.195
ST-LCSS-WTW	0.188	1	8	0.273	0.172	0.872	1.084	*Misma*	0.196
		2	9	0.268	0.168	0.897	1.096	*Misma*	0.181
		Media		0.270	0.170	0.885	1.090	*Misma*	0.188

representan una nube de puntos donde es difícil distinguir cualquier grupo definido de pacientes. Con estas imágenes corroboramos visualmente que el CS promedio de los clústeres no es muy alto, como muestra la Tabla 4.1. Este se encuentra entre 0.12 y 0.21. Una interpretación [26] del CS es que un valor superior a 0.7 es "fuerte", superior a 0.5 es "razonable", y superior a 0.25 es "débil" y representa que los puntos están distribuidos uniformemente.

En **DTW**, los pacientes quedan divididos en **8 clústeres**. Se trata de clústeres pequeños (cada uno tenía entre 1 y 3 pacientes) y con una alta desviación estándar (0.2), siendo la similitud promedio de sus VS de 0.3. Hay clústeres que presentan una distancia intra-clúster baja, como los clústeres 6 y 7, y otros alta, como los clústeres 1 y 2. La mínima distancia inter-clúster promedio de cada clúster también es muy variable. En consecuencia, hay clústeres con un CS más "fuerte" que otros, dando un CS promedio de todos los clústeres de 0.25.

De manera similar a DTW, en **ST-DTW** hay algunos pacientes cuyos VS son similares entre sí, pero diferentes al resto. Sin embargo, a diferencia de DTW, en ST-DTW los pacientes han quedado agrupados en **3 clústeres**: dos de los clústeres abarcan casi la mitad de los pacientes cada uno, y hay también un clúster con solo un

paciente. La diversidad entre los pacientes de cada clúster se puede ver reflejada en su alta desviación estándar media respecto a la similitud media de sus VS (0.2 y 0.5, respectivamente), y que sus distancias intra-clúster e inter-clúster medias son similares. En consecuencia, el CS de cada clúster es bajo, con un valor promedio de 0.116 (inferior al calculado en DTW). Cabe señalar que el CS del clúster 3 es cero porque solo tenía un paciente.

En contraste con DTW y ST-DTW, en **STLC** y **JSTLC** no hay VS con valores notablemente más bajos o altos respecto a los demás. En consecuencia, los pacientes han sido agrupados en **2 clústeres**. Sin embargo, sí que hay una diferencia apreciable en el tamaño de cada clúster: tanto en STLC como en JSTLC hay un clúster que alberga casi todos los pacientes (15 y 14, respectivamente) y un clúster con el resto (2 y 3, respectivamente). Cabe señalar que, en ambos TDMA, los clústeres más pequeños están formados por aquellos pacientes con la mayor similitud global: *1222* y *1632* en STLC, y también *1626* en JSTLC. Podemos destacar que estos tres pacientes no comparten los pacientes a los que han estado más próximos, es decir, los valores más altos de sus VS no están en las mismas posiciones.

En la Tabla 4.1 podemos observar que, en ambos clústeres de cada TDMA, los valores de la distancia intra-clúster, las distancias inter-clúster mínima y máxima y la desviación estándar son similares. Solo ligeras diferencias entre las distancias intra-clúster e inter-clúster llevan a un CS promedio de 0.21 en STLC y 0.13 en JSTLC. Nótese que las matrices de similitud en STLC y JSTLC presentaban valores similares y sin cambios destacables, dando a lugar a puntos distribuidos uniformemente en las Figuras 4.14.c y 4.14.d.

En el caso de **ST-LCSS** y **ST-LCSS-WTW**, podemos destacar como rango distintivo que, a pesar de que las MS presentaban valores bajos en general, en los clústeres obtenidos es posible distinguir algunos pacientes que habrían estado próximos. En ambos TDMA, los pacientes se han dividido en **2 clústeres** que coinciden con las dos áreas cálidas mostradas en las Figuras 4.13.e y 4.13.f. El hecho de que haya algunos VS que puedan agruparse en base al valor de algunas de sus celdas también se indica con el hecho de que la desviación estándar de los clústeres es alta respecto a la media de las similitudes: en los dos clústeres de ambos TDMA, la primera tiene un valor aproximado de 0.17 y la segunda de 0.27.

Respecto al CS, tanto en ST-LCSS como en ST-LCSS-WTW, su valor medio es similar: 0.195 y 0.188, respectivamente. Podemos destacar que, con la excepción de DTW, estos son los valores más altos del CS. Éste es ligeramente inferior en ST-LCSS-WTW debido a la mayor distancia intra-clúster de sus clústeres. Esta situación podría explicarse debido a que las similitudes entre los pacientes cercanos son más altas en ST-LCSS-WTW que en ST-LCSS, mientras que las similitudes entre pacientes distantes tienen valores similares en ambos TDMA.

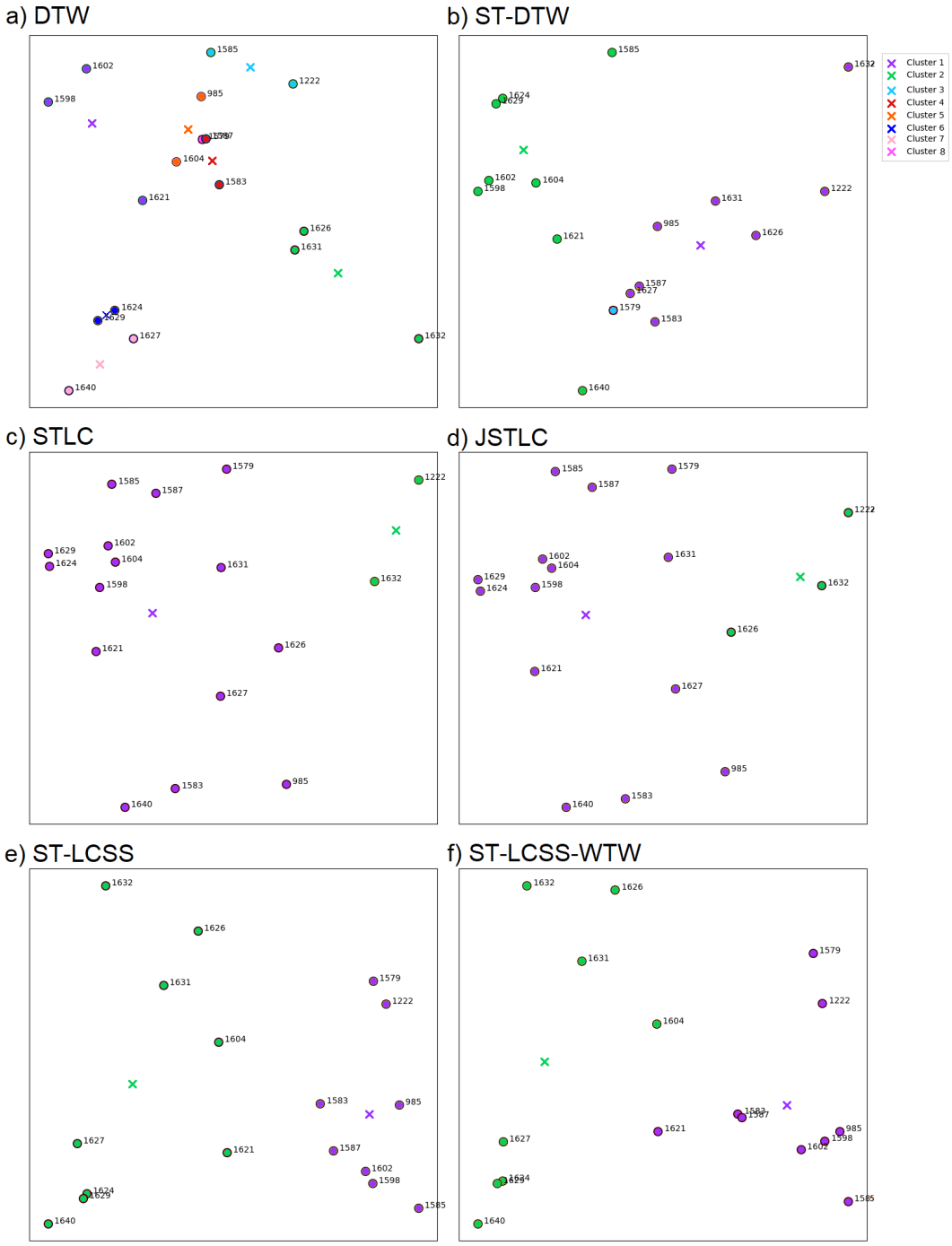


FIGURA 4.14. **Representación de los clústeres de cada TDMA en un espacio bidimensional**. Los vectores de similitud de cada paciente han sido reducidos linealmente a dos dimensiones. Los centroides de los clústeres se representan como una ×.

4.5 Discusión

En esta sección comentamos varios temas relacionados con los resultados de cada TDMA y su *clustering*.

4.5.1. DTW y ST-DTW

DTW es un algoritmo en el que debe haber un emparejamiento entre el primer punto de cada trayectoria y un emparejamiento entre el último de cada una. Además, cada uno de los puntos intermedios debe coincidir al menos una vez con otro. Estas restricciones hacen que la longitud de las trayectorias de intersección (TTII) sea un factor destacable al calcular su similitud. Aunque DTW solo evalúa la dimensión espacial, la propia definición del algoritmo lleva a que no solo el hecho de que haya camas próximas en ambas trayectorias influya en su similitud total, sino que también influye el momento en el que estas camas cercanas fueron utilizadas. Sin embargo, no existe una garantía de que solo se pueda conseguir una similitud alta debido al hecho de que en ambas trayectorias haya puntos que están alineados espacial y temporalmente.

Vamos a explicar el comportamiento de DTW mediante tres ejemplos representados en la Figura 4.15:

- En el primer ejemplo (Figura 4.15.a), las habitaciones de T_1 y T_2 están lo suficientemente cerca como para que casi todos sus puntos coincidan uno a uno.
- El segundo ejemplo (Figura 4.15.b) muestra un caso límite donde T_1^1 y T_2^1 ocurren en la misma habitación, quedándose p_1 en la misma habitación hasta el penúltimo punto. Por su parte, p_2 se cambia de habitación durante toda su trayectoria, coincidiendo con p_1 en su último punto. Como consecuencia, hay coincidencias entre T_2^1 y todos los puntos de T_1, excepto el último, y entre el último punto de T_1 y todos los puntos de T_2, excepto el primero.
- El tercer ejemplo (Figura 4.15.c) es una versión del segundo ejemplo en el que T_1^2 y T_2^2 están en habitaciones cercanas, pero alejadas de T_1^1 y T_2^1. Además, los puntos intermedios de ambas trayectorias ocurren en habitaciones con una distancia que podríamos definir como "intermedia". Con estas modificaciones, el número de coincidencias y su similitud total son más bajos que en Figura 4.15.b.

En los dos primeros casos, la similitud sería alta. Sin embargo, en el segundo ejemplo hay más emparejamientos, por lo que se devolvería una similitud mayor que en el primero. En nuestro experimento podemos detectar una situación similar a la mostrada en el segundo ejemplo entre los pacientes *1579* y *1632*. En el caso del tercer ejemplo, su similitud sería la menor de los tres casos mostrados: hay un menor número de emparejamientos y la similitud entre los puntos que los forman es menor.

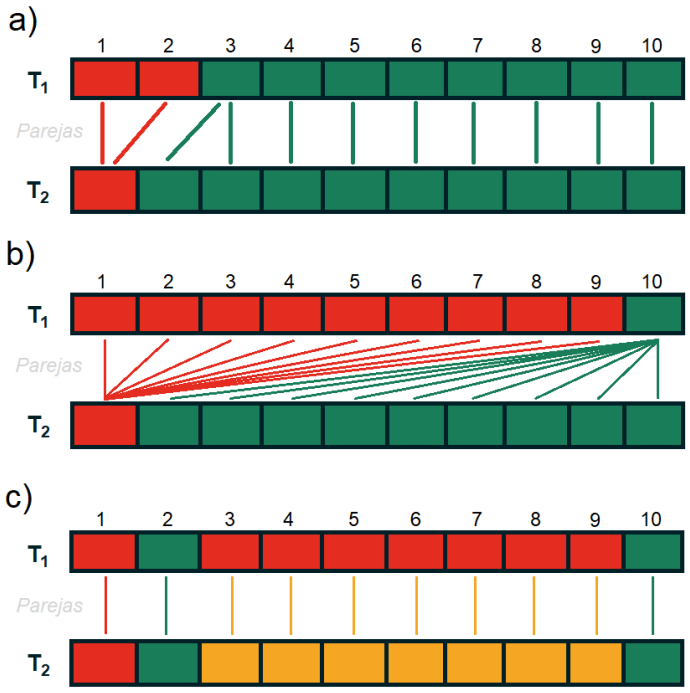


FIGURA 4.15. **Representación de tres casos donde se utiliza DTW para evaluar la similitud epidemiológica espacio-temporal entre dos trayectorias,** T_1 **y** T_2. Cada trayectoria está representada por una línea temporal en la que el color de las celdas representa la *Habitación* en la que ha estado el paciente. Vamos a suponer que la distancia espacial entre dos celdas del mismo color es la misma y la más baja. La distancia entre celdas naranjas y verdes es la misma que entre celdas naranjas y rojas. La distancia entre celdas verdes y rojas es mayor que las otras dos. Los emparejamientos se representan mediante una línea que une los dos puntos que los forman.

ST-DTW sigue el mismo esquema que DTW, pero también evalúa la dimensión temporal además de la espacial. Si continuamos con los ejemplos de la Figura 4.15, nos encontraríamos con los siguientes resultados:

- El resultado del primer ejemplo no se vería muy afectado, pues los emparejamientos seguirían siendo los mismos. La similitud final se vería aumentada o reducida en función del valor del parámetro α.
- En el segundo ejemplo, debido a inclusión de la dimensión temporal en el cálculo de la similitud, habrá un punto en el cual los puntos intermedios de T_1 dejen de emparejarse con T_2^1. Por tanto, el número de emparejamientos se verá reducido y, con ello, la similitud total.
- En el tercer ejemplo, consideramos que los emparejamientos serían los mismos que al utilizar DTW, pero la inclusión de la similitud temporal aumentaría la similitud espacio-temporal entre puntos y, en consecuencia, la total.

En DTW, los puntos que están lejos espacialmente pero cerca temporalmente tienen poco impacto, salvo en casos similares a los del segundo ejemplo, en los que muchos emparejamientos con una baja similitud dan lugar a una similitud final alta. En el caso de ST-DTW, los emparejamientos entre puntos lejanos en una dimensión y cercanos en la otra sí que pueden presentar una similitud mayor (dependiendo de α). Esto explicaría el aumento general en las similitudes calculadas con ST-DTW, siendo el incremento mayor cuanto más largas son las TTII. Por ejemplo, en nuestro experimento, la similitud entre los pacientes *1222* y *1579* es 0.4 puntos mayor en ST-DTW que en DTW.

A modo de conclusión, DTW y ST-DTW son dos algoritmos con fuertes restricciones respecto al orden en que los puntos pueden emparejarse, pero no en cuanto a la distancia máxima (o similitud mínima) que debe haber entre ellos para permitir el emparejamiento. En consecuencia, la MS está formada por muchos pares de pacientes con alta similitud entre sí y baja similitud con el resto, lo que provoca la aparición de pequeños clústeres, especialmente en DTW.

4.5.2. STLC y JSTLC

STLC y JSTLC presentan un esquema en el que cada punto de cada trayectoria debe emparejarse al menos una vez con un punto de la otra trayectoria, sin depender un emparejamiento de los anteriores. Esto implica que, cuanto más largas sean las trayectorias, mayor será la similitud. Por ejemplo, los pacientes *1222*, *1579*, *1585* y *985* presentan algunas de las trayectorias más largas y similitudes más altas entre sí y con otros pacientes. En el caso contrario se encuentran los pacientes *1621* y *1640*. Cabe señalar que no consideramos esto un problema: cuanto más tiempo y espacio compartan dos personas, mayor será la probabilidad de que ocurra una infección. Sin embargo, el hecho de que la longitud de las trayectorias influencie la similitud final puede tener un efecto negativo en los resultados generales.

En **STLC**, para cada punto, buscamos por separado los puntos más cercanos espacial y temporalmente. En la práctica, el punto más cercano temporalmente es siempre el que tiene la misma marca de tiempo. Dado que las TTII entre dos trayectorias tienen el mismo número de puntos y las mismas fechas de inicio y fin, la similitud temporal es una medida de la longitud de las TTII: a mayor longitud de la TTII, mayor será la "aportación" de la dimensión temporal. En la práctica, solo se evalúa la similitud espacial. La falta de restricciones en la selección del punto más cercano espacialmente llevaría a que, en los ejemplos mostrados en la Figura 4.15, $sim_{STLC}(T_1, T_2)$ fuese alta y con el mismo valor en el primer y segundo casos. En el tercer ejemplo, como $sim_{STLC}(T_1, T_2)$ se calcula como el promedio de las similitudes de cada trayectoria con la otra, habría una compensación entre la alta similitud de T_1 y T_2 y la baja similitud de T_2 y T_1.

Por el contrario, **JSTLC** busca para cada punto su punto más cercano espacio-temporalmente. Por tanto, en base al parámetro βA, se busca un equilibrio entre ambas dimensiones, de manera que disminuya la proporción de puntos cuyos

emparejamientos tengan una similitud cercana a la máxima. Esto podemos observarlo en la Figura 4.13.d, donde los valores de la MS son, en general, ligeramente inferiores a los de la MS de la Figura 4.13.c. Si consideramos que todas las TTII tienen longitudes similares, las similitudes más altas se obtendrían al comparar aquellas TTII que comparten una alta proporción de puntos cercanos en el tiempo y el espacio. Sin embargo, habrá emparejamientos cuyos puntos solo estén cerca en una de las dos dimensiones y, al igual que en ST-DTW, la suma de varios puntos en los que una de sus similitudes es máxima acaba resultando en una similitud total en un rango medio-alto.

A modo de resumen, podemos encontrar tanto en STLC como en JSTLC dos hechos que conducen a que la similitud general de las MS sea media-alta: el requerimiento de que todos los puntos tengan al menos un emparejamiento y la falta de límites para definir la similitud mínima que se debe cumplir en cada dimensión para permitir un emparejamiento. En nuestro experimento, dado que no hay diferencias marcadas entre los vectores de similitud de los pacientes, la mayoría de ellos han sido agrupados en el mismo clúster, separando a aquellos que difieren del resto (no necesariamente son parecidos entre sí) en otros clústeres. Este hecho se refleja en que los clústeres obtenidos en STLC y JSTLC presentan las distancias intra-clúster más altas.

4.5.3. ST-LCSS y ST-LCSS-WTW

A diferencia de los otros TDMA, ST-LCSS y ST-LCSS-WTW implementan dos parámetros, ε y δ, para limitar espacial y temporalmente el emparejamiento entre puntos. Los valores de estos parámetros son determinantes pues la efectividad del algoritmo depende en gran medida de ellos. Para su elección, deberían tenerse en cuenta factores como la estructura física del hospital y cómo la literatura científica caracteriza la transmisión de la infección.

Si consideramos la dimensión espacial, podríamos aprovechar la jerarquía de *Localizaciones* del modelo de datos para ajustar la evaluación de la similitud entre las trayectorias sin necesidad de cambiar los valores de la propiedad *coste*. Por ejemplo, se podría hacer una primera comparativa permitiendo los emparejamientos hasta un nivel alto de la jerarquía, como *Planta*. La comparativa permitiendo como máximo un nivel bajo de la jerarquía puede dar como resultado que la mayoría de las similitudes de la MS sean cero. Por tanto, proponemos empezar con un nivel alto para que no se quede un posible brote sin detectar y, en caso de que lo haya, realizar una comparación más precisa restringiendo espacialmente los emparejamientos a niveles inferiores como *ZonaLógica*, *Área* o *Servicio*.

Cabe recordar que además de la jerarquía espacial, el modelo de datos también cuenta con otros tipos de relaciones de vecindad espacial, como *contiguo* y *opuesto*. Además de permitir una mayor precisión al calcular la distancia entre *Camas*, podemos utilizarlas para representar diferentes niveles de proximidad entre puntos dentro de la misma *Planta*.

Con respecto a los resultados de ST-LCSS y ST-LCSS-WTW, podemos observar en las Figuras 4.13.c y 4.13.d que, en general, la similitud entre pacientes presenta valores bajos o medios. Si tenemos en cuenta el uso de los parámetros ε y γ y el no requerimiento de que todos los puntos deban tener un emparejamiento en la solución final, podemos tener una confianza en que una similitud alta o baja entre dos pacientes está relacionada con que estos hayan compartido o no una *Localización* cercana en un tiempo próximo (dentro de los límites establecidos). En ST-LCSS se restringe el hacer un emparejamiento con puntos que pertenezcan a otro emparejamiento o que estén antes de otro punto de su misma trayectoria que ya haya sido emparejado. En ST-LCSS-WTW intentamos aliviar la segunda restricción, obteniendo un ligero aumento en las similitudes entre algunos pacientes, especialmente entre aquellos que ya tenían una mayor similitud que los demás.

Un aumento localizado en el valor de las similitudes beneficia tanto el análisis por parte de personas de los resultados como la aplicación del algoritmo de *clustering*. ST-LCSS-WTW presenta los clústeres con la distancia intra-clúster más baja (exceptuando DTW) y una distancia inter-clúster similar a la obtenida con el resto de TDMA. En consecuencia, el *clustering* sobre la MS de ST-LCSS-WTW tiene uno de los valores más altos de CS (más alto que el de ST-LCSS). Además, los clústeres en ST-LCSS-WTW y ST-LCSS son los más similares a los grupos de pacientes observados en nuestra interpretación visual de las MS.

4.5.4. Otros aspectos relacionados con los TDMA

Hay algunos aspectos que, aunque no estén directamente relacionados con el funcionamiento de los TDMA o sus resultados, cabe señalar.

Desde una perspectiva del coste computacional, ST-LCSS-WTW es el algoritmo más costoso, ya que tiene una ejecución ramificada que implica un orden de complejidad exponencial en función de la longitud de las trayectorias. Este nivel de ramificación depende en gran medida de ω y de la frecuencia de muestreo utilizada para crear las trayectorias. En contraste, STLC y JSTLC son algoritmos con un orden de complejidad cuadrático. En el caso de DTW (y, por extensión, también podríamos aplicar a ST-DTW), a pesar de su definición recursiva, se han estudiado varias optimizaciones para su implementación. Podemos destacar su implementación basada en programación dinámica [140], que presenta un orden de complejidad cuadrático.

Otro aspecto por discutir sería la facilidad de modificar la implementación del algoritmo para su transformación en un algoritmo compatible con la computación paralela o para añadir nuevas condiciones con las que guiar la búsqueda de los emparejamientos óptimos, como establecer límites en la similitud espacial y la temporal. En este sentido, JSTLC es un algoritmo cuya simplicidad puede facilitar estas tareas. Por ejemplo, las búsquedas del punto más cercano a cada punto de T_1 y de T_2 son independientes, por lo que podrían ejecutarse en paralelo. Además, al calcular la similitud entre dos puntos, podríamos añadir los parámetros ε y δ para comprobar

si las similitudes espaciales y temporales están por encima del mínimo definido y, si no lo están, entonces definir $sim_{ST}(p_1, p_2) = 0$. Sería necesario estudiar si se puede aplicar una modificación similar a DTW y ST-DTW.

Con respecto a esta posible versión modificada de JSTLC con restricciones espaciales y temporales, consideramos que sus resultados serían similares a los de ST-LCSS-WTW, ya que ambos algoritmos buscan maximizar los emparejamientos entre los puntos y su similitud total. Una diferencia relevante entre los dos algoritmos sería el número de coincidencias que un punto puede tener. En ST-LCSS-WTW, un punto puede coincidir, como máximo, con un punto de la otra trayectoria. Mientras que en JSTLC, un punto puede ser el más cercano a varios puntos de la otra trayectoria y, por tanto, formar parte de varios emparejamientos.

Ponderar una alternativa por encima de la otra depende de si consideramos que la probabilidad de transmisión solo aumenta con una estancia cercana y prolongada de ambos pacientes (ST-LCSS-WTW) o si también queremos tener en cuenta situaciones en las que el paciente susceptible haya podido estar en un entorno posiblemente contaminado (aire, muebles y equipamiento médico, personal sanitario), no necesariamente estando al mismo tiempo con el paciente infeccioso (JSTLC modificado). En esta decisión, sería esencial tener en cuenta las características del patógeno.

Como último aspecto a considerar y, aunque no hemos realizado una comparación explícita, consideramos que la interpretabilidad humana de los resultados de cada TDMA también es un factor relevante, especialmente si queremos que StESPT sea una herramienta de apoyo para los epidemiólogos en la detección de brotes entre pacientes. Aparte de la similitud, algunos TDMA devuelven información adicional que puede servir para este propósito. ST-LCSS y ST-LCSS-WTW devuelven el número de emparejamientos entre las trayectorias comparadas, y también podrían devolver cuáles son estos emparejamientos y su similitud. Esto podría ayudar a determinar las fuentes de transmisión con mayor precisión. De manera similar, STLC y JSTLC podrían devolver, para cada trayectoria, cuáles son los emparejamientos entre sus puntos y los más cercanos en la otra trayectoria. En el caso de DTW y ST-DTW, no sería posible obtener este tipo de información.

4.5.5. Aspectos relacionados con el tiempo

En esta sección discutimos aspectos relacionados con la definición de las trayectorias de intersección.

Al evaluar la similitud entre dos trayectorias, T_1 y T_2, definimos las trayectorias de intersección entre ellas y comenzamos a buscar emparejamientos desde el primer hasta el último punto de cada TI. Sin embargo, podríamos proponer que se pudiesen emparejar puntos dentro de una TI con puntos que estén fuera de la otra TI, pero dentro del TBTP. Vamos a utilizar un ejemplo para ilustrar una situación en el que

esta comparación podría ser beneficiosa. Suponemos dos pacientes, p_1 y p_2, cuyas hospitalizaciones comienzan en diferentes momentos, siendo p_1 el primero en entrar al hospital. Por tanto, T_1 tiene algunos puntos antes del inicio de las TTII entre ambos pacientes (consideramos que T_1 y T_2 están dentro del TBTP). p_2 ingresa al hospital en la misma habitación en la que estaba p_1, pero en ese momento p_1 se traslada a otra habitación. Si p_1 hubiese estado infectado (e infeccioso), el entorno en el que se encuentra p_2 podría estar contaminado. Sin embargo, esta situación no se considera cuando se evalúa la similitud entre sus TTII. Otro caso similar podría ocurrir si la hospitalización de p_1 hubiese terminado antes que la de p_2, y, durante el tiempo en que p_1 ya no estaba en el hospital, p_2 hubiese estado en una *Localización* cercana a la última *Localización* de p_1.

Podríamos establecer un margen para definir la diferencia máxima permitida en el tiempo entre dos puntos cuando uno está dentro de una TI y el otro solo dentro del TBTP. Implementar esta modificación sería sencillo en STLC y JSTLC, pues los emparejamientos entre puntos son independientes. Para el resto de TDMA, la implementación de esta modificación sería más compleja. Por ejemplo, los emparejamientos (T_1^1, T_2^1) y (T_1^m, T_2^m) son obligatorios en DTW y ST-DTW, respectivamente. Dado que los beneficios de esta modificación son situacionales, sería necesario analizar la conveniencia de su aplicación en función de las características de la infección.

Otro aspecto para discutir es la definición del punto final de las TTII entre T_1 y T_2. Este se define en función del final del TBTP o del último punto de una de las trayectorias, dependiendo de cuál ocurra primero. Sin embargo, el punto final de las TTII podría plantearse como el punto de muestreo en el que ocurre el primer *TestMicro* en T_1 o T_2. Esta opción sería apropiada para una investigación epidemiológica centrada en identificar qué pacientes tienen más probabilidades de haber transmitido la infección a un paciente en particular. En nuestro caso, queremos que el marco de estudio de nuestro método sea más amplio, cubriendo las relaciones espacio-temporales entre pacientes más allá del contagio, ya que estas pueden ser clave para determinar dónde se encuentran las fuentes del brote. Por ejemplo, si un paciente tenía un *TestMicro* positivo mientras estaba en una habitación, pero luego fue trasladado a otra, los pacientes cercanos a ambas habitaciones podrían estar en riesgo de contagiarse.

4.5.6. Método de *clustering*

Los agrupamientos de los pacientes al aplicar el algoritmo de *clustering* k-means han sido muy variados en función del TDMA con el que hemos calculado la MS de entrada. Sin embargo, las seis soluciones presentan coeficientes de Silhouette similares, dentro del rango de $[0.13,\ 0.21]$.

K-means es un algoritmo que crea k clústeres minimizando la suma de las distancias euclidianas entre cada paciente del clúster y su centroide. Esto puede ser una desventaja cuando tenemos un conjunto de datos de entrada en el que no hay diferencias

significativas entre sus elementos, como es el caso de las MS en nuestro experimento. Esta situación puede deberse porque la mayoría de los datos comparten valores similares en cada una de las dimensiones que los definen (como en STLC y JSTLC) o porque existen pequeños grupos de datos que son muy similares entre sí pero que no están significativamente diferenciados del resto (como en DTW y ST-DTW).

En la búsqueda de un equilibrio entre la cohesión y la separación de los clústeres, los pacientes pueden ser divididos en pocos clústeres que contengan a la mayoría de los pacientes o en muchos clústeres pequeños que los agrupen en base a un grano más fino. Cabe destacar que el *clustering* en DTW con 8 clústeres presenta el CS más alto. Sin embargo, si consideramos que el TBTP es de ocho días, que son 17 los pacientes a agrupar en clústeres, y que éstos estuvieron en ubicaciones cercanas, un número menor de clústeres sería más razonable.

En el caso de STLC y JSTLC, podemos interpretar el hecho de que el algoritmo de *clustering* haya separado en un clúster diferente a los pacientes que presentaban las similitudes más altas (cuando la mayor parte de la MS están en un rango medio-alto) como una señal de que estos algoritmos son más específicos para representar altas similitudes, enfocándose en aquellos pacientes que están particularmente cercanos. Sin embargo, como se muestra en las Figuras 4.13.c y 4.13.d, las similitudes entre los pacientes que forman estos pequeños clústeres (pacientes *1222*, *1632* y *1626*) no son las más altas en cada VS. Además, si nos fijamos en la Figura 4.12, sus trayectorias no parecen representar un grupo de pacientes que hayan estado especialmente cercanos durante gran parte del TBTP. Por tanto, basándonos en nuestro experimento, no podemos asegurar que STLC y JSTLC puedan funcionar como una "lupa" para resaltar a los pacientes con una relación muy estrecha.

Si solo consideramos las medidas cuantitativas presentadas en la Sección 4.4.2.d, no podríamos determinar si los clústeres obtenidos en ST-LCSS y ST-LCSS-WTW son mejores que los obtenidos en los otros TDMA. Sin embargo, desde una perspectiva visual, podemos notar que los clústeres obtenidos en ST-LCSS y ST-LCSS-WTW son los más similares a los dos grupos de pacientes que, en el paso 4, conjeturamos que podrían estar espacio-temporalmente cerca. Por lo tanto, basándonos en los resultados de ST-LCSS y ST-LCSS-WTW, concluimos que la fuente de la infección por *C. diff* habría sido el *Servicio* de *A&E* en la *Planta 0*, y que, posteriormente, ocurrieron dos brotes en la *Planta 2*.

Un último aspecto para tener en cuenta sería que, en todos los TDMA, los clústeres tienen un CS medio de 0.2, el cual puede interpretarse como bajo o "débil". Sin embargo, hay estudios en los que se ha comprobado que un aumento en la dimensionalidad de los datos de entrada dificulta la obtención de valores altos de CS, ya que las distancias promedio entre los elementos se vuelven similares: hay un aumento del espacio que define los elementos a agrupar, por lo que parecen estar dispersos, impidiendo la detección eficiente de áreas con propiedades similares [26]. Cabe destacar que, en nuestro experimento, cada paciente está representado por un vector de 17 dimensiones (los vectores de similitud tienen tantas dimensiones como pacientes a agrupar). Además, hemos normalizado las similitudes espaciales y temporales entre

dos puntos a $[0, 1]$ desde un rango más amplio.

4.6 Conclusiones

Este capítulo tenía como objetivo desarrollar una herramienta software de apoyo en la investigación de las relaciones epidemiológicas entre pacientes hospitalizados que se han contagiado de una infección nosocomial bacteriana. Con este fin, diseñamos el método StESPT, en el que definimos un conjunto de cinco pasos para la consecución de dos objetivos principales: cuantificar cómo han estado espacio-temporalmente conectados los pacientes infectados durante su estancia en el hospital y, en base a este resultado, determinar si estos pacientes forman parte de uno o varios brotes y cuáles han sido las posibles rutas de transmisión. Estos pasos consisten en obtener los pacientes infectados y sus trayectorias, evaluar la similitud entre dos trayectorias de pacientes basándonos en una ecuación que combina linealmente la similitud espacial y la similitud epidemiológica-temporal entre sus puntos, y agrupar a los pacientes en función de sus similitudes mediante un algoritmo de *clustering* (concretamente, k-means).

Diseñamos un experimento en el que se representa cómo un epidemiólogo podría utilizar StESPT para analizar las trayectorias de varios pacientes infectados con Clostridium difficile en un hospital. El uso de StESPT nos ha permitido determinar el origen de la infección y la existencia de dos brotes.

Para calcular la similitud entre dos trayectorias, modificamos tres TDMA de propósito general (DTW, STLC y ST-LCSS) y creamos una extensión de cada uno de ellos con cambios más significativos con el fin de adaptarlos mejor a la semántica del problema.

Cada par de algoritmos presenta características distintivas que influyen en sus resultados. En DTW y ST-DTW, todos los puntos de ambas trayectorias deben coincidir, estando cada nuevo emparejamiento condicionado por los anteriores. Con este par también comprobamos los resultados proporcionados al evaluar únicamente la dimensión espacial. En ST-LCSS y ST-LCSS-WTW, no es necesario que todos los puntos coincidan, ya que definimos una similitud espacial y temporal mínimas para determinar cuándo es posible un emparejamiento. Sin embargo, en ST-LCSS sigue existiendo cierta dependencia de los nuevos emparejamientos respecto a los anteriores, restricción que intentamos atenuar en ST-LCSS-WTW con una ejecución más ramificada y la inclusión de una ventana temporal para permitir la exploración de varias soluciones en cada paso de la ejecución del algoritmo. Por último, en STLC y JSTLC, los emparejamientos entre puntos son independientes. Como en DTW y ST-DTW, todos los puntos deben formar parte de algún emparejamiento y, a diferencia de ST-LCSS y ST-LCSS-WTW, no tiene definidos unos límites en las similitudes espacial y temporal entre los puntos de los emparejamientos.

El modo en que cada TDMA refleja las relaciones espacio-temporales entre cada par de trayectorias tiene una gran influencia en el resultado global de las similitudes

entre los pacientes y el agrupamiento de estos en el algoritmo de *clustering*.

Observamos que evaluar la proximidad entre los pacientes combinando información espacial y temporal proporciona mejores resultados que evaluar únicamente la dimensión espacial (DTW y, en la práctica, STLC). También identificamos dos casos diferentes cuando existe el requerimiento de emparejamiento de todos los puntos: imponer restricciones que hagan que las nuevas coincidencias dependan de las anteriores puede generar un aumento excesivo de la similitud entre trayectorias en algunos casos específicos (DTW y ST-DTW), mientras que la ausencia total de estas restricciones puede conducir a una similitud media-alta de forma general entre todos los pacientes en la que no se presentan variaciones significativas (STLC y JSTLC).

Establecer límites espaciales y temporales para definir cuándo permitimos un emparejamiento y, en consecuencia, no forzar el emparejamiento de todos los puntos de ambas trayectorias (ST-LCSS y ST-LCSS-WTW) nos ha permitido representar más fielmente las relaciones espacio-temporales entre las trayectorias. Al comparar ST-LCSS y ST-LCSS-WTW, el mayor grado de independencia para el emparejamiento de puntos en ST-LCSS-WTW permite una mayor precisión en la determinación de qué puntos están más cerca espacio-temporalmente. Estos límites espaciales y temporales también permiten evaluar las trayectorias en distintos niveles de precisión. Por ejemplo, los límites espaciales pueden apoyarse en un modelo flexible y multinivel para representar la dimensión espacial, como el propuesto en el Capítulo 3.

Para concluir, en k-means, para lograr un CS alto, los datos no deben estar uniformemente dispersos (como en STLC y JSTLC) ni contener algunos datos con características muy similares (DTW y ST-DTW). Aunque con un CS promedio similar al resto, al aplicar k-means sobre los resultados de ST-LCSS y ST-LCSS-WTW podemos encontrar los clústeres de pacientes en los que se refleja mejor lo ocurrido en base a las relaciones espacio-temporales entre los pacientes.

4.7 *Open-science*

Para el desarrollo de este capítulo hemos implementado diferentes herramientas de software y conjuntos de datos, cuyo acceso es público a través de repositorios en GitHub.

- Implementación del método StESPT y el conjunto de datos utilizados en el experimento de la Sección 4.4 como un KG en formato RDF*:

  `https://github.com/LorenaPujante/StESPT`

- Software para adaptar la salida del modelo de simulación hospitalario [98] a un grafo de conocimiento en formato RDF y RDF* que siga nuestro modelo de datos: `https://github.com/LorenaPujante/HospitalGeneratorRDF_V2`

Capítulo 5

Conclusiones

5.1 Conclusiones

La hipótesis planteada en esta tesis doctoral fue la demostración de que es posible combinar de forma efectiva las tecnologías basadas en grafos con el análisis espacio-temporal para la resolución de tareas epidemiológicas en el contexto de la vigilancia de infecciones nosocomiales causadas por bacterias multirresistentes.

Para abordar la hipótesis, propusimos como objetivo general el diseño de un modelo de datos y un conjunto de problemas clínicos epidemiológicos, una formalización mediante un enfoque basado en grafos que pudiese servir de base para la futura investigación epidemiológica de infecciones nosocomiales basada en el análisis espacial y temporal de los movimientos y contactos de pacientes en un hospital. En consecuencia, describimos siete objetivos específicos (ver Sección 1.2).

Tras la realización de la investigación aquí mostrada, las principales conclusiones de esta tesis doctoral en relación con los objetivos propuestos son:

- **Objetivo 1**: *Análisis bibliográfico y estudio del uso de grafos en los modelos computacionales para la simulación de brotes epidémicos.*
  - Encontramos un interés predominante en la simulación de enfermedades infecciosas virales, como la gripe estacional o el COVID-19, que suelen tener un periodo de latencia y en donde el individuo queda inmunizado hasta el final de la simulación.
  - Para representar la historia natural de la infección se utilizan modelos epidemiológicos compartimentales, destacando el uso de SEIR y SIR, así como con variaciones de estos para adaptarlos a escenarios más complejos y análisis de otros aspectos relacionados con el brote.
  - Hemos identificado un creciente interés en el uso de redes como apoyo para la simulación de los brotes, siendo utilizadas como bases de conocimiento de las relaciones físicas o sociales entre los de individuos sobre los que se simula la propagación del brote epidémico.

- Desde un punto de vista semántico, las redes se pueden clasificar en: redes de contactos, redes de relaciones, redes multicapa y redes de metapoblaciones. Las más utilizadas son las redes de contactos y las redes de relaciones.
- Desde un punto de vista topológico, hemos encontrado una preferencia en el uso de redes complejas ya que cuentan con varios parámetros ajustables que facilitan la representación de conexiones sociales como núcleos familiares y grupos de amigos. Salvo en el caso de las redes de contactos, las redes son estáticas, es decir, sus aristas y nodos no cambian durante la simulación.
- Hemos identificado el uso habitual de tres modelos computacionales para la simulación de brotes infecciosos: modelo determinista, modelo estocástico y modelo basado en agentes.
- Desde el punto de vista espacial, la tendencia mayoritaria consiste en simulaciones en áreas extensas como ciudades (se centran en las relaciones sociales y movimiento de las personas entre edificios) y regiones y países (centradas en el movimiento colectivo de las personas entre localidades). La simulación de brotes dentro de edificios es minoritaria, destacando la simulación de brotes nosocomiales. También existe una tendencia a no definir un espacio de simulación concreto, centrándose el estudio en otros aspectos relacionados con el brote, como el comportamiento social o las medidas preventivas y de control.
- Hemos encontrado una relación en el uso de los modelos computacionales, el tipo semántico de red y la escala espacial de las simulaciones.
  - En espacios interiores, se usan modelos estocásticos con redes de contactos dinámicas (en cada paso, cambian las aristas que unen los nodos que se mantienen fijos).
  - En entornos urbanos, se exploran múltiples combinaciones, destacando el uso de modelos basados en agentes para el rastreo detallado de contactos en escalas de tiempo reducidas.
  - En escalas mayores, como una región o país, destaca el uso de redes de metapoblaciones en combinación con un modelo estocástico.
  - Cuando las simulaciones se realizan sin modelar un espacio concreto, el tipo de red más utilizado es la red de relaciones que se combina con un modelo determinista o estocástico.
- La unidad temporal de las simulaciones rara vez se especifica claramente. En los pocos casos donde se menciona, la unidad temporal más común es de un día o fracciones de día.
- En cuanto al uso de datos, se detecta una carencia en el uso de datos individualizados con los que registrar los movimientos detallados o la posición de personas. En términos generales, se utilizan datos agregados (como censos) para definir las características de la población sobre las que se basa la topología de la red.

- En el ámbito hospitalario, hemos encontrado dificultades en el acceso a fuentes de datos relevantes (historia clínica electrónica, sistema de información hospitalario), así como una escasez de modelos de simulación de los movimientos de los pacientes y trabajadores sanitarios dentro del hospital.

- **Objetivo 2**: *Modelado espacio-temporal para el análisis de contactos entre pacientes orientado a la transmisión epidemiológica de infecciones nosocomiales.*
  - Constatamos la necesidad de una representación diferenciada pero interconectada de las dimensiones espacial y temporal para modelar con precisión los movimientos de los pacientes dentro del hospital y su relación con la transmisión de infecciones.
  - En cuanto a la dimensión espacial, la integración de tanto aspectos físicos (*Localizaciones*) como lógicos (*Servicios* y *Unidades*) del hospital, la estructura jerárquica de la parte física así como la introducción de relaciones horizontales entre las clases que la componen resultaron claves para identificar rutas de contagio no evidentes y permitieron evaluar de manera más rica la proximidad entre pacientes, más allá de la mera co-localización física.
  - Encontramos útil la inclusión en el modelo de la propiedad *coste* en las relaciones entre las clases que forman la dimensión espacial para permitir cuantificar la proximidad entre pacientes. Esta cuantificación se puede utilizar de base para la posterior aplicación de distintas técnicas para analizar los contactos entre pacientes.
  - En cuanto a la dimensión temporal, la formalización de los eventos asociados a cada paciente (desde movimientos hasta pruebas diagnósticas) ofreció un marco versátil para analizar la coincidencia en tiempo y espacio entre pacientes, y así detectar posibles contactos relevantes.
  - La definición flexible de contacto entre pacientes, basada en el modelo espacio-temporal, permitió identificar diversos escenarios plausibles de transmisión.
  - En conjunto, el modelo contribuyó a una comprensión más detallada de las dinámicas de contacto en entornos hospitalarios, ofreciendo una base sólida para la vigilancia y el análisis epidemiológico de infecciones nosocomiales.

- **Objetivo 3**: *Diseño y formalización de las tareas epidemiológicas fundamentales en la vigilancia y detección de brotes nosocomiales.*
  - El análisis epidemiológico necesita de diversas tareas cuya complejidad requiere que en formalización sean descompuestas en varias consultas como la detección de brotes en una ubicación física (*Localización*) o lógica (*Servicio*), el análisis de contactos entre un conjunto de pacientes definido o de un paciente con cualquier otro paciente del hospital que cumpla unos requisitos epidemiológicos y espacio-temporales, y la detección del paciente índice.
  - Comprobamos la idoneidad del uso encadenado de las consultas para la identificación de posibles fuentes de contagio (localización o servicio) cadenas de transmisión entre pacientes de la infección.
  - Constatamos que la modelización en formato de grafo no solo simplifica la construcción de consultas en las cuales es preciso el recorrido a través de datos altamente conectados, sino que la representación visual de los resultados en forma de grafo facilita la interpretación y análisis de estos por partes del personal de epidemiología del hospital. Por ejemplo, es factible la identificación de cúmulos de pacientes en torno a unos pocos nodos pertenecientes a la dimensión espacial, o el seguimiento de la trayectoria cronológica seguida por varios pacientes y detección del nexo entre ellas.
  - Probamos que el establecimiento de una definición de contacto entre pacientes laxa (estar en una misma área de habitaciones en algún punto del mismo día) que se basa sobre un modelo de datos altamente detallado ofrece dos ventajas principales: el descubrimiento de rutas que habrían quedado ocultas con una definición más estricta y la posibilidad de establecer diferentes niveles en la fuerza del contacto, pudiendo filtrar los resultados en base a ellos.

- **Objetivo 4**: *Estudio y selección de tecnologías basadas en grafos para el almacenamiento e inferencia de información. Además, implementación del modelo espacio-temporal y las tareas epidemiológicas en la tecnología seleccionada.*
  - En la comparación de las bases de datos orientadas a grafos (BDOG) comprobamos que:
    - Neo4j presenta un consumo de memoria más dependiente del subgrafo recorrido y no del tamaño total del grafo que contiene el conjunto de datos. Aunque presenta un menor consumo de memoria y tiempo que GraphDB en consultas simples, su escalabilidad es reducida respecto al aumento en el número de aristas y nodos a recorrer. En cuanto al almacenamiento del conjunto de datos, si bien precisa de alrededor de

10 veces menos de espacio, este no crece de manera lineal respecto al tamaño del grafo.
    - GraphDB ofrece un tiempo de ejecución y consumo de memoria mayor que Neo4j, pero más constante (apenas variable conforme aumenta el tamaño del grafo), con un comportamiento más predecible y escalable, incluso en grafos grandes.
  - El uso de formatos y lenguajes de consulta estándar, como RDF y SPARQL (y sus extensiones), proporciona una mayor flexibilidad en la selección de la tecnología para el almacenamiento, administración y consulta de los datos, así como más posibilidades de integración con otras herramientas y conjuntos de datos ya existentes.
  - El uso del formato RDF* ha sido preferible sobre RDF por su implementación más simple de las aristas con propiedades, lo que reduce la complejidad del modelo, así como el consumo de espacio de almacenamiento y tiempo y memoria de ejecución. En consonancia, SPARQL* también presenta una sintaxis más sencilla.
  - Destacamos que el carácter modular implícito en la sintaxis de SPARQL* (definición de los nodos y aristas a recorrer mediante tripletas) permite una fácil adaptación de las consultas para la adición de nuevos caminos a recorrer, filtros a aplicar o la permisión de definir consultas que puedan ser ejecutadas sobre un conjunto de datos que no esté completo respecto al modelo (un caso concreto relacionado con nuestro modelo puede ser que no se complete la jerarquía de *Localizaciones*).
  - Seleccionamos como tecnología más prometedora los grafos de conocimiento en formato RDF*, siendo SPARQL* el lenguaje utilizado para la implementación de las consultas epidemiológicas.

- **Objetivo 5**: *Cuantificación de la similitud espacio-temporal entre pacientes infectados a partir del análisis de contactos derivado de sus movimientos en el hospital.*
  - Probamos que la transformación en trayectoria de los movimientos de los pacientes dentro del hospital a partir de la información modelada en forma de eventos y su posterior análisis espacio-temporal mediante algoritmos para medir la distancia entre trayectorias (TDMA) nos permite una evaluación de la cercanía entre los pacientes basada en sus relaciones espacio-temporales.
  - Logramos una cuantificación semántica (basada en los tipos de aristas y clases de nodos definidos en el modelo de datos) de la similitud entre los puntos que conforman las trayectorias de los pacientes que aúna tanto la dimensión espacial como la temporal, así como las características epidemiológicas del brote.

- Del estudio comparativo de 6 TDMA (3 clásicos -DTW, STLC y ST-LCSS- y 3 extensiones nuevas de la tesis -ST-DTW, JSTLC y ST-LCSS-WTW-), observamos que:
  - En el caso de DTW y ST-DTW, encontramos dos algoritmos con fuertes restricciones respecto al orden en que los puntos pueden emparejarse, pero no en cuanto a la distancia máxima para permitir el emparejamiento entre dos puntos. En consecuencia, no existe una garantía de que la obtención de una similitud alta entre dos trayectorias sea únicamente debido a que sus puntos estén alineados espacio-temporalmente.
  - En el caso de STLC y JSTLC, obtenemos de forma general una similitud alta y sin diferencias significativas entre cada par de trayectorias. Esto se debe principalmente a dos motivos: el requerimiento de que todos los puntos de cada trayectoria tengan al menos un emparejamiento y el no establecimiento de una similitud mínima a cumplir en cada dimensión (espacial y temporal) para permitir el emparejamiento. En STLC se aúna el hecho de que cada dimensión es evaluada por separado, y cabe mencionar que, dado un punto, su punto más cercano temporalmente no es necesariamente el mismo que su punto más cercano espacialmente.
  - En el caso de ST-LCSS y ST-LCSS-WTW, el hecho de establecer límites espaciales y temporales para definir cuándo permitimos un emparejamiento y, en consecuencia, no forzar el emparejamiento de todos los puntos de ambas trayectorias permite una evaluación de la similitud entre las trayectorias que es más fiel a las relaciones espacio-temporales entre ellas.
  - Al comparar ST-LCSS y ST-LCSS-WTW, el mayor grado de independencia para el emparejamiento de puntos en ST-LCSS-WTW permite una mayor precisión para determinar qué puntos están más cerca espacio-temporalmente.
  - Establecer límites espaciales y temporales en el emparejamiento entre puntos también permite la evaluación de las trayectorias en distintos niveles de precisión. Por ejemplo, los límites espaciales pueden apoyarse en un modelo flexible y multinivel para representar la dimensión espacial como el propuesto en el Objetivo 3, de manera que se haga una evaluación inicial definiendo el contacto mediante una clase superior en la jerarquía, y posteriormente filtrar el resultado mediante definiciones más específicas del contacto entre pacientes.
- Observamos que la representación visual de los resultados de aplicar un TDMA sobre cada par de trayectorias entre pacientes como un mapa de calor puede ayudar para dar una interpretación inicial de las relaciones entre pacientes, siendo necesario información de apoyo, como puede ser la representación de las trayectorias en forma de líneas temporales. Sin embargo, la identificación de grupos de pacientes es dependiente de la

ordenación de la matriz, por lo que es preciso el análisis de los resultados con un algoritmo de *clustering*.

- **Objetivo 6**: *Agrupamiento de pacientes infectados en base a su conexión espacio-temporal para la detección de posibles brotes y potenciales rutas de transmisión.*
  - El algoritmo de clustering k-means busca minimizar la suma de las distancias entre cada elemento del clúster y el centroide de este. En el caso concreto de nuestro problema, la aplicación de k-means sobre una nube de datos uniforme en la que no hay diferencias significativas entre sus elementos conlleva a la obtención de clústeres en los que es difícil encontrar relaciones espacio-temporales significativas entre los pacientes pertenecientes al mismo clúster y los de otros clústeres.
  - Detectamos que, en el caso de las matrices de similitud (MS) obtenidas con DTW y ST-DTW, obtenemos una gran cantidad de pequeños clústeres que agrupan a los pacientes en base a un grano muy fino. Es decir, encontramos clústeres cuyos elementos son muy similares entre sí, pero que no están significativamente diferenciados del resto.
  - En el caso de las MS obtenidas con STLC y JSTLC, en general, presentan un rango de valores de reducido y similares entre ellos. Esto conlleva a la obtención de grandes clústeres que engloban la mayor parte de los pacientes, junto con otros pequeños clústeres que abarcan los elementos más diferenciados, pero no necesariamente parecidos entre sí.
  - Demostramos que aplicar la técnica de clustering k-means sobre la similitud espacio-temporal medida por pares de trayectorias de pacientes mediante los TDMA ST-LCSS y ST-LCSS-WTW nos permite la detección de brotes epidémicos, así como la identificación de posibles fuentes de contagio.

- **Objetivo 7**: *Una garantía de la reproducibilidad de la investigación realizada mediante repositorios de acceso público.*
  - Dentro del uso de datos de acceso libre, tanto MIMIC-III como el modelo de simulación de datos pueden ser usados para la realización de investigaciones epidémicas en entornos hospitalarios cuando no hay otras fuentes de datos disponibles, como pueden ser los sistemas de información hospitalaria o el registro de posición mediante sensores. Ambos conjuntos ofrecen información de gran detalle (podemos destacar la alta parametrización del modelo de simulación), tanto clínica como relacionada con los movimientos de pacientes dentro del hospital, que puede ser adaptado al modelo de datos propuesto.
  - En favor del cumplimiento de esta tesis con los principios de la ciencia abierta, el software implementado y los conjuntos de datos utilizados en la

validación de las propuestas es de acceso es público a través de repositorios en GitHub.

Presentadas las conclusiones, podemos afirmar que hemos cumplido todos los objetivos propuestos y, por tanto, la hipótesis de partida queda demostrada.

5.2 Trabajo futuro

Esta tesis y toda la investigación realizada en ella pueden servir como punto de partida para el siguiente trabajo futuro:

- La vinculación y adaptación del modelo espacio-temporal con otras ontologías relacionadas con la atención sanitaria y terminologías clínicas estándar, como *SNOMED CT* (*Systematized Nomenclature of Medicine - Clinical Terms*; en español, Nomenclatura Sistematizada de la Medicina – Términos Clínicos) o *ICD* (*International Classification of Diseases*; en español, Clasificación de Enfermedades Internacional). Por ejemplo, los *Episodios* y *Eventos* podrían conectarse con una jerarquía de clases de estas terminologías, permitiendo una descripción más precisa e interoperable de las enfermedades y causas de hospitalización, así como un filtro extra de los *Pacientes* en las consultas.

- Profundización del modelo en aspectos relacionados con la información genética de las bacterias detectadas en las pruebas microbiológicas. Por ejemplo, sería interesante combinar el análisis de las relaciones espacio-temporales entre pacientes como un análisis de la variación en el genoma de las bacterias detectada.

- Como mejora inmediata del trabajo realizado en el Capítulo 4, proponemos mejorar el algoritmo JSTLC añadiendo unos límites mínimos de similitud espacial y temporal para permitir el emparejamiento entre puntos, así como no requerir el emparejamiento de todos los puntos. Sería interesante la comparación entre los resultados de este algoritmo y los de ST-LCSS-WTW.

- Planeamos el uso de otros tipos de algoritmos de *clustering* para el agrupamiento de pacientes en base a su similitud. Entre las alternativas a estudiar se encuentra el *clustering jerárquico*, que se basa en la sucesiva unificación de clústeres, y cuyo resultado representado en forma de dendrograma puede proporcionar un soporte visual para comprender la estructura de las relaciones espacio-temporales; y el *clustering fuzzy* (en español, *difuso*), como *Fuzzy C-Means* [6], en el que un elemento del conjunto de datos de entrada puede ser asignado a más de un clúster con una determinada probabilidad de pertenencia.

- Plateamos la integración de la ecuación de $sim_{ST}(p_1, p_2)$ y algunos TDMA dentro del algoritmo de *clustering*, de forma que podamos realizar el análisis en un único paso.

- La aplicación de otras técnicas que puedan aprovechar la representación del modelo espacio-temporal en forma de grafo para la detección y análisis de brotes nosocomiales. Entre las opciones barajadas se encuentra el análisis de redes sociales (por ejemplo, algoritmos para la detección de comunidades) y la minería de patrones espacio-temporales.
- Creación de una herramienta *software* que permita la integración del modelo espacio-temporal, las consultas epidemiológicas y el método StESPT. En relación con el Objetivo 4, recomendamos que no se dependa de herramientas de terceros en lo máximo de lo posible y, concretamente, en relación con el Objetivo 3, planteamos crear una herramienta de visualización de los resultados de las consultas epidemiológicas en la que los nodos próximos sean representados como cúmulos cercanos.

Bibliografía

[1] M. H. Abbasi, F. Karimipour, and S. Gholipour. Detection of the association rules of the occurrence of brucellosis in humans using spatial data mining. Depiction of Health, 11:20–30, 03 2020. doi: 10.34172/doh.2020.03.

[2] V. Abhishek and V. Srivastava. Sis epidemic model under mobility on multi-layer networks. In 2020 American Control Conference (ACC, 2020.

[3] M. Alexander and R. Kobes. Effects of vaccination and population structure on influenza epidemic spread in the presence of two circulating strains. BMC public health, 11(1):8,, 2011.

[4] J. F. Allen. Maintaining knowledge about temporal intervals. Communications of the ACM, 26:832–843, 11 1983. ISSN 15577317. doi: 10.1145/182.358434. URL `https://dl.acm.org/doi/10.1145/182.358434`.

[5] C. Alota, C. Arceo, and A. Reyes V. An edge-based model of seir epidemics on static random networks. Bulletin of Mathematical Biology, 82(7):96,, 2020.

[6] J. C. D. and. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 3(3):32–57, 1973. doi: 10.1080/01969727308546046. URL `https://doi.org/10.1080/01969727308546046`.

[7] M. Andrews and C. Bauch. The impacts of simultaneous disease intervention decisions on epidemic outcomes. Journal of Theoretical Biology, 395:1–10,, 2016.

[8] N. Andrienko and G. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. Data Mining and Knowledge Discovery, 27(1):55–83, Jul 2013. ISSN 1573-756X. doi: 10.1007/s10618-012-0285-7.

[9] R. Angles, J. B. Antal, A. Averbuch, A. Birler, P. Boncz, M. Búr, O. Erling, A. Gubichev, V. Haprian, M. Kaufmann, J. L. L. Pey, N. Martínez, J. Marton, M. Paradies, M.-D. Pham, A. Prat-Pérez, D. Püroja, M. Spasić, B. A. Steer, D. Szakállas, G. Szárnyas, J. Waudby, M. Wu, and Y. Zhang. The ldbc social network benchmark, 2024. URL `https://arxiv.org/abs/2001.02299`.

[10] J. Angulo, H.-L. Yu, A. Langousis, A. Kolovos, J. Wang, A. Madrid, and G. Christakos. Spatiotemporal infectious disease modeling: a bme-sir approach. PloS One, 8(9):72168,, 2013.

[11] A. Antelmi, G. Cordasco, C. Spagnuolo, and V. Scarano. A design-methodology for epidemic dynamics via time-varying hypergraphs. In Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, Richland, SC, 2020.

[12] A. Apolloni, C. Poletto, J. Ramasco, P. Jensen, and V. Colizza. Metapopulation epidemic models with heterogeneous mixing and travel behaviour. Theoretical Biology and Medical Modelling, 11:3,, 2014.

[13] D. Arndt, J. Broekstra, B. DuCharme, O. Lassila, P. F. Patel-Schneider, E. Prud'hommeaux, T. Thibodeau Jr., and B. Thompson. RDF-star and SPARQL-star. W3C Community Group. W3C, Dec. 2021. URL `https://www.w3.org/2021/12/rdf-star.html`.

[14] G. Atemezing and F. Amardeilh. Benchmarking commercial rdf stores with publications office dataset. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11155 LNCS, page 379–394. Springer, 2018. doi: 10.1007/978-3-319-98192-5_54.

[15] S. Baker Effendi, B. van der Merwe, and W.-T. Balke. Suitability of graph database technology for the analysis of spatio-temporal data. Future Internet, 12 (5), 2020. ISSN 1999-5903. doi: 10.3390/fi12050078. URL `https://www.mdpi.com/1999-5903/12/5/78`.

[16] D. Balcan, B. Gonçalves, H. Hu, J. Ramasco, V. Colizza, and A. Vespignani. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. Journal of Computational Science, 1(3):132–145,, 2010.

[17] A. Banos, N. Corson, B. Gaudou, V. Laperrière, and S. Coyrehourcq. The importance of being hybrid for spatial epidemic models: A multi-scale approach. Systems, 3(4):309–329,, 2015.

[18] R. Barnard, I. Kiss, L. Berthouze, and J. Miller. Edge-based compartmental modelling of an sir epidemic on a dual-layer static–dynamic multiplex network with tunable clustering. Bulletin of Mathematical Biology, 80(10):2698–2733,, 2018.

[19] S. Barnes, B. Golden, and E. Wasil. A dynamic patient network model of hospital-acquired infections. In Proceedings of the 2010 Winter Simulation Conference, Baltimore, Maryland, 2010.

[20] J. I. Barrasa-Villar, C. Aibar-Remón, P. Prieto-Andrés, R. Mareca-Doñate, and J. Moliner-Lahoz. Impact on morbidity, mortality, and length of stay of hospital-acquired infections by resistant microorganisms. Clinical Infectious Diseases, 65(4):644–652, Aug. 2017.

[21] D. Beckett. RDF 1.1 N-Triples. W3C Community Group. W3C, Feb. 2014. URL https://www.w3.org/TR/n-triples/.

[22] D. Beckett, T. Berners-Lee, E. Prud'hommeaux, and G. Carothers. RDF 1.1 Turtle. W3C Community Group. W3C, Feb. 2014. URL https://www.w3.org/TR/turtle/.

[23] S. D. Bella, G. Sanson, J. Monticelli, V. Zerbato, L. Principe, M. Giuffrè, G. Pipitone, and R. Luzzati. Clostridioides difficile infection: history, epidemiology, risk factors, prevention, clinical manifestations, treatment, and future options. Clinical Microbiology Reviews, 37(2):e00135–23, 2024. doi: 10.1128/cmr.00135-23. URL https://journals.asm.org/doi/abs/10.1128/cmr.00135-23.

[24] P. Bellini and P. Nesi. Performance assessment of rdf graph databases for smart city services. Journal of Visual Languages and Computing, 45:24–38, 2018. doi: 10.1016/J.JVLC.2018.03.002.

[25] T. Berry and T. Sauer. Consistent manifold representation for topological data analysis. Foundations of Data Science, 1(1):1–38, 3 2019. doi: 10.3934/fods.2019001. URL https://www.aimsciences.org/article/id/2556e6c9-b4b9-455a-9d9e-886ef0cd166f.

[26] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In C. Beeri and P. Buneman, editors, Database Theory — ICDT'99, pages 217–235, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-49257-3.

[27] N. Bifolchi, R. Deardon, and Z. Feng. Spatial approximations of network-based individual level infectious disease models. Spatial and Spatio-Temporal Epidemiology, 6:59–70,, 2013.

[28] C. Bizer and R. Cyganiak. RDF 1.1 TriG. W3C Community Group. W3C, Feb. 2014. URL https://www.w3.org/TR/trig/.

[29] D. Brickley and R. Guha. RDF Schema 1.1. W3C Recommendation. W3C, Feb. 2014. URL https://www.w3.org/TR/rdf-schema/.

[30] D. Brockmann and D. Helbing. The hidden geometry of complex, network-driven contagion phenomena. Science, 342(6164):1337–1342,, 1979. doi: 10.1126/SCIENCE.1245200/SUPPL_FILE/BROCKMANN.SM.PDF.

[31] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pages 108–122, 2013.

[32] W. Cao. Comparative study on ventilation and air conditioning system schemes based on virus pollution control in hospital infusion room. Aerobiologia, 2023. doi: 10.1007/s10453-023-09801-x.

[33] N. Carnegie. Effects of contact network structure on epidemic transmission trees: implications for data required to estimate network structure. Statistics in Medicine, 37(2):236–248,, 2018.

[34] G. Carothers. RDF 1.1 N-Quads. W3C Community Group. W3C, Feb. 2014. URL https://www.w3.org/TR/n-quads/.

[35] S. Carr and S. Roberts. Planning for infectious disease outbreaks: A geographic disease spread, clinic location, and resource allocation simulation. In Proceedings of the 2010 Winter Simulation Conference, 2010.

[36] J. Chen, A. Marathe, and M. Marathe. Coevolution of epidemics, social networks, and individual behavior: A case study. In S.-K. Chai, J. J. Salerno, and P. L. Mabry, editors, Advances in Social Computing, pages 218–227, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-12079-4.

[37] P. Chen, D. Zhang, J. Liu, and I. Jian. Assessing personal exposure to covid-19 transmission in public indoor spaces based on fine-grained trajectory data: A simulation study. Build Environ, 218:109153,, 06 2022. doi: 10.1016/j.buildenv.2022.109153.

[38] C.-H. Cheng, Y.-H. Kuo, and Z. Zhou. Tracking nosocomial diseases at individual level with a real-time indoor positioning system. Journal of Medical Systems, 42 (11):222,, 2018.

[39] P. Cimiano and H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. Semant. Web, 8(3):489–508, Jan. 2017. ISSN 1570-0844. doi: 10.3233/SW-160218. URL https://doi.org/10.3233/SW-160218.

[40] O. Cliff, N. Harding, M. Piraveenan, E. Erten, M. Gambhir, and M. Prokopenko. Investigating spatiotemporal dynamics and synchrony of influenza epidemics in australia: An agent-based modelling approach. Simulation Modelling Practice and Theory, 87:412–431,, 2018.

[41] J. Codella, N. Safdar, R. Heffernan, and O. Alagoz. An agent-based simulation model for clostridium difficile infection control. Medical Decision Making, 35(2): 211–229,, 02 2015. doi: 10.1177/0272989X14545788/ASSET/IMAGES/LARGE/10.1177_0272989X14545788-FIG6.JPEG.

[42] P. Coletti, P. Libin, O. Petrof, L. Willem, S. Abrams, S. Herxog, C. Faes, E. Kuylen, J. Wambua, P. Beutels, and N. Hens. A data-driven metapopulation model for the belgian covid-19 epidemic: assessing the impact of lockdown and exit strategies. BMC infectious diseases, 21(1):503,, 2021.

[43] E. Colman, P. Holme, H. Sayama, and C. Gershenson. Efficient sentinel surveillance strategies for preventing epidemics on networks. PLos Coputational Biology, 15(11), 2019.

[44] L. Cowley. Genomics, social media and mobile phone data enable mapping of sars-cov-2 lineages to inform health policy in bangladesh. Nat Microbiol, 6(10): 1271–1278,, 10 2021. doi: 10.1038/s41564-021-00955-3.

[45] S. Curry. Incubation period of clostridioides difficile infection in hospitalized patients and long-term care facility residents: a prospective cohort study. Antimicrobial Stewardship and Healthcare Epidemiology : ASHE, 4(1):144,, 09 2024. doi: 10.1017/ASH.2024.392.

[46] M. Cusumano-Towner, D. Li, S. Tuo, G. Krishnan, and D. Maslove. A social network of hospital acquired infection built from electronic medical record data. Journal of the American Medical Informatics Association: JAMIA, 20 (3):427–434,, 2013.

[47] R. Cyganiak, D. Wood, and M. Lanthaler. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. W3C, Feb. 2014. URL `https://www.w3.org/TR/rdf11-concepts/`.

[48] 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, A. A. da Silva Machado, and o. b. o. N. of Hospital Research Study Collaborators. Geographical information system and spatial–temporal statistics for monitoring infectious agents in hospital: a model using klebsiella pneumoniae complex. Antimicrobial Resistance & Infection Control, 10(1):92, Jun 2021. ISSN 2047-2994. doi: 10.1186/s13756-021-00944-5.

[49] L. Danon, A. Ford, T. House, C. Jewell, M. Keeling, G. Roberts, J. Ross, and M. Vernon. Networks and the epidemiology of infectious disease. Interdisciplinary Perspectives on Infectious Diseases, 2011(284909), 2011.

[50] P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, and K. H. Jacobsen. Complexity of the basic reproduction number (r0). Emerging Infectious Diseases, 25:1, 1 2019. ISSN 10806059. doi: 10.3201/EID2501.171901. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6302597/`.

[51] E. Drezen, T. Guyet, and A. Happe. From medico-administrative databases analysis to care trajectories analytics: an example with the french snds. Fundamental & Clinical Pharmacology, 32(1):78–80, 2018. doi: https://doi.org/10.1111/fcp.12323. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/fcp.12323`.

[52] R. D'Mello, J. Melcher, and J. Torous. Similarity matrix-based anomaly detection for clinical intervention. Sci Rep, 12(1):9162,, 06 2022. doi: 10.1038/s41598-022-12792-3.

[53] L. Ehrlinger and W. Wöß. Towards a definition of knowledge graphs. In International Conference on Semantic Systems, volume 1695, 9 2016. URL https://ceur-ws.org/Vol-1695/paper4.pdf.

[54] A. El-Gilany. Covid-19 caseness: An epidemiologic perspective. J Infect Public Health, 14(1):61,, 01 2021. doi: 10.1016/J.JIPH.2020.11.003.

[55] J. Erber. Infection control measures and prevalence of sars-cov-2 igg among 4,554 university hospital employees, munich, germany. Emerg Infect Dis, 28(3): 572–581,, 03 2022. doi: 10.3201/eid2803.204436.

[56] J. M. Escamilla Molgora, L. Sedda, and P. M. Atkinson. Biospytial: spatial graph-based computing for ecological big data. GigaScience, 9(5), 05 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa039. URL https://doi.org/10.1093/gigascience/giaa039.

[57] European Centre for Disease Prevention and Control. Surveillance of healthcare-associated infections in intensive care units. Publications Office, LU, 2017. URL https://data.europa.eu/doi/10.2900/833186.

[58] K. Farrahi, R. Emonet, and M. Cebrian. Epidemic contact tracing via communication traces. PloS One, 9(5):95133,, 2014.

[59] S. Fast, M. González, J. Wilson, and N. Markuzon. Modelling the propagation of social response during a disease outbreak. Journal of the Royal Society, Interface, 12(104):20141105,, 2015.

[60] Y. Fatima-Zohra and H. Djamila. A surveillance and spatiotemporal visualization model for infectious diseases using social network. International Journal of Decision Support System Technology, 7(4):1–19,, 2015.

[61] S. Feng and Z. Jin. Infectious diseases spreading on an adaptive metapopulation network. IEEE Access, 8:153425–153435,, 2020.

[62] N. Ferro and L. Sinico. Graph databases benchmarking on the italian business register. In CEUR Workshop Proceedings, page 2161, 2018. URL https://ceur-ws.org/Vol-2161/paper43.pdf.

[63] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011.

[64] M. Friesen and R. Mcleod. A survey of agent-based modeling of hospital environments. IEEE Access, 2:227–233, 01 2014. doi: 10.1109/ACCESS.2014.2313957.

[65] M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, and A. Rinaldo. Spread and dynamics of the covid-19 epidemic in italy: Effects of emergency containment measures. Proceedings of the National Academy of Sciences of the United States of America, 117(19):10484–10491,, 2020.

[66] B. Gaudou, N. Huynh, D. Philippon, A. Brugière, K. Chapuis, P. Taillandier, P. Larmande, and A. Drogoul. Comokit: A modeling kit to understand, analyze, and compare the impacts of mitigation policies against the covid-19 epidemic at the scale of a city. Frontiers in Public Health, 8:563247,, 2020.

[67] J. Golbeck. Analyzing the Social Web. Morgan Kaufmann Publishers Inc, San Francisco, 2013.

[68] J. Gomez, J. Prieto, E. Leon, and A. Rodríguez. Infekta-an agent-based model for transmission of infectious diseases: The covid-19 case in bogotá, colombia. PloS One, 16(2):0245787,, 2021.

[69] A. Gordon. The hospital water environment as a reservoir for carbapenem-resistant organisms causing hospital-acquired infections-a systematic review of the literature. Clin Infect Dis, 64(10):1436–1444,, 05 2017. doi: 10.1093/CID/CIX132.

[70] D. Greene, P. Cunningham, and R. Mayer. Unsupervised Learning and Clustering, pages 51–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-75171-7. doi: 10.1007/978-3-540-75171-7_3. URL `https://doi.org/10.1007/978-3-540-75171-7_3`.

[71] B. Gross and S. Havlin. Epidemic spreading and control strategies in spatial modular network. Applied Network Science, 5(1):95,, 2020.

[72] G. Großmann, M. Backenköhler, and V. Wolf. Heterogeneity matters: Contact structure and individual variation shape epidemic dynamics. PloS One, 16(7): 0250050,, 2021.

[73] Q. Guo, Y. Lei, X. Jiang, Y. Ma, G. Huo, and Z. Zheng. Epidemic spreading with activity-driven awareness diffusion on multiplex network. Chaos, 26(4):043110,, 2016.

[74] H. Haddad, B. Moulin, M. Thériault, and D. Navarro-Velazquez. Integrated epidemiologic simulation for person to person contagion through urban mobility within gis. In Proceedings of the First ACM SIGSPATIAL International Workshop on Use of GIS in Public Health, New York, NY, USA, 2012.

[75] H. Haddad, B. Moulin, and M. Thériault. A fully gis-integrated simulation approach for analyzing the spread of epidemics in urban areas. SIGSPATIAL Special, 8(1):34–41,, 2016.

[76] A. Hagberg, P. J. Swart, and D. A. Schult. Exploring network structure, dynamics, and function using networkx. In Proceedings of the 7th Python in Science Conference. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 06 2008. doi: 10.25080/TCWV9851. URL https://www.osti.gov/biblio/960616.

[77] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. J Intell Inf Syst, 17(2–3):107–145,, 12 2001. doi: 10.1023/A:1012801612483/METRICS.

[78] W. Hamilton. Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 14(3):1–159,, 2020.

[79] D. Han, Q. Shao, and D. Li. Exploring the epidemic spreading in a multilayer metapopulation network by considering individuals' periodic travelling. Complexity, 2020:6782018,, 2020.

[80] P. Hernández, C. Pena, A. Ramos, and J. Gómez-Cadenas. A new formulation of compartmental epidemic modelling for arbitrary distributions of incubation and removal times. PloS One, 16(2):0244107,, 2021.

[81] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann. Knowledge graphs. ACM Comput. Surv., 54(4), July 2021. ISSN 0360-0300. doi: 10.1145/3447772. URL https://doi.org/10.1145/3447772.

[82] U. Hohenstein and M. Jergler. Database performance comparisons: An inspection of fairness. In DATA 2019 - Proceedings of the 8th International Conference on Data Science, Technology and Applications, page 243–250, 2019. doi: 10.5220/0007926602430250.

[83] M. Hornbrook, A. Hurtado, and R. Johnson. Health care episodes: Definition, measurement and use. Medical Care Research and Review, 42(2):163–218,, 1985. doi: 10.1177/107755878504200202.

[84] A. Hota and K. Gupta. A generalized sis epidemic model on temporal networks with asymptomatic carriers and comments on decay ratio. In 2021 American Control Conference (ACC, 2021.

[85] 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

Transactions on Knowledge and Data Engineering, 36(05):2191–2212,, 05 2024. doi: 10.1109/TKDE.2023.3323535.

[86] E. Hunter, B. Namee, and J. Kelleher. A model for the spread of infectious diseases in a region. International Journal of Environmental Research and Public Health, 17(9):3119,, 2020.

[87] J. D. Hunter. Matplotlib: A 2d graphics environment. Computing in Science and Engineering, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[88] G. Hwang, P. Mahoney, J. James, G. Lin, A. Berro, M. Keybl, D. Goedecke, J. Mathieu, and T. Wilson. A model-based tool to predict the propagation of infectious disease via airports. Travel Medicine and Infectious Disease, 10(1): 32–42,, 2012.

[89] A. Iqbal, M. K. Gangwani, A. Beran, D. S. Dahiya, A. H. Sohail, W. Lee-Smith, M. Aziz, and M. Hassan. Nosocomial vs healthcare associated vs community acquired spontaneous bacterial peritonitis: Network meta-analysis. The American Journal of the Medical Sciences, 366:305–313, 10 2023. ISSN 00029629. doi: 10.1016/j.amjms.2023.06.014.

[90] K. Jitkajornwanich, N. Pant, M. Fouladgar, and R. E. and. A survey on spatial, temporal, and spatio-temporal database research and an original example of relevant applications using sql ecosystem and deep learning. Journal of Information and Telecommunication, 4(4):524–559, 2020. doi: 10.1080/24751839.2020.1774153. URL `https://doi.org/10.1080/24751839.2020.1774153`.

[91] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. Mark. Mimic-iii, a freely accessible critical care database. Scientific Data, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35.

[92] J. Juang and Y.-H. Liang. The impact of vaccine success and awareness on epidemic dynamics. Chaos, 26(11):113105,, 2016.

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

[94] J.-Q. Kan, C. Ma, H.-F. Zhang, and B.-B. Xiang. Interplay of epidemic spreading and strategy-mixed awareness diffusion on multiplex networks. International Journal of Modern Physics C, 31(6):2050085,, 2020.

[95] P. Kasaie, D. Dowdy, and W. Kelton. An agent-based simulation of a tuberculosis epidemic: understanding the timing of transmission. In Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World, Washington, D.C, 2013.

[96] C. P. Kelly and J. T. LaMont. Clostridium difficile - more difficult than ever. New England Journal of Medicine, 359(18):1932–1940, 2008. doi: 10.1056/NEJMra0707500. URL https://www.nejm.org/doi/full/10.1056/NEJMra0707500.

[97] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical character, 115(772):700–721, 1927. ISSN 09501207. URL http://www.jstor.org/stable/94815.

[98] D. Kim, B. Canovas-Segura, A. Jimeno-Almazán, M. Campos, and J. Juarez. Spatial-temporal simulation for hospital infection spread and outbreaks of clostridioides difficile. Scientific Reports, 13(1):1–11, 2023. doi: 10.1038/s41598-023-47296-1.

[99] K. Kim, S. Lee, D. Lee, and K. Lee. Coupling effects on turning points of infectious diseases epidemics in scale-free networks. BMC bioinformatics, 18(7):250,, 2017.

[100] K. Kim, S. Yoo, S. Lee, D. Lee, and K.-H. Lee. Network analysis to identify the risk of epidemic spreading. Applied Sciences, 11(7):2997,, 2021.

[101] T. Kovács, G. Simon, and G. Mezei. Benchmarking graph database backends - what works well with wikidata? Acta Cybernetica, 24(1):43–60, 2019. doi: 10.14232/ACTACYB.24.1.2019.5.

[102] A. Krisnadhi and P. Hitzler. A core pattern for events. In Advances in Ontology Design and Patterns, volume 32, page 29–37. IOS Press, Amsterdam, 2017. doi: 10.3233/978-1-61499-826-6-29.

[103] C.-L. Kuo, W. Chan, and M. Chen. Impact of vaccination strategies for epidemic node-level svir probabilistic model : Pandemic simulation on social networks under various vaccination strategies. In 2020 International Conference on Public Health and Data Science (ICPHDS, 2020.

[104] A. Kuzdeuov, D. Baimukashev, B. Karabay, B. Ibragimov, A. Mirzakhmetov, M. Nurpeiissov, M. Lewis, and H. Varol. A network-based stochastic epidemic simulator: Controlling covid-19 with region-specific policies. IEEE Journal of Biomedical and Health Informatics, 24(10):2743–2754,, 2020.

[105] P.-C. Lai, C. Chow, H. Wong, K. Kwong, Y. Kwan, S. Liu, W. Tong, W. Cheung, and W. Wong. An early warning system for detecting h1n1 disease outbreak – a spatio-temporal approach. International Journal of Geographical Information Science, 29(7):1251–1268,, 2015.

[106] M. Lau, G. Gibson, H. Adrakey, A. McClelland, S. Riley, J. Zelner, G. Streftaris, S. Funk, J. Metcalf, B. Dalziel, and B. Grenfell. A

mechanistic spatio-temporal framework for modelling individual-to-individual transmission-with an application to the 2014-2015 west africa ebola outbreak. PLoS computational biology, 13(10):1005798,, 2017.

[107] T. Le, L. Wang, C. Zeng, L. Fu, Z. Liu, and J. Hu. Clinical and microbiological characteristics of nosocomial, healthcare-associated, and community-acquired klebsiella pneumoniae infections in guangzhou, china. Antimicrobial Resistance & Infection Control, 10(1):41, Feb. 2021.

[108] X. Le, M. Bui, and J. Cohen. A computational paradigm for the simulation of complex epidemic diseases. In Proceedings of the Seventh Symposium on Information and Communication Technology, New York, 2016.

[109] N. Leavitt. Will nosql databases live up to their promise? Computer, 43(2): 12–14, 2010. doi: 10.1109/MC.2010.58.

[110] B. Leclère, D. L. Buckeridge, P.-Y. Boëlle, P. Astagneau, and D. Lepelletier. Automated detection of hospital outbreaks: A systematic review of methods. PLOS ONE, 12(4):1–16, 04 2017. doi: 10.1371/journal.pone.0176438. URL `https://doi.org/10.1371/journal.pone.0176438`.

[111] B. Lee, S. Brown, P. Cooley, R. Zimmerman, W. Wheaton, S. Zimmer, J. Grefenstette, T.-M. Assi, T. Furphy, D. Wagener, and D. Burke. A computer simulation of employee vaccination to mitigate an influenza epidemic. American Journal of Preventive Medicine, 38(3):247–257,, 2010.

[112] D. Lee and M. Zhu. Epidemic spreading in a social network with facial masks wearing individuals. IEEE Transactions on Computational Social Systems, 8(6): 1393–1406,, 2021.

[113] D. Levick and J. Osheroff. A clinical decision support implementation guide: Practical considerations. In R. A. Greenes, editor, Clinical Decision Support (Second Edition), pages 689–709. Academic Press, Oxford, second edition edition, 2014. ISBN 978-0-12-398476-0. doi: https://doi.org/10.1016/B978-0-12-398476-0.00025-7. URL `https://www.sciencedirect.com/science/article/pii/B9780123984760000257`.

[114] R. Levins. Some demographic and genetic consequences of environmental heterogeneity for biological control. Bulletin of the Entomological Society of America, 15(3):237–240,, 1969.

[115] C. Li, G.-P. Jiang, Y. Song, L. Xia, Y. Li, and B. Song. Modeling and analysis of epidemic spreading on community networks with heterogeneity. Journal of Parallel and Distributed Computing, 119:136–145,, 2018.

[116] J. Li, Z. Jin, Y. Yuan, and G.-Q. Sun. A non-markovian sir network model with fixed infectious period and preventive rewiring. Computers and Mathematics with Applications, 75(11):3884–3902,, 2018.

[117] W. Li, S. Wang, S. Wu, Z. Gu, and Y. Tian. Performance benchmark on semantic web repositories for spatially explicit knowledge graph applications. Computers, Environment and Urban Systems, 98, 2022. doi: 10.1016/J.COMPENVURBSYS.2022.101884.

[118] J. Liang, H.-Y. Yuan, L. Wu, and D. Pfeiffer. Estimating effects of intervention measures on covid-19 outbreak in wuhan taking account of improving diagnostic capabilities using a modelling approach. BMC infectious diseases, 21(1):424,, 2021.

[119] J. Liao, X. Hu, M. Wang, M. Leeson, E. Hines, and E. Paolo. A ripple-spreading network model for the study of infectious disease transmission. In 2012 5th International Conference on BioMedical Engineering and Informatics, 2012.

[120] A. Liberati, D. Altman, J. Tetzlaff, C. Mulrow, P. Gøtzsche, J. Ioannidis, M. Clarke, P. Deveraux, J. Kleijnen, and D. Moher. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLOS Medicine, 6(7):1–28,, 2009.

[121] K. Linka, M. Peirlinck, F. Costabal, and E. Kuhl. Outbreak dynamics of covid-19 in europe and the effect of travel restrictions. Computer Methods in Biomechanics and Biomedical Engineering, 23(11):710–717,, 2020.

[122] T. Liu, P. Li, Y. Chen, and J. Zhang. Community size effects on epidemic spreading in multiplex social networks. PloS One, 11(3):0152021,, 2016.

[123] A. Lombardi, N. Amoroso, A. Monaco, S. Tangaro, and R. Belloti. Complex network modelling of origin–destination commuting flows for the covid-19 epidemic spread analysis in italian lombardy region. Applied Sciences, 11(10):4381,, 2021.

[124] W. Luo. Visual analytics of geo-social interaction patterns for epidemic control. International Journal of Health Geographics, 15(1):28,, 2016.

[125] I. Lymperopoulos. #stayhome to contain covid-19: Neuro-sir – neurodynamical epidemic modeling of infection patterns in social networks. Expert Systems with Applications, 165, 2021.

[126] J. Ma. Estimating epidemic exponential growth rate and basic reproduction number. Infectious Disease Modelling, 5:129–141, 1 2020. ISSN 2468-0427. doi: 10.1016/J.IDM.2019.12.009.

[127] A. Machens, F. Gesualdo, C. Rizzo, A. Tozzi, A. Barrat, and C. Cattuto. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. BMC infectious diseases, 13:185,, 2013.

[128] F. Majid, M. Gray, A. M. Deshpande, S. Ramakrishnan, M. Kumar, and S. Ehrlich. Non-pharmaceutical interventions as controls to mitigate the spread of epidemics: An analysis using a spatiotemporal pde model and covid–19 data. ISA Transactions, 124:215–224, 2022. ISSN 0019-0578. doi: https://doi.org/10.1016/j.isatra.2021.02.038.

[129] L. Mao. Cost-effectiveness of workplace closure and travel restriction for mitigating influenza outbreaks: a network-based simulation. In Proceedings of the Second ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health, New York, NY, USA, 2013.

[130] L. Mao and Y. Yang. Coupling infectious diseases, human preventive behavior, and networks–a conceptual framework for epidemic modeling. Social Science and Medicine, 74(2):167–175,, 1982.

[131] G. Martín, M.-C. Marinescu, D. Singh, and J. Carretero. Leveraging social networks for understanding the evolution of epidemics. BMC systems biology, 5 (3):14,, 2011.

[132] R. W. Mathes, R. Lall, A. Levin-Rector, J. Sell, M. Paladini, K. J. Konty, D. Olson, and D. Weiss. Evaluating and implementing temporal, spatial, and spatio-temporal methods for outbreak detection in a local syndromic surveillance system. PLOS ONE, 12(9):1–19, 09 2017. doi: 10.1371/journal.pone.0184419. URL `https://doi.org/10.1371/journal.pone.0184419`.

[133] A. Matsuki and G. Tanaka. Intervention threshold for epidemic control in susceptible-infected-recovered metapopulation models. Physical Review. E, 100 (2-1):022302,, 2019.

[134] L. Mcdonald. Vital signs: Preventing clostridium difficile infections. JAMA The Journal of the American Medical Association, 307:1684–1687,, 2012. doi: 10.15585/mmwr.mm6435a9.

[135] R. McFee and G. Abdelsayed. Clostridium difficile. Disease-a-Month, 55(7): 439–470,, 07 2009. doi: 10.1016/J.DISAMONTH.2009.04.010.

[136] S. Merler, M. Ajelli, L. Fumanelli, M. Gomes, A. Piontti, L. Rossi, D. Chao, I. Longini, M. Halloran, and A. Vespignani. Spatiotemporal spread of the 2014 outbreak of ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. The Lancet. Infectious Diseases, 15(2):204–211,, 2015.

[137] H. Mills and S. Riley. The spatial resolution of epidemic peaks. PLoS Computational Biology, 10(4):1003561,, 2014.

[138] Y. Min, X. Jin, Y. Ge, and J. Chang. The role of community mixing styles in shaping epidemic behaviors in weighted networks. PloS One, 8(2):57100,, 2013.

[139] R.-X. Ming, J.-M. Liu, W. Cheung, and X. Wan. Stochastic modelling of infectious diseases for heterogeneous populations. Infectious Diseases of Poverty, 5(1):107,, 2016.

[140] E. Mizutani and S. Dreyfus. On using dynamic programming for time warping in pattern recognition. Inf Sci, N Y), vol. 580:684–704,, 11 2021. doi: 10.1016/J.INS.2021.08.075.

[141] B. Mo, K. Feng, Y. Shen, C. Tam, D. Li, Y. Yin, and J. Zhao. Modeling epidemic spreading through public transit using time-varying encounter network. Transportation Research Part C: Emerging Technologies, 122:102893,, 2021.

[142] N. A. Mohd Asri, S. Ahmad, R. Mohamud, N. Mohd Hanafi, N. F. Mohd Zaidi, A. A. Irekeola, R. H. Shueb, L. C. Yee, N. Mohd Noor, F. H. Mustafa, C. Y. Yean, and N. Y. Yusof. Global prevalence of nosocomial Multidrug-Resistant klebsiella pneumoniae: A systematic review and Meta-Analysis. Antibiotics (Basel), 10 (12), Dec. 2021.

[143] D. Moher, A. Liberati, J. Tetzlaff, D. Altman, and T. P. Group. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. PLOS Medicine, 6(7):332–339,, 2009.

[144] J. Monteiro, F. Sá, and J. Bernardino. Experimental evaluation of graph databases: Janusgraph, nebula graph, neo4j, and tigergraph. Applied Sciences (Switzerland, 13(9), 2023. doi: 10.3390/APP13095770.

[145] 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. Trop Med Infect Dis, 6(3), 09 2021. doi: 10.3390/TROPICALMED6030141.

[146] R. Mukhamadiarov, S. Deng, S. Serrao, R. Priyanka, L. Yao, and U. Täuber. Social distancing and epidemic resurgence in agent-based susceptible-infectious-recovered models. Scientific Reports, 11(1):130,, 2021.

[147] C. J. L. Murray, K. S. Ikuta, F. Sharara, L. Swetschinski, G. R. Aguilar, A. Gray, and et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet, 399(10325):629–655, 2022. doi: 10.1016/S0140-6736(21)02724-0.

[148] A. Myall. Network memory in the movement of hospital patients carrying antimicrobial-resistant bacteria. Appl Netw Sci, 6(1):34,, 12 2021. doi: 10.1007/s41109-021-00376-5.

[149] A. Myall. Prediction of hospital-onset covid-19 infections using dynamic networks of patient contact: an international retrospective cohort study. Lancet Digit Health, 4(8):573– 583,, 08 2022. doi: 10.1016/S2589-7500(22)00093-0.

[150] A. Myall, R. 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. URL https://europepmc.org/article/PPR/PPR309081.

[151] R. E. Nelson, K. M. Hatfield, H. Wolford, and et al. National estimates of healthcare costs associated with multidrug-resistant bacterial infections among hospitalized patients in the united states. Clinical Infectious Diseases, 72 (Supplement 1):S17–S26, 01 2021. ISSN 1058-4838. doi: 10.1093/cid/ciaa1581. URL https://doi.org/10.1093/cid/ciaa1581.

[152] Neo4j. The neo4j python driver manual v4.4, 2023. URL https://neo4j.com/docs/python-manual/4.4/.

[153] V. Nguyen, R. Mikolajczyk, and E. Hernandez-Vargas. High-resolution epidemic simulation using within-host infection and contact data. BMC public health, 18 (1):886,, 2018.

[154] C. Nowzari, V. Preciado, and G. Pappas. Stability analysis of generalized epidemic models over directed networks. In 53rd IEEE Conference on Decision and Control, 2014.

[155] C. Nowzari, M. Ogura, V. Preciado, and G. Pappas. Optimal resource allocation for containing epidemics on time-varying networks. In 2015 49th Asilomar Conference on Signals, Systems and Computers, 2015.

[156] Ontotext. What is rdf-star?, 2021. URL https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-star/. [Online; accessed 17-March-2025].

[157] K. Paarporn, C. Eksin, J. Weitz, and J. Shamma. The effect of awareness on networked sis epidemics. In 2016 IEEE 55th Conference on Decision and Control (CDC, 2016.

[158] K. Paarporn, C. Eksin, J. Weitz, and J. Shamma. Networked sis epidemics with awareness. IEEE Transactions on Computational Social Systems, 4(3):93–103,, 2017.

[159] S. Pai, P. Polgreen, A. Segre, D. Sewell, and S. Pemmaraju. Spatiotemporal clustering of in-hospital clostridioides difficile infection. Infection Control & Hospital Epidemiology, 41(4):418–424,, 04 2020. doi: 10.1017/ICE.2019.350.

[160] J.-O. Palacio-Niño and F. Berzal. Evaluation metrics for unsupervised learning algorithms, 05 2019. URL https://arxiv.org/abs/1905.05667v2. Accessed: Nov. 25, 2024. [Online]. Available:.

[161] M. Peirlinck, K. Linka, F. Costabal, and E. Kuhl. Outbreak dynamics of covid-19 in china and the united states. Biomechanics and Modeling in Mechanobiology, 19(6):2179–2193,, 2020.

[162] X.-L. Peng, X.-J. Xu, M. Small, X. Fu, and Z. Jin. Prevention of infectious diseases by public vaccination and individual protection. Journal of Mathematical Biology, 73(6-7):1561–1594,, 2016.

[163] K. Petersen, S. Vakkalanka, and L. Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 64:1–18,, 2015.

[164] G. Peñalva, R. Cantón, and M. T. e. a. Pérez-Rodríguez. s. The Lancet Regional Health - Europe, 51, Jan 2025. ISSN 2666-7762. doi: 10.1016/j.lanepe.2025.101220. URL https://doi.org/10.1016/j.lanepe.2025.101220.

[165] D. U. Pfeiffer and K. B. Stevens. Spatial and temporal epidemiological analysis in the big data era. Prev Vet Med, 122(1-2):213–220, June 2015.

[166] M. Piotrowska, K. Sakowski, A. Karch, H. Tahir, J. Horn, M. Kretzschmar, and R. Mikolajczyk. Modelling pathogen spread in a healthcare network: Indirect patient movements. PLoS computational biology, 16(11):1008442,, 2020.

[167] A. Piyush Shanker and A. Rajagopalan. Off-line signature verification using dtw. Pattern Recognition Letters, 28(12):1407–1414, 2007. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2007.02.016. URL https://www.sciencedirect.com/science/article/pii/S0167865507000669.

[168] M. Porta. Operational Research. Oxford University Press, 2016. ISBN 9780199390069. doi: 10.1093/acref/9780199976720.013.1352. URL https://www.oxfordreference.com/view/10.1093/acref/9780199976720.001.0001/acref-9780199976720-e-1352.

[169] F. Prestinaci, P. Pezzotti, and A. Pantosti. Antimicrobial resistance: a global multifaceted phenomenon. Pathog Glob Health, 109(7):309,, 10 2015. doi: 10.1179/2047773215Y.0000000030.

[170] J. Price. Development and delivery of a real-time hospital-onset covid-19 surveillance system using network analysis. Clinical Infectious Diseases, 72(1):82–89,, 01 2021. doi: 10.1093/CID/CIAA892.

[171] J. Pujara, H. Miao, L. Getoor, and W. Cohen. Knowledge graph identification. In Proceedings of the 12th International Semantic Web Conference - Part I, ISWC '13, page 542–557, Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 9783642413346. doi: 10.1007/978-3-642-41335-3_34. URL https://doi.org/10.1007/978-3-642-41335-3_34.

[172] I. Pulido-Valdeolivas, D. Gómez-Andrés, J. A. Martín-Gonzalo, I. Rodríguez-Andonaegui, J. López-López, S. I. Pascual-Pascual, and E. Rausell. Gait phenotypes in paediatric hereditary spastic paraplegia revealed by dynamic time warping analysis and random forests. PLOS ONE, 13(3):1–28, 03 2018. doi: 10.1371/journal.pone.0192345. URL https://doi.org/10.1371/journal.pone.0192345.

[173] RDFLib. Sparql endpoint interface to python, 2022. URL https://sparqlwrapper.readthedocs.io/en/latest/.

[174] L. Rocha and V. Blondel. Bursts of vertex activation and epidemics in evolving networks. PLoS computational biology, 9(3):1002974,, 2013.

[175] L. Rocha, V. Singh, M. Esch, T. Lenaerts, F. Liljeros, and A. Thorson. Dynamic contact networks of patients and mrsa spread in hospitals. Scientific Reports, 10 (1):9336,, 2020.

[176] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math, 20:53–65,, 1987. doi: doi:. URL https://doi.org/10.1016/0377-0427(87)90125-7.

[177] A. Ruiz-Herrera and P. Torres. The role of movement patterns in epidemic models on complex networks. Bulletin of Mathematical Biology, 83(10):98,, 2021.

[178] G. Röst, F. Bartha, N. Bogya, P. Boldog, A. Dénes, T. Ferenci, K. Horváth, A. Juhász, C. Nagy, T. Tekeli, Z. Vizi, and B. Oroszi. Early phase of the covid-19 outbreak in hungary and post-lockdown scenarios. Viruses, 12(7):708,, 2020.

[179] A. Sahasranaman and H. Jensen. Poverty in the time of epidemic: A modelling perspective. PloS One, 15(11):0242042,, 2020.

[180] F. Sahneh and C. Scoglio. Epidemic spread in human networks. In 2011 50th IEEE Conference on Decision and Control and European Control Conference, 2011.

[181] F. Sahneh, C. Scoglio, and F. Chowdhury. Effect of coupling on the epidemic threshold in interconnected complex networks: A spectral analysis. In Proceedings of the American Control Conference, 2013.

[182] F. Sahneh, A. Vajdi, J. Melander, and C. Scoglio. Contact adaption during epidemics: A multilayer network formulation approach. IEEE Transactions on Network Science and Engineering, 6(1):16–30,, 2019.

[183] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

[184] L. Schaposnik and A. Zhang. Modeling epidemics on d-cliqued graphs. Letters in Biomathematics, 5:49–69, 01 2018. doi: 10.1080/23737867.2017.1419080.

[185] C. Seibold and H. Highlander. Modeling epidemics on a regular tree graph. Letters in Biomathematics, 3:59–74, 01 2016. doi: 10.1080/23737867.2016.1185979.

[186] S. Shang, R. Ding, K. Zheng, C. Jensen, P. Kalnis, and X. Zhou. Personalized trajectory matching in spatial networks. VLDB Journal, 23(3):449–468,, 07 2014. doi: 10.1007/S00778-013-0331-0/METRICS.

[187] S. Shang, L. Chen, Z. Wei, C. Jensen, K. Zheng, and P. Kalnis. Trajectory similarity join in spatial networks. Proceedings of the VLDB Endowment, 10 (11):1178–1189,, 08 2017. doi: 10.14778/3137628.3137630.

[188] Y. Shang. Modeling epidemic spread with awareness and heterogeneous transmission rates in networks. Journal of Biological Physics, 39(3):489–500,, 2013.

[189] R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In Proceedings of the 4th Asian Conference on The Semantic Web, ASWC '09, page 153–167, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 9783642108709. doi: 10.1007/978-3-642-10871-6_11.

[190] B. Shneiderman, C. Plaisant, and B. W. Hesse. Improving healthcare with interactive visualization. Computer, 46(5):58–66, 2013. doi: 10.1109/MC.2013.38.

[191] J. D. Siegel, E. Rhinehart, M. Jackson, and L. Chiarello. 2007 guideline for isolation precautions: Preventing transmission of infectious agents in health care settings. American Journal of Infection Control, 35(10):S65–S164, Dec 2007. ISSN 0196-6553. doi: 10.1016/j.ajic.2007.10.007.

[192] 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, 29 (6):2300335,, 02 2024. doi: 10.2807/1560-7917.ES.2024.29.6.2300335/CITE/REFWORKS.

[193] H. Skutkova, M. Vitek, P. Babula, R. Kizek, and I. Provaznik. Classification of genomic signals using dynamic time warping. BMC Bioinformatics, 14(10):S1, Aug 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-S10-S1. URL `https://doi.org/10.1186/1471-2105-14-S10-S1`.

[194] M. Small and D. Cavanagh. Modelling strong control measures for epidemic propagation with networks—a covid-19 case study. IEEE Access, 8: 109719–109731,, 2020.

[195] T. Smieszek, M. Balmer, J. Hattendorf, K. Axhausen, J. Zinsstag, and R. Scholz. Reconstructing the 2003/2004 h3n2 influenza epidemic in switzerland with a spatially explicit, individual-based model. BMC infectious diseases, 11:115,, 2011.

[196] K. Smith, M. Goldberg, S. Rosenthal, L. Carlson, J. Chen, C. Chen, and S. Ramachandran. Global rise in human infectious disease outbreaks. Journal of the Royal Society Interface, 11(101):20140950,, 2014.

[197] M. Sporny, D. Longley, G. Kellog, M. Lanthaler, P.-A. Champin, and N. Lindström. JSON-LD 1.1. A JSON-based Serialization for Linked Data. W3C Community Group. W3C, July 2020. URL `https://www.w3.org/TR/json-ld/`.

[198] H. Su, S. Liu, B. Zheng, X. Zhou, and K. Zheng. A survey of trajectory distance measures and performance evaluation. VLDB Journal, 29(1):3–32,, 01 2020. doi: 10.1007/S00778-019-00574-9/METRICS.

[199] 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. Journal of Clinical Medicine, 11 (11):3204,, 2022. doi: 10.3390/JCM11113204.

[200] A. Teslya, T. Pham, N. Godijk, M. Kretzschmar, M. Bootsma, and G. Rozhnova. Impact of self-imposed prevention measures and short-term government-imposed social distancing on mitigating and delaying a covid-19 epidemic: A modelling study. PLoS medicine, 17(7):1003166,, 2020.

[201] The Matplotlib Development Team. Matplotlib: Visualization with python, Aug. 2024. URL `https://doi.org/10.5281/zenodo.13308876`.

[202] D. S. Thompson. Methicillin-resistant staphylococcus aureus in a general intensive care unit. J R Soc Med, 97(11):521–526, Nov. 2004.

[203] M. Tizzoni, P. Bajardi, C. Poletto, J. Ramasco, D. Balcan, B. Gonçalves, N. Perra, V. Colizza, and A. Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. BMC medicine, 10:165,, 2012.

[204] Y.-S. Tsai, C.-T. Sun, T.-H. Wen, and M.-Y. Yen. Integrating epidemic dynamics with daily commuting networks: Building a multilayer framework to assess influenza a (h1n1) intervention policies. SIMULATION, 87(5):385–405,, 2011.

[205] Y.-J. Tseng, J.-H. Wu, X.-O. Ping, H.-C. Lin, Y.-Y. Chen, R.-J. Shang, M.-Y. Chen, F. Lai, and Y.-C. Chen. A web-based multidrug-resistant organisms surveillance and outbreak detection system with rule-based classification and clustering. Journal of medical Internet research, 14:e131, 10 2012. doi: 10.2196/jmir.2056.

[206] S. Tuarob, C. Tucker, M. Salathe, and N. Ram. Modeling individual-level infection dynamics using social network information. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, New York, NY, USA, 2015.

[207] A. Turner, M. Doxa, and D. O'Sullivan. From isovists to visibility graphs: A methodology for the analysis of architectural space. Environment and Planning B: Planning and Design, 28:103–121, 02 2001. doi: 10.1068/b2684.

[208] E. Vergu, H. Busson, and P. Ezanno. Impact of the infection period distribution on the epidemic spread in a metapopulation model. PloS One, 5(2):9371,, 2010.

[209] M. Vilain and H. Kautz. Constraint propagation algorithms for temporal reasoning. Proceedings of the National Conference on Artificial Intelligence, 1: 377 – 382, 1 1986.

[210] E. Volz, J. Miller, A. Galvani, and L. Meyers. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS computational biology, 7(6):1002042,, 2011.

[211] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. Commun. ACM, 57(10):78–85, Sept. 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

[212] Y. Wan. 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. Journal of Infectious Diseases, 230(1):159– 170,, 07 2024. doi: 10.1093/infdis/jiae019.

[213] J. Wang, X. Wang, and J. Wu. Inferring metapopulation propagation network for intra-city epidemic control and prevention. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2018.

[214] R. Wang, Z. Yang, W. Zhang, and X. Lin. An empirical study on recent graph database systems. In G. Li, H. T. Shen, Y. Yuan, X. Wang, H. Liu, and X. Zhao, editors, Knowledge Science, Engineering and Management, pages 328–340, Cham, 2020. Springer International Publishing. ISBN 978-3-030-55130-8.

[215] 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 Min Knowl Discov, 26(2):275–309,, 03 2013. doi: 10.1007/S10618-012-0250-5/METRICS.

[216] D. Weatherall, B. Greenwood, H. L. Chee, and P. Wasi. Science and technology for disease control: Past, present, and future. In Disease Control Priorities in Developing Countries. Oxford University Press, New York, 2nd edition, 2006.

[217] West Virginia Bureau for Public Health. Guidelines for clostridium difficile (c. diff) outbreaks in long-term care facilities (ltcfs), 2013. URL https://oeps.wv.gov/toolkits/documents/cdi/CDiff-Guidelines.pdf.

[218] WHO Regional Office for Europe and European Centre for Disease Prevention and Control. Antimicrobial Resistance Surveillance in Europe 2022 – 2020 Data. WHO Regional Office for Europe, Copenhagen, 2022.

[219] S. Winkelmann, J. Zonker, C. Schütte, and N. Conrad. Mathematical modeling of spatio-temporal population dynamics and application to epidemic spreading. Mathematical Biosciences, 336:108619,, 2021.

[220] C.-E. A. Winslow. The untilled fields of public health. Science, 51(1306):23–33, 1920. doi: 10.1126/science.51.1306.23. URL https://www.science.org/doi/abs/10.1126/science.51.1306.23.

[221] World Health Organization. Prevention of hospital-acquired infections: a practical guide, 2002. URL https://iris.who.int/handle/10665/67350. Accessed: Dec. 05, 2024. [Online].

[222] World Health Organization. Characterising contact in disease outbreaks via a network model of spatial-temporal proximity. WHO Library Cataloguing-in-Publication Data, Geneva, 2012. ISBN 978 92 4 150318 1. URL https://iris.who.int/bitstream/handle/10665/44812/9789241503181_eng.pdf.

[223] World Health Organization. Regional Office for Europe. Ottawa Charter for Health Promotion. World Health Organization. Regional Office for Europe, 1986. URL https://www.who.int/publications/i/item/ottawa-charter-for-health-promotion.

[224] Q. Wu, X. Fu, M. Small, and X.-J. Xu. The impact of awareness on epidemic spreading in networks. Chaos: An Interdisciplinary Journal of Nonlinear Science, 22(1):013101,, 2012.

[225] Z. Xu, Z. Zu, T. Zheng, W. Zhang, Q. Xu, and J. Liu. Long-distance travel behaviours accelerate and aggravate the large-scale spatial spreading of infectious diseases. Computational and Mathematical Methods in Medicine, 2014:295028,, 2014.

[226] K. Yashima and A. Sasaki. Epidemic process over the commute network in a metropolitan area. PloS One, 9(6):98518,, 2014.

[227] B. Yi, H. Jagadish, and C. Faloutsos. Efficient retrieval of similar time sequences under time warping. Proc Int Conf Data Eng, pages 201–208,, 1998. doi: 10.1109/ICDE.1998.655778.

[228] Q. Yin, T. Shi, C. Dong, and Z. Yan. The impact of contact patterns on epidemic dynamics. PloS One, 12(3):0173411,, 2017.

[229] Z. Zaplotnik, A. Gavrić, and L. Medic. Simulation of the covid-19 epidemic on the social network of slovenia: Estimating the intrinsic forecast uncertainty. PloS One, 15(8):0238090,, 2020.

[230] H.-F. Zhang, Z. Yang, Z.-X. Wu, B.-H. Wang, and T. Zhou. Braess's paradox in epidemic game: better condition results in less payoff. Scientific Reports, 3: 3292,, 2013.

[231] M. Zhang, A. Verbraeck, R. Meng, B. Chen, and X. Qiu. Modeling spatial contacts for epidemic prediction in a large-scale artificial city. Journal of Artificial Societies and Social Simulation, 19(4):3,, 2016.

[232] N. Zhang, W. Chen, P.-T. Chan, H.-L. Yen, J.-T. Tang, and Y. Li. Close contact behavior in indoor environment and transmission of respiratory infection. Indoor Air, 30(4):645–661,, 2020. doi: 10.1111/ina.12673.

[233] T. Zhang, M. Lees, C. Kwoh, X. Fu, G. Lee, and R. Goh. A contact-network-based simulation model for evaluating interventions under ẅhat-ifs̈cenarios in epidemic. In Proceedings of the Winter Simulation Conference, Berlin, Germany, 2012.

[234] X. Zhang, B. Ge, Q. Wang, J. Jiang, H. You, and Y. Chen. Epidemic spreading characteristics and immunity measures based on complex network with contact strength and community structure. Mathematical Problems in Engineering, 2015: 1–12,, 2015.

[235] Z. Zhang, H. Wang, C. Wang, and H. Fang. Modeling epidemics spreading on social contact networks. IEEE Transactions on Emerging Topics in Computing, 3(3):410–419,, 2015.

[236] C. Zheng, Z. Wang, and C. Xia. A novel epidemic model coupling the infectious disease with awareness diffusion on multiplex networks. In 2018 Chinese Control And Decision Conference (CCDC, 2018.

[237] L. Zheng and L. Tang. A node-based sirs epidemic model with infective media on complex networks. Complexity, 2019:2849196,, 2019.

[238] 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, 300:112207,, 09 2024. doi: 10.1016/j.knosys.2024.112207.

[239] L. Zino, A. Rizzo, and M. Porfiri. On assessing control actions for epidemic models on temporal networks. IEEE Control Systems Letters, 4(4):797–802,, 2020.

[240] C. Zuo, A. Wang, F. Zhu, Z. Meng, and X. Zhao. A new coupled awareness-epidemic spreading model with neighbor behavior on multiplex networks. Complexity, 2021:6680135,, 2021.

[241] Q. Zuo. Research on spatial-temporal spread and risk profile of the covid-19 epidemic based on mobile phone trajectory data. Front Big Data, 5:705698,, 04 2022. doi: 10.3389/fdata.2022.705698.

Apéndice A

Asignación de valores para la propiedad *coste*

En el experimento de la Sección 4.4, la asignación de los valores de la propiedad *coste* de las relaciones entre las distintas clases que conforman la dimensión espacial del modelo de datos hemos realizado en base a las dimensiones (en metros) de una *Cama* hospitalaria y una *Habitación* para hospitalizaciones con dos *Camas*, una al lado de la otra. A partir de estas dimensiones y de cómo se organizan las *Habitaciones* para hospitalizaciones (queda excluida para el cálculo la *Planta 0*) en el hospital que hemos diseñado para el experimento, hemos establecido un proceso metódico con el que calcular el valor de *coste*. Organizamos este apéndice por tipo de relación y clases de los nodos origen y destino.

Las dimensiones base son las siguientes:

- Una *Cama* mide 1×2 metros.
- Una *Habitación* mide $3 \times 6,5$ metros (ver Figura A.1).

A.1 Cama $\rightarrow^{contiguo}$ Cama

Dos *Camas* que están contiguas en una misma *Habitación* tienen una separación de 1 metro.

A.2 Cama $\rightarrow^{opuesto}$ Cama

En caso de que haya dos *Camas* contrapuestas, entre ellas hay un pasillo de 1,5 metros.

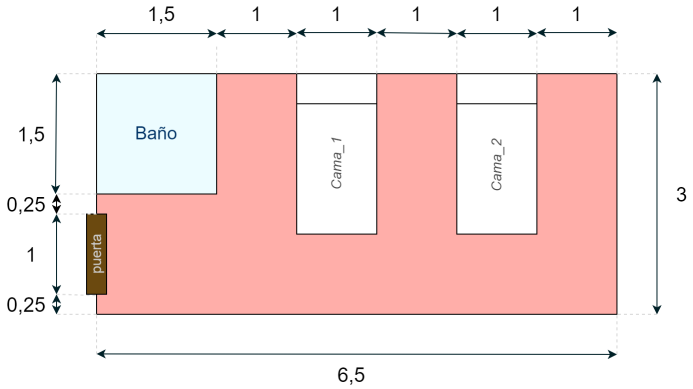


FIGURA A.1. Representación esquemática de una *Habitación* con sus dimensiones en metros.

A.3 Cama $\rightarrow^{situadoEn}$ Habitación

Consideramos esta relación como la distancia que hay desde los pies de la cama hasta la puerta. La calculamos como la media de la distancia de cada cama:

- $Cama\_1 : 1,5 + 1 = 2,5$
- $Cama\_2 : 1,5 + 1 \cdot 3 = 4,5$
- $Media : 3,5$

Esta distancia se corresponde con el recorrido completo $Cama \rightarrow Habitación \rightarrow Cama$, por lo que el valor de la propiedad *coste* (α) es:

$$2\alpha = 3,5 \quad \rightarrow \quad \alpha = 3,5/2 \quad \rightarrow \quad \alpha = 1,75$$

A.4 Habitación $\rightarrow^{contiguo}$ Habitación

Las *Habitaciones* contiguas se disponen en espejo, como muestra la Figura A.2. La media de la distancia entre las puertas de dos *Habitaciones* es:

$$(1,5 + 4,5)/2 = 3$$

A.5 Habitación $\rightarrow^{opuesto}$ Habitación

Entre dos *Habitaciones* contrapuestas hay un pasillo de 2,5 metros de ancho, de modo que puedan pasar dos camas con algo de holgura.

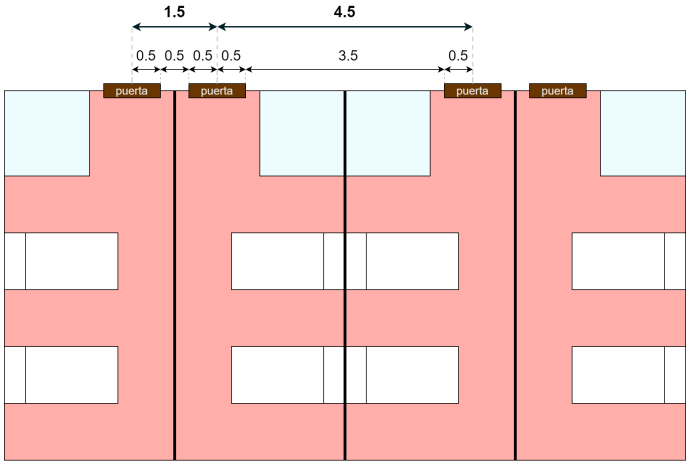


FIGURA A.2. Representación esquemática de cuatro habitaciones contiguas con sus dimensiones en metros.

A.6 Habitación $\rightarrow^{situadoEn}$ Pasillo

El cálculo de la distancia entre dos *Habitaciones* cualesquiera situadas en el mismo *Pasillo* se basa en la media de la distancia máxima en base a aristas *contiguo* entre dos *Habitaciones* que estén en el mismo *Pasillo* y lado de este. Nótese que tenemos en cuenta la organización del hospital.

$$(3 \times 10 \ + \ 6 \times 40 \ + \ 9 \times 15 \ + \ 12 \times 15 \ + \ 15 \times 20)/100 \ = \ 885/100 \ = \ 8,85 \ \approx \ 9$$

Esta distancia se corresponde con el recorrido completo *Habitación* $\rightarrow$ *Pasillo* $\rightarrow$ *Habitación*, por lo que el valor de la propiedad *coste* (α) es:

$$2\alpha = 9 \quad \rightarrow \quad \alpha = 9/2 \quad \rightarrow \quad \alpha = 4,5$$

TABLA A.1
Distancia entre las dos Habitaciones más alejadas del mismo Pasillo que estén en el mismo lado Pasillo . La columna proporción se refiere a cuántos Pasillos del tipo indicado en la columna de la izquierda hay (es decir, pasillos que en cada lado tienen las Habitaciones indicadas).

Nº *Habitaciones* consecutivas	**Proporción**	**Nº de aristas *contiguo* atravesadas entre las *Habitaciones* más alejadas**	**Distancia entre las *Habitaciones* más alejadas**
2	10 %	1	$1 \times 3 = 3$
3	40 %	2	$2 \times 3 = 6$
4	15 %	3	$3 \times 3 = 9$
5	15 %	4	$4 \times 3 = 12$
6	20 %	5	$5 \times 3 = 15$

Nótese que la distancia entre dos *Habitaciones* situadas en un *Pasillo* con 2 o 3 *Habitaciones* se calcula recorriendo aristas del tipo *contiguo*.

A.7 Pasillo $\rightarrow^{contiguo}$ Pasillo

Si la distancia media entre dos *Habitaciones* situadas en el mismo *Pasillo* es 9, la distancia media entre una *Habitación* y otra *Habitación* situada en un pasillo contiguo es el doble. Es decir:

$$9 \times 2 = 18$$

Esta distancia se corresponde con el recorrido completo $Habitación \rightarrow Pasillo \rightarrow Área \rightarrow Pasillo \rightarrow Habitación$, por lo que el valor de la propiedad *coste* (α) es:

$$4,5 + 2 \cdot \alpha + 4,5 = 18 \quad \rightarrow \quad 2 \cdot \alpha + 9 = 18 \quad \rightarrow \quad \alpha = 9/2 \quad \rightarrow \quad \alpha = 4,5$$

A.8 Pasillo $\rightarrow^{situadoEn}$ Área

El hospital está organizado de manera que todo *Pasillo* de un *Área* es vecino de todos los otros *Pasillo* que pueda haber en el mismo *Área*. Por tanto, el *coste* de ir de un *Pasillo* a otro *Pasillo* dentro del mismo *Área* es siempre el mismo. Es decir, este peso es el mismo que el calculado para dos *Pasillos* contiguos: 18.

A.9 Área $\rightarrow^{contiguo}$ Área

La distancia entre dos *Áreas* (sean contiguas o no) se calcula en base al número medio de *Pasillos* a recorrer para llegar de un *Pasillo* del *Área* origen a un *Pasillo* del *Área* destino. Llamamos "*salto*" al número de *Pasillos* a atravesar.

En la Figura A.3 mostramos cuántos saltos hay que dar desde un *Pasillo* "interior" (Caso 1) y otro exterior (Caso 2) de un *Área* hacia cualquier otro *Pasillo* en la misma *Planta*. Las *Áreas* en rosa son las *Áreas* contiguas al *Área* de origen.

Calculamos la distancia entre dos *Áreas* contiguas en base al promedio de saltos que hay que dar para llegar desde un *Pasillo* de un área hasta un *Pasillo* del *Área* contigua. Esta distancia es totalmente dependiente del tamaño de las cuadrículas de las plantas del hospital. En nuestro experimento, las tres *Plantas* para hospitalizaciones tienen 2 *Unidades* y 4 *Bloques*, como las *Plantas* de la figura A.3. No tenemos en cuenta el número y disposición concretos de los *Pasillos* de cada *Área*, sino que consideramos que todas las *Áreas* tienen 4 *Pasillos* dispuestos en cruz como en la figura A.3.

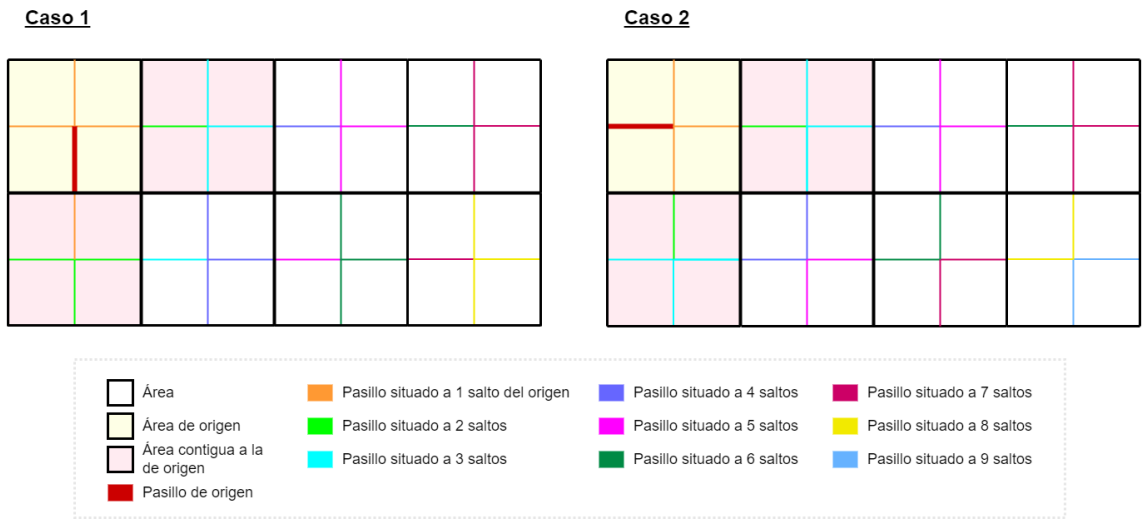


FIGURA A.3. Representación esquemática de la *Planta* de un hospital con 6 *Áreas* dispuestas en 2 *Unidades* (filas) y 4 *Bloques* (columnas). Los *Pasillos* están marcados en diferentes colores según el número de saltos que hay que dar desde el *Pasillo de origen* (marcado con una línea roja de mayor grosor) hasta ellos como *Pasillo de destino*. El *Pasillo de destino* se incluye en la cuenta del número de saltos. En el Caso 1 se parte desde un *Pasillo* "interior", y en el Caso 2 se parte desde un *Pasillo* "exterior".

TABLA A.2
Distancia entre dos Pasillos pertenecientes a dos Áreas contiguas.

Nº Saltos	**Proporción**	**Saltos × Coste × Proporción**
1	$(1/16) \times 100 = 6,25\,\%$	$1 \times 9 \times 0,0625 = 0,56$
2	$(6/16) \times 100 = 37,5\,\%$	$2 \times 9 \times 0,375 = 6,75$
3	$(9/16) \times 100 = 56,25\,\%$	$3 \times 9 \times 0,5625 = 15,19$

Para el cálculo de la distancia entre *Áreas* contiguas consideramos como *Pasillos destino* aquellos *Pasillos* que estén en las *Áreas* marcadas en rosa en la Figura A.3. En la Tabla A.2 está la base del cálculo para el *coste* de la arista *contiguo* entre *Áreas*, cuya suma es $22,5$.

Esta distancia se corresponde con el recorrido completo *Pasillo* $\rightarrow$ *Área* $\rightarrow^{contiguo}$ *Área* $\rightarrow$ *Pasillo*, por lo que el valor de la propiedad *coste* (α) es:

$$4,5 + \alpha + 4,5 = 22,5 \quad \rightarrow \quad \alpha + 9 = 22,5 \quad \rightarrow \quad \alpha = 13,5$$

A.10 Área $\rightarrow^{situadoEn}$ Planta

El cálculo del *coste* entre dos *Áreas* no contiguas de una misma *Planta* se realiza de manera similar al de dos *Áreas* contiguas. Es decir, en base al número medio de saltos entre sus Pasillos. Para este cálculo consideramos los *Pasillos* de todas las *Áreas* de la *Planta* salvo los del *Área de origen*.

En la Tabla A.3 está la base del cálculo para el *coste* de la arista *situadoEn* entre un *Área* y una *Planta*, cuya suma es $44,88$ (de forma aproximada, 45).

Esta distancia se corresponde con el recorrido completo *Pasillo* $\rightarrow$ *Área* $\rightarrow$ *Planta* $\rightarrow$ *Área* $\rightarrow$ *Pasillo*, por lo que el valor de la propiedad *coste* (α) es:

$$4,5 + 2 \cdot \alpha + 4,5 = 45 \quad \rightarrow \quad 2 \cdot \alpha + 9 = 45 \quad \rightarrow \quad \alpha = 36/2 \quad \rightarrow \quad \alpha = 18$$

Tabla A.3
Distancia entre dos Pasillos pertenecientes a dos Áreas no contiguas.

Nº Saltos	**Proporción**	**Saltos × Coste × Proporción**
1	$(1/56) \times 100 = 1,8\,\%$	$1 \times 9 \times 0,018 = 0,16$
2	$(6/56) \times 100 = 10,6\,\%$	$2 \times 9 \times 0,106 = 1,91$
3	$(10/56) \times 100 = 18\,\%$	$3 \times 9 \times 0,18 = 4,86$
4	$(7/56) \times 100 = 12,5\,\%$	$4 \times 9 \times 0,125 = 4,5$
5	$(9/56) \times 100 = 16\,\%$	$5 \times 9 \times 0,16 = 7,2$
6	$(7/56) \times 100 = 12,5\,\%$	$6 \times 9 \times 0,125 = 6,75$
7	$(9/56) \times 100 = 16\,\%$	$7 \times 9 \times 0,16 = 10,1$
8	$(5/56) \times 100 = 9\,\%$	$8 \times 9 \times 0,09 = 6,48$
9	$(2/56) \times 100 = 3,6\,\%$	$9 \times 9 \times 0,036 = 2,92$

A.11 Planta $\rightarrow^{situadoEn}$ Edificio

Ir de una *Planta* a otra *Planta* supone, en la práctica, recorrer la *Planta de origen* y recorrer la *Planta de destino*, tal y como se muestra en la Figura A.4. Es decir, ir de un *Pasillo* cualquiera de la *Planta de origen* a un *Pasillo* cualquiera de la *Planta de destino*.

Basándonos en el cálculo del *coste* ya realizado para las *Áreas* y *Plantas*, el *coste* está basado en el siguiente recorrido: $(Pasillo \rightarrow Área \rightarrow Planta \rightarrow Área \rightarrow Pasillo) \times 2$

Por tanto, tenemos el siguiente *coste*:

$$(4,5 + 18 + 2 \cdot \alpha + 18 + 4,5) = 45 \times 2 \quad \rightarrow \quad 45 + 2 \cdot \alpha = 90 \quad \rightarrow \quad \alpha = 45/2 \quad \rightarrow \quad \alpha = 22,5$$

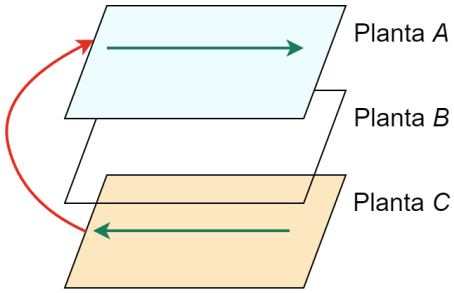


FIGURA A.4. Representación esquemática del camino a recorrer entre dos *Plantas*.

A.12 ZonaLógica $\rightarrow^{tieneÁrea}$ Área

Queremos que la distancia entre dos *Áreas* que pertenecen a la misma *ZonaLógica* sea significativamente menor a la distancia entre dos *Áreas* cualesquiera de la misma *Planta*. Para ello, vamos a dividir el *coste* de la relación *Área* $\rightarrow^{situadoEn}$ *Planta* entre 2. Esto da un valor para la propiedad *coste* de 9.

A.13 Aclaraciones

Como hemos podido observar a lo largo del apéndice, los valores de la propiedad *coste* para cada tipo de arista y clases de nodos no son independientes. Están pensados para cuantificar la distancia que hay de un nodo *Localización* a otro en función de los nodos de clases de nivel inferior de la jerarquía que hay que atravesar. Teniendo en cuenta que la distancia entre dos *Localizaciones* se calcula mediante un algoritmo de búsqueda del camino más corto en un grafo, podemos ver el valor de la propiedad *coste* como "el aumento que hay que sumar al *coste* del camino en el paso actual para conseguir el *coste* deseado".

El hecho de utilizar un algoritmo de búsqueda del camino más corto conlleva que no para todos los pares de *Localizaciones* (en el caso del experimento del Capítulo 4, *Lechos* o *Camas*) este camino más corto recorra necesariamente los caminos aquí descritos, sino que es posible que encuentre "atajos" a través de la concatenación de aristas *contiguo* y *opuesto*. Por ejemplo, dadas dos *Camas* en dos Áreas no contiguas, el camino según la jerarquía de *Localizaciones* es:

$$Cama \rightarrow^{situadoEn} Habitación \rightarrow^{situadoEn} Pasillo \rightarrow^{situadoEn} Área \rightarrow^{situadoEn} Planta^{situadoEn} \leftarrow Área^{situadoEn} \leftarrow Pasillo^{situadoEn} \leftarrow Habitación^{situadoEn} \leftarrow Cama$$

La distancia total de este camino es: $(1,75+4,5+4,5+18)\times 2 = 57,5$. Sin embargo, es posible que los *Pasillos* de esas Áreas no tengan mucha longitud y mediante varias aristas *contiguo* entre *Habitaciones* se llegue hasta la *Habitación* con la *Cama de destino*; o que los *Pasillos* en los que estén las *Camas* sean los extremos más cercanos de ambas *Áreas* y el camino más corto se define mediante aristas *contiguo* entre *Pasillos*;

o también puede ser una combinación de ambas opciones. Es en este tipo de ejemplos donde se puede resaltar el valor que aporta la semántica definida mediante la jerarquía de *Localizaciones* y los tipos de relaciones que hemos añadido extra, ya que permiten establecer diferentes niveles de proximidad aun estando en una misma *Localización* amplia de nivel superior.