

Universitat Rovira i Virgili
Escola Tècnica Superior de Enginyeria
Tesis para la obtención del grado de:
Màster en Enginyeria Informàtica i Seguretat

**Estudio de métodos para proteger y explotar datos
de seguimientos de turistas.**

Efraín Ricardo Sosa Gámez

Tarragona 5 de Junio del 2009

A mis padres con mucho cariño, por la oportunidad de esta experiencia.

A mis hermanos y hermanas por el apoyo en todo momento.

A Gabriela, por estar siempre conmigo.

A mi asesora la Dra. Aïda Vall's por su colaboración y asesoría en todos los aspectos de este proyecto.

Contenido

1. Introducción.	6
1.1 Minería de Datos (Data Mining).....	6
1.1.1 Datos espaciales.	8
1.1.2 Datos temporales.	9
1.1.3 Datos secuenciales.	9
1.2 Privacidad.	10
1.3 Turismo.....	11
1.4 Proyecto URV-Turisme-02.	12
1.5 Objetivos de este trabajo de máster.....	13
1.5 Estructura del documento.....	13
2. Los datos de localización y seguimiento de individuos.	15
2.1 Introducción.	15
2.2 Trayectorias.....	17
2.3 Espacio.....	19
2.4 Tiempo.....	21
2.5 Entidades en movimiento y sus actividades.	22
3. Anonimización de Trayectorias.	24
3.1 Introducción.	24
3.2 Métodos de Anonimización de Trayectorias.....	25
3.2.1 Protegiendo la privacidad de la localización a través de confusión de trayectorias.	26
3.2.2 Ofuscación de localizaciones.....	31
3.2.3 Modificación de proyecciones.	32
3.2.4 Anonimización por generalización.	35
3.3 Aplicaciones y usos en Turismo.....	40
3.4 Grupos de investigación	44
4. Generación de perfiles en datos secuenciales.	46
4.1 Data Mining espacio-temporal.....	46
4.2 Clustering	48
4.2.1 Clustering basado en distancias de trayectorias.	48
4.2.2 Clustering de Trayectorias específicas.	50
4.3 Patrones Locales Espacio-Temporales.	53

4.3.1 Extrayendo patrones frecuentes.....	54
4.3.2 Recuperación de ocurrencia.....	57
4.4 Predicción.....	58
4.4.1 Predicción de Ubicaciones y Trayectorias.....	58
4.4.2 Predicción de Densidad.....	59
4.4.3 Extrapolación y Predicción de Alcance.....	60
4.4.4 Predicción de Eventos.....	60
4.4.5 Predicción en Series de Tiempo Geo Referenciadas.....	60
4.4.6 Clasificación de Trayectorias.....	61
4.5 El papel de la incertidumbre en la Minería de Datos Espacio-Temporales.....	62
5. Análisis de los estudios realizados.....	63
5.1 Análisis sobre los métodos de anonimización.....	63
5.2 Análisis sobre los métodos de minería de datos.....	66
5.3 Otras consideraciones.....	67
6. Realización del trabajo dentro del proyecto de investigación.....	68
7. Conclusiones y contribuciones del trabajo.....	69
7. Conclusiones y contribuciones del trabajo.....	69
Bibliografía.....	71

Tabla de Ilustraciones

Ilustración 1. Ejemplo de un "árbol r"	8
Ilustración 2. Cubo de espacio y tiempo	15
Ilustración 3. Prisma espacio-temporal.....	16
Ilustración 4. Base de datos de Trayectorias.....	24
Ilustración 5. Dos usuarios moviéndose en paralelo.	29
Ilustración 6. Distancia preservando la ofuscación de dos posiciones $p(e1,t)$ y $p(e2,t)$	32
Ilustración 7. Anonimización de dos trayectorias.	33
Ilustración 8. Trayectorias originales, anonimizadas y reconstruidas respectivamente.	41
Ilustración 9. Mapa de la ciudad con 3 grupos, cada uno con 2 trayectorias.....	41
Ilustración 10. LCM para anonimización - Datos Milán.....	41
Ilustración 11. Resultados del Agrupamiento - Datos Milán.....	42
Ilustración 12. Resultados del Agrupamiento - Datos sintéticos.	43
Ilustración 13. Trayectorias en coordenadas espacio temporales.....	49
Ilustración 14. Agrupaciones sobre un intervalo de tiempo.....	52
Ilustración 15. Una agrupación en movimiento.....	52
Ilustración 16. Simple trayectoria (a), Eventos derivados espacio temporales (b), Variación de la velocidad (c).	54
Ilustración 17. Ejemplo básico de patrones espacio-temporales.	56
Ilustración 18. Liderazgo de patrones de movimiento.	58

Tablas

Tabla 1 Grupos de Investigación.	45
Tabla 2. Métodos de Anonimización vs Información Espacio-Temporal	64
Tabla 3. Características de la protección de la privacidad.	65

1. Introducción.

El estudio de movilidad de las personas es un área de investigación emergente que plantea retos importantes a la comunidad científica.

Las nuevas tecnologías permiten obtener información precisa de las secuencias de acciones realizadas por una persona. El estudio de este tipo de datos se prevé que será muy importante porque permite que se puedan diseñar y desarrollar nuevos tipos de aplicaciones avanzadas.

En los últimos años el sector turístico es una industria que ha incorporado las nuevas tecnologías en diversos ámbitos. En este dominio resulta interesante disponer de datos de seguimiento y localización de turistas para poder proporcionar servicios nuevos y avanzados. Por ejemplo el seguimiento de un grupo de personas (mediante datos espacio - temporales) conlleva a poder obtener secuencias de datos que contienen información detallada sobre el comportamiento personal o vehicular del individuo, con lo que se pueden recomendar rutas personalizadas, realizar recomendaciones basadas en el itinerario seguido por el turista, planificar y gestionar el flujo de turistas en una zona concreta, etc.

Para poder desarrollar este tipo de aplicaciones, se debe estudiar cómo obtener los datos de seguimiento y localización de turistas (tracking) usando la tecnología ya existente (Dispositivos GPS, Teléfonos móviles).

Debido a las grandes cantidades de información que se puede obtener utilizando estos dispositivos, se requiere métodos automáticos para el tratamiento y análisis de los datos. En el área de investigación de Inteligencia Artificial se trata este problema, estos métodos se conocen como técnicas de minería de datos (*Data Mining*). Recientemente se están desarrollando métodos para el análisis de secuencias de datos, denominados *Temporal Data Mining*.

No obstante el seguimiento y localización de los turistas atenta directamente contra su privacidad, por lo tanto es necesario que los sistemas que ofrezcan estos servicios incorporen mecanismos para proteger la privacidad de los usuarios, como por ejemplo anonimizar los datos sensibles obtenidos de los usuarios.

A continuación se introducen los conceptos básicos de minería de datos y de protección de la privacidad en bases de datos, para pasar a explicar los objetivos de este trabajo de máster, que forma parte de un proyecto de investigación financiado por la Universidad Rovira i Virgili, en colaboración con el Parque Científico y Tecnológico de Turismo y Ocio de Tarragona.

1.1 Minería de Datos (Data Mining).

La minería de datos se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semi-automático (asistido) para permitir tratar grandes volúmenes de datos, además el conocimiento descubierto debe ayudar a tomar mejores decisiones que reporten, por tanto, algún beneficio a la organización. (Hernandez, Ramírez, & Ferri, 2004).

Por lo tanto dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos...), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil.

En muchos casos la utilidad del conocimiento está íntimamente relacionada con la comprensibilidad del modelo inferido. No debemos olvidar que, generalmente, el usuario final no tiene por qué ser experto en las técnicas de minería de datos, ni tampoco puede perder mucho tiempo interpretando los resultados. Por ello, en muchas aplicaciones es importante hacer que la información descubierta sea más comprensible por los humanos (por ejemplo, usando representaciones gráficas, convirtiendo los patrones a lenguaje natural o utilizando técnicas de visualización de los datos).

De una manera simplista pero ambiciosa, podríamos decir que el objetivo de la minería de datos es convertir datos en conocimiento.

Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación. (Hernández et al., 2004)

¿A qué tipos de datos puede aplicarse la minería de datos? En principio, puede aplicarse a cualquier tipo de información, siendo las técnicas de minería diferentes para cada una de ellas.

El proceso global de análisis de datos, es el proceso de aplicar a una determinada base de datos las operaciones requeridas de selección, exploración, muestreo, transformación y métodos de modelado para extraer los patrones y posteriormente evaluarlos para identificar el conjunto de ellos que representarán el conocimiento u objetivo final de la necesidad del usuario. Como todo proceso, representa una conjunción de pasos: (Pérez, 2006)

1. **Selección del conjunto de datos**, tanto en lo que se refiere a las variables dependientes, como a las variables objetivo, como posiblemente al muestreo de los registros disponibles.
2. **Análisis de las propiedades de los datos**, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
3. **Transformación del conjunto de datos de entrada**, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema.
4. **Seleccionar y aplicar la técnica de minería de datos**, se construye el modelo de predicción, de clasificación o segmentación.
5. **Evaluar los resultados** contrastándolos con un conjunto de datos previamente reservado para validar la generalidad del modelo.

En este trabajo nos centraremos en estudiar las técnicas de minería de datos que existen para datos secuenciales espacio – temporales.

1.1.1 Datos espaciales.

Una base de datos espacial es una base de datos que contiene datos pertenecientes a un determinado espacio, ejemplos de bases de datos espaciales son: una base de datos sobre los inmuebles de una ciudad, las habitaciones de un hotel etc. Un concepto clave en las bases de datos espaciales es, lógicamente, la dimensión espacio. Esta dimensión establece el marco de referencia donde ubicar los datos. En las bases de datos espaciales normalmente se requiere de técnicas para tratar con una gran cantidad de datos, así como de métodos específicos para tratar datos espaciales.

Un sistema de información espacial o geográfica SIG (en inglés GIS Geographic Information Systems) es un conjunto de métodos, herramientas y datos que permiten capturar, almacenar, analizar, transformar y presentar toda la información geográfica y de sus atributos almacenada en una base de datos espacial.

Dadas las características especiales de este tipo de información se han desarrollado estructuras de datos que permiten un mejor almacenamiento y tratamiento. Una de las estructuras más conocidas y utilizadas para almacenar información con estructura espacial son los "árboles r" (*r-trees*, de *region trees*). (Guttman, 1984) Este tipo de árboles permiten que los datos sean almacenados en la base de datos con respecto a la posición que ocupan en el espacio real. En los árboles r las hojas están formadas por punteros a los datos, mientras que los nodos intermedios son los rectángulos mínimos que contienen todos sus sub-nodos y los punteros a los nodos hijos. Todas las hojas están al mismo nivel del árbol.

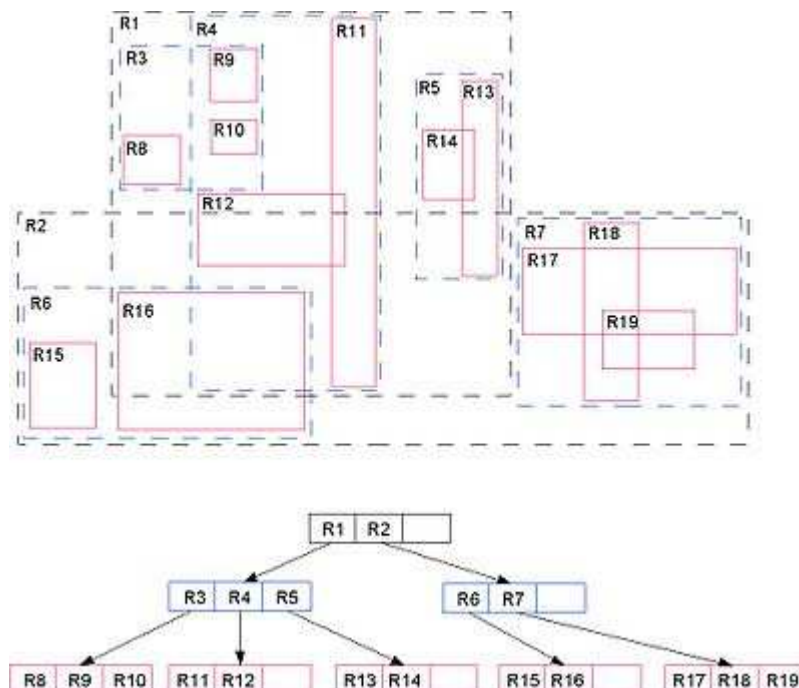


Ilustración 1. Ejemplo de un "árbol r"

1.1.2 Datos temporales.

Un tipo de datos que ha despertado interés especial en su investigación por la gran cantidad de aplicaciones que pueden derivarse son los datos que tienen componente temporal. Muchas bases de datos están formadas por series con observaciones de carácter cronológico que normalmente se realizan de forma repetida y con la misma frecuencia. Este tipo de series se denominan series temporales. Existen muchos ejemplos de series temporales, por ejemplo, la evolución diaria de la cotización en bolsa de un determinado valor, el número de vehículos que han transitado por una determinada carretera medidos hora a hora y por qué no el número de turistas que visitan un punto de interés turístico a ciertas horas.

Existen dos grandes objetivos que han impulsado el estudio de las series temporales: identificar la naturaleza del sistema que genera la secuencia de los datos, y predecir los valores futuros que tomará la serie temporal. Para conocer el comportamiento que tiene una serie temporal normalmente se descompone en cuatro elementos o movimientos principales. (Han & Kamber, 2001)

- **Movimiento a largo término o tendencias:** estos movimientos indican el comportamiento general de la serie en un período largo de tiempo. Ayudan a identificar cuál es la tendencia que sigue o ha seguido la serie.
- **Variaciones cíclicas:** representan ciclos que presentan las series, pueden o no ser periódicas. Es decir, los ciclos pueden no ser completamente iguales después de periodos de tiempos idénticos.
- **Movimientos estacionales:** estos movimientos se deben a eventos que ocurren con una frecuencia establecida y constante.
- **Movimientos aleatorios o irregulares:** estos movimientos representan el comportamiento de la serie debido a eventos aleatorios o semi- aleatorios.

1.1.3 Datos secuenciales.

En un contexto de datos secuenciales, los datos de entrenamiento representan una secuencia de ejemplos donde datos contiguos presentan algún tipo de relación. La minería de datos secuenciales se define como la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. A diferencia de las series temporales el momento preciso no es tan relevante, sino que estos eventos se producen secuencialmente. Además, el objetivo no suele ser cómo seguir a una serie sino analizar muchos individuos que tienen comportamiento secuencial.

Gran cantidad de datos pueden considerarse como datos secuenciales: transacciones comerciales, acceso a sitios web por un cliente, recorrido de un cliente por las secciones de un supermercado, transacciones comerciales, secuencias de proteínas, etc.

Las técnicas más habituales para datos secuenciales son el agrupamiento de secuencias y las reglas de asociación.

Agrupamiento de patrones secuenciales. La tarea de agrupamiento se define como la tarea de separar en grupos a los datos, de manera que los miembros de un mismo grupo sean muy similares entre sí, y al mismo tiempo, sean diferentes a los objetos de otros grupos. Para el caso de datos secuenciales, la tarea de agrupamiento se convierte en la búsqueda de grupos de secuencias con alta cohesión. Utilizar técnicas de agrupamiento para identificar grupos

significativos ayuda a identificar patrones interesantes, lo que puede dar lugar a aplicaciones novedosas, como se comentará más adelante.

Reglas de asociación con datos secuenciales. Existe gran interés desde los últimos años en el desarrollo de algoritmos de aprendizaje de reglas de asociación desde datos secuenciales. La motivación es la gran cantidad de aplicaciones prácticas que tienen estas técnicas.

1.2 Privacidad.

La **privacidad** puede ser definida como el ámbito de la vida personal de un individuo que se desarrolla en un espacio reservado y debe mantenerse confidencial (RAE, 2005).

Los límites y el contenido de lo que se considera privado difieren entre las culturas y los individuos, pero comparten temas comunes básicos. La privacidad es a veces relacionada con el anonimato, el deseo de permanecer inadvertidos o no identificados en la esfera pública. Cuando algo es privado de una persona, generalmente significa que hay algo en ellos que se considera intrínsecamente especiales o personalmente sensibles. El grado en que la información privada está expuesta, por lo tanto, depende de la forma en que el público recibirá esta información.

El derecho a la privacidad es un derecho que les corresponde por excelencia a todos los seres humanos, incluso desde su nacimiento, y esto es un punto que todos los doctrinarios comparten. Todos y cada uno de nosotros nacemos con el derecho de que sea protegida por el ordenamiento jurídico esa esfera de nuestra vida que compone todos los datos y acontecimientos que conforman nuestra vida privada.

Privacidad Informacional.

La necesidad de privacidad informacional o privacidad de los datos está relacionada con la evolución de la tecnología y las expectativas públicas de la intimidad en la recogida y difusión de datos acerca de individuos. Se debe proteger la intimidad en datos de identificación relativos a una persona o personas que están almacenados en formato digital o de otro tipo. En algunos casos, esas preocupaciones se refieren a cómo los datos se recogen, almacenan, y se asocian. En otros casos, la cuestión es que se da acceso a la información.

Por diversas razones, los individuos pueden no desear que información personal como su religión, orientación sexual, afiliación política, personal o actividades sean expuestas. Esto puede ser para evitar la discriminación, la vergüenza personal, o daños a su reputación profesional.

En el campo informático y sus implicaciones en sistemas de información, se trata información confidencial de individuos la cual debe ampararse dentro del marco de la privacidad, como indica la Declaración Universal de los Derechos Humanos (AGNU, 1948).

Bajo esta concepción se han desarrollado métodos para la “protección de datos” o “protección de la privacidad”.

La privacidad es un tema que muchas veces no se le da la atención que requiere al principio, es hasta que se ve uno envuelto en un problema de privacidad cuando en verdad se le empieza a poner atención a dicho tema.

Para darnos una idea de lo que esto refiere citemos el caso de la nueva tecnología de Google: Google Street View, es una característica de Google Maps y de Google Earth que proporciona imágenes esféricas a nivel de calle, permitiendo a los usuarios ver partes de las ciudades seleccionadas y sus áreas metropolitanas circundantes. Suena muy bien, en principio es una herramienta bastante útil si así lo queremos ver pero mucha gente ha empezado a notar que hay un aspecto muy importante que Google no tomó en cuenta al principio: La privacidad, cuando el servicio se puso en marcha la gente empezó a notar que había sido captada por las cámaras de Google y es cuando empezaron los problemas, hubo personas que fueron captadas en lugares donde no se suponía que debían estar o en situaciones bochornosas y ha Google le han empezado a llover las quejas e inclusive demandas de personas que alegan les fue violada su privacidad al ser captadas por este servicio.

En este caso, los datos son imágenes, pero ocurre lo mismo si se toman datos de otro tipo, por ejemplo, a partir del uso del móvil o de la tarjeta bancaria.

Por tanto, podemos afirmar la importancia que tiene la privacidad, por eso mismo las nuevas tecnologías por desarrollarse deben tomar en cuenta esto para evitar la diseminación de datos confidenciales o sensibles de las personas.

1.3 Turismo.

La aplicación en el sector turístico de las tecnologías de la información y comunicaciones ha tenido una evolución acorde con las etapas de informatización e información de la sociedad. Dos hechos permiten pensar en una estrecha interrelación entre la industria turística y la informática, por un lado el hecho de que el turismo al ser la primera industria mundial se convierte, potencialmente en el principal cliente de la industria informática, y por otro lado el hecho de que la información está en los mismos cimientos de la industria turística. (Lacramioara et al., 2007)

Por ejemplo, hoy en día es imposible imaginar que llegamos a una ciudad nueva y no sabemos nada de ella. Antes de iniciar el viaje, el sector turístico ha puesto a nuestra disposición gran cantidad de recursos que nos permiten obtener información de nuestro destino. Por ejemplo, a través de páginas Web. Pero además, en los últimos años, gracias a nuevas formas de difusión de información, existen métodos que nos permiten obtener información on-line del sitio que estamos visitando, por ejemplo a través de bluetooth. Cada vez son más las ciudades que tienen disponible información turística que nos puede ser enviada a un dispositivo móvil a través de esta tecnología.

Así pues, con los avances tecnológicos que existen, el abanico de oportunidades que existen para poder mejorar la estancia de los turistas en sus destinos es muy amplio. En este sentido, el sector turístico está especialmente interesado en nuevas tecnologías que permitan ofrecer servicios personalizados a turistas utilizando su localización (Hawking et al, 2005; Schmidt-Belz et al., 2003; Tchetchik et al., 2009).

Conocer la localización o el recorrido de un turista nos permitiría ofrecer información relativa al lugar donde se encuentran, tal como, servicios cercanos, estaciones de policía, servicios de emergencia, bancos, restaurantes, sitios de interés turístico, hoteles y sitios de descanso, espacios recreativos y lugares de ocio. Además podemos desarrollar información personalizada de acuerdo a los gustos de cada viajero reuniendo información que permita generar un perfil

del usuario y nos permita sugerir puntos de interés o sitios para realizar actividades específicas, todo este tipo de servicios ya está empezando a ofrecerse por ejemplo en dispositivos GPS, hay compañías que ya desarrollaron los servicios de algunas ciudades y cada vez son más las que se agregan, esto es lo que ya existe pero debemos ser conscientes que lo que está por venir debe ser mejor, la inversión en este par de sectores (turismo y tecnología) cada vez es mayor por lo que debemos saber que en la actualidad este es un tema de mucho interés y de muchas oportunidades.

1.4 Proyecto URV-Turisme-02.

El Parque Científico y Tecnológico del Turismo y el Ocio (PCTTO) de Tarragona es una organización que surge de la oportunidad de colaboración entre los agentes públicos y privados de una de las zonas del litoral mediterráneo con una mayor orientación al negocio del turismo: la Costa Dorada.

El PCTTO tiene como misión llegar a ser un polo de "*knowledge-economy*" de turismo con una proyección internacional, así como generar un entorno ideal para la inversión turística en I+D+I en los temas estratégicos para este sector.

Esta entidad ha concedido, conjuntamente con la Universidad Rovira i Virgili, un proyecto de investigación sobre los temas planteados en las secciones anteriores, titulado "Servicios Avanzados para turistas basados en su localización o seguimiento", en el que participan tres grupos de investigación de la URV: ITAKA, CRISES y GRATET.

El objetivo principal es estudiar el estado actual de provisión de servicios personalizados a turistas, a partir de información de su localización o su seguimiento, así como de otras variables que pudieran conocerse.

Este objetivo puede dividirse en sub objetivos.

1.-Tecnologías.

Estudiar los proyectos y las tecnologías existentes para realizar una localización de turistas.

2.-Tipos de Servicios.

Estudiar los diferentes tipos de servicios que se pueden ofrecer a partir de la localización y/o seguimiento de turistas.

3.-Técnicas de privacidad.

Estudiar cuales técnicas para garantizar la privacidad de los turistas se están desarrollando y en qué casos se aplican; en función de los servicios a ofrecer o en función del origen de los datos. Se pueden considerar orígenes en entornos cerrados (dentro de un parque temático), en una zona concreta del territorio o directamente de todos los turistas que visitan el país.

4.-Técnicas de Análisis inteligente.

Estudiar cuales técnicas de análisis inteligente de datos se están desarrollando y aplicando para modelar el comportamiento de los turistas y generar perfiles (clustering).

5.-Nuevas tendencias.

Estudiar cuales son las nuevas tendencias en los campos de la privacidad, de la obtención del comportamiento de los turistas y de la provisión de nuevos servicios avanzados que ofrecen las nuevas tecnologías. Esto debe servir para identificar las líneas de investigación que pueden surgir en este ámbito.

1.5 Objetivos de este trabajo de máster.

El trabajo de investigación que se presenta en este trabajo de máster es parte del proyecto URV-Turisme02. Por ello, los objetivos de este trabajo están relacionados con los objetivos detallados en la sección anterior. Concretamente, en este trabajo de máster se han abordado los siguientes objetivos:

- Estudio en profundidad del estado del arte de los métodos que se están desarrollando para la anonimización de los datos de seguimiento de itinerarios de personas (*Tracking*). Identificando y analizando métodos concretos de anonimización, así como, grupos de investigación líderes en este ámbito.
- Estudio preliminar del estado del arte en técnicas de inteligencia artificial para el análisis de datos espacio-temporales. Con especial interés en métodos de generación de perfiles de usuario. Por ejemplo, técnicas basadas en construcción automática de grupos de usuarios con características comunes (*Clustering*).

El primer objetivo forma parte de los objetivos 3 y 5 del proyecto, que incluye además el estudio de los métodos para protección de localizaciones en servicios LSB (*Location based Services*). El segundo objetivo corresponde al objetivo 4 del proyecto.

Puesto que este trabajo forma parte de un proyecto más grande en el que participan diversos grupos, a nivel más organizativo, un objetivo ha sido la realización y documentación de estos estudios de la literatura de forma colaborativa con el resto del equipo del proyecto.

1.5 Estructura del documento.

Este documento se estructura de la siguiente forma. En primer lugar el capítulo 2 define el tipo de datos objeto del estudio en este trabajo: las secuencias espacio-temporales. A continuación se exponen los resultados de los dos objetivos principales del proyecto. Por un lado, el capítulo 3 está dedicado a la anonimización de datos secuenciales. Por otro lado, el capítulo 4, presenta los métodos de minería de datos que se pueden usar para generación de perfiles a partir de secuencias espacio-temporales. Una vez presentado el estado del arte, el capítulo 5 analiza los resultados, identificando líneas de interés para la investigación en este tema, que pueden dar lugar a futuros proyectos financiados.

Después de presentar la parte más importante del trabajo, en el capítulo 6 se explica cómo se ha realizado el trabajo dentro del proyecto de investigación. Siguiendo el objetivo indicado en la sección anterior, se ha trabajado de forma que el material recopilado durante los estudios de la literatura estuviera disponible para el resto de miembros del equipo. Así mismo, se

explican los mecanismos para compartir los resultados parciales que se obtenían en los distintos grupos.

Finalmente el capítulo 7 expone las contribuciones principales de este trabajo y presenta las conclusiones.

2. Los datos de localización y seguimiento de individuos.

2.1 Introducción.

Desde tiempos antiguos, las personas han observado el movimiento de varias entidades, desde insectos y peces hasta planetas y estrellas, y se ha investigado la conducta de sus movimientos. Sin embargo los métodos de observación y seguimiento usados anteriormente son muy diferentes a los usados ahora con las nuevas tecnologías, aunque se puede aprender mucho de los estudios anteriores. Primero, está la exhaustiva atención puesta a los múltiples aspectos del movimiento, esto no solo incluye trayectorias en espacio, también características de movimientos tales como la velocidad y la dirección, y su dinámica sobre el tiempo, así como características y actividades de las entidades en movimiento. Segundo, está el esfuerzo para relacionar los movimientos a las propiedades que los rodean y a otros fenómenos o eventos.

En los estudios de desplazamientos (trayectorias) es importante tener en cuenta el contexto en que los desplazamientos se llevan a cabo y las características de las entidades (individuos) que se desplazan. Además, los desplazamientos como tal no son siempre el aspecto principal de estudio. Uno además puede analizar los desplazamientos con el objetivo de adquirir conocimiento sobre las entidades en movimiento o del ambiente de los movimientos. Así en el área conocida como geografía temporal, la observación de los movimientos cotidianos de individuos humanos es, primordialmente, el medio del estudio de las actividades en diferentes categorías de personas. A nivel global, la geografía temporal busca tendencias en la sociedad.

Las ideas de geografía temporal originadas por (Hagerstrand T., 1970), una característica destacada de la geografía temporal es la contemplación del espacio y tiempo como inseparables. La idea principal era considerar trayectorias espacio- temporales en un espacio tridimensional donde los ejes horizontales representaban espacio geográfico y los ejes verticales representaban el tiempo. Esta representación es conocida como **cubo de espacio y tiempo**.

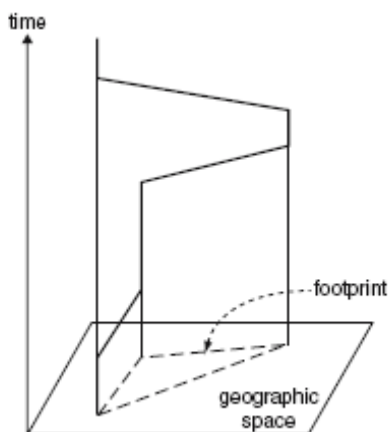


Ilustración 2. Cubo de espacio y tiempo

La línea representa los movimientos de una entidad, por ejemplo una persona que trabaja, primero está en casa, después se mueve al supermercado de compras y después de pasar un tiempo ahí, regresa a casa. Las líneas verticales representan las estancias en ciertas ubicaciones (hogar, lugar de trabajo, supermercado). Los segmentos de líneas inclinadas indican movimientos, cuánto más lento el movimiento más empinada la línea. La rectitud de

las líneas en la ilustración supone que la persona se desplaza con una velocidad constante que es usualmente solo una aproximación del comportamiento real. La trayectoria espacio-temporal puede ser proyectada en un mapa, resultando en la *huella digital* de la trayectoria.

Otro concepto importante en la geografía temporal es la noción del **prisma espacio-temporal**.

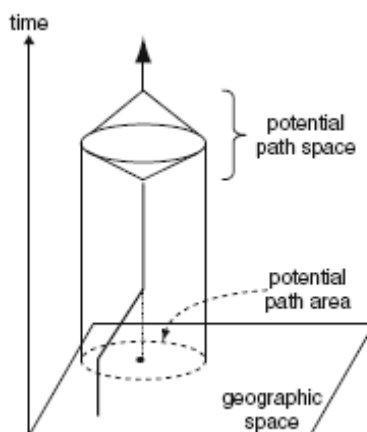


Ilustración 3. Prisma espacio-temporal

En la representación tridimensional, este es el volumen en espacio y tiempo que una persona puede alcanzar en un intervalo particular de tiempo, comenzando y terminando en la misma ubicación (por ejemplo, el lugar a donde una persona va a comer en su descanso de comida en el trabajo). La medida más amplia es llamada la trayectoria potencial de espacio y su huella digital es llamada área de trayectoria potencial. En la ilustración anterior, está representada con un círculo, asumiendo que sea posible alcanzar todas las ubicaciones dentro del círculo, en realidad el entorno físico no siempre permitirá esto. En general las trayectorias espacio temporales de individuos son influenciadas por ciertas limitaciones. Uno puede distinguir entre limitaciones de capacidad (por ejemplo, modo de transporte y necesidad de dormir), las limitaciones de acoplamiento (por ejemplo, estar trabajando o en el club deportivo) y limitaciones de autoridad (por ejemplo, accesibilidad a edificios o parques en espacio y tiempo).

En la era previa a los gráficos por computadora, se consumía demasiado tiempo y era demasiado caro producir visualizaciones de cubos espacio-temporales para apoyar la exploración del comportamiento de movimientos, sin embargo con el desarrollo de la nueva tecnología de visualización e interactividad los investigadores han vuelto a recuperar este concepto. Además, la geografía temporal moderna no está basada enteramente en representaciones visuales y descripciones cualitativas. Así en (Miller H., 2005) se sugiere una teoría de medición para entidades básicas y sus relaciones la cual incluye definiciones formales de conceptos básicos y características de movimientos.

En la siguiente sección se presenta una síntesis de los conceptos básicos y características de movimientos. Los desplazamientos suceden en espacio y tiempo así que se analizan también las propiedades de espacio y tiempo.

2.2 Trayectorias.

La definición de movimiento se refiere a la noción de cambiar de posición física una entidad con respecto a un algún sistema de referencia, en el cual cada uno asume posiciones conocidas. Frecuentemente el sistema referenciado es un espacio geográfico.

Una trayectoria es el trayecto o camino creado por la entidad en movimiento a través del espacio donde se desplaza. El trayecto nunca se crea instantáneamente dado que requiere una cierta cantidad de tiempo. Por lo tanto el tiempo es un aspecto inseparable de las trayectorias. Veamos si t_0 es el momento en que el trayecto comienza y t_{end} es el momento en que termina, para cualquier momento t_i entre t_0 y t_{end} hay una posición en el espacio que está siendo ocupada por una entidad en ese momento (aunque en la práctica esta posición no siempre es conocida). Por lo tanto una trayectoria puede ser vista como una función que empareja momentos de tiempo con momentos en espacio. También puede ser visto como una consistencia de pares (tiempo y ubicación), dado que el tiempo es continuo, hay un número infinito de dichos pares en una trayectoria. Por razones prácticas sin embargo, las trayectorias tienen que ser representadas por secuencias finitas de ubicaciones referenciadas en tiempo. Dichas secuencias pueden ser obtenidas de de varias formas que son usadas para observar movimientos y coleccionar datos de movimientos.

- Registro basado en tiempo: las posiciones de las entidades son registradas en momentos de tiempo en el espacio regulares (Por ejemplo, cada 5 minutos).
- Registro basado en cambios: un registro es realizado cuando la posición de la entidad difiere de la anterior.
- Registros basados en la localización: los registros son realizados cuando la entidad se aproxima a ubicaciones específicas (Por ejemplo, donde los sensores son instalados).
- Registros basados en eventos: las posiciones y el tiempo son registrados cuando ciertos eventos suceden, en particular, actividades realizadas por la entidad en movimiento (Por ejemplo, llamando desde un teléfono móvil).
- Varias combinaciones de estas aproximaciones básicas.

Típicamente, las posiciones son medidas con incertidumbre, algunas ocasiones es posible refinar las posiciones tomando en cuenta condicionantes físicas (Por ejemplo, la red de calles). En estudios del movimiento, un analista atiende a cierto número de características que pueden ser agrupadas dependiendo si refieren a estados en momentos individuales o a movimientos en intervalos de tiempo. Las características relacionadas al movimiento incluyen lo siguiente:

- Tiempo, posición del momento es la escala de tiempo.
- Posición de la entidad en el espacio.
- Dirección del movimiento de la entidad.
- Velocidad del movimiento, que es cero cuando la entidad está en un mismo lugar.
- Cambio de dirección.
- Cambio de velocidad.
- Tiempo de viaje y distancia acumulados.

Las características totales de una trayectoria completa o un fragmento de trayectoria realizadas durante un sub intervalo $[t_1, t_2]$ del lapso de tiempo completo $[t_0, t_{end}]$ incluyen lo siguiente:

- Forma geométrica de la trayectoria en el espacio.
- Distancia recorrida.

- Duración de la trayectoria.
- Vector de movimiento o dirección mayor.
- Velocidad media, mediana y máxima.
- Dinámica de la velocidad (comportamiento de la velocidad).
 - Periodos de velocidad constante, aceleración y desaceleración.
 - Características de estos periodos: tiempo de inicio y de final, duración, posición inicial y final, velocidad inicial y final.
 - Orden de estos periodos de tiempo.
- Dinámica de las direcciones (comportamiento de direcciones)
 - Vueltas mayores (puntos de vuelta) con sus características: tiempo, posición, ángulo, dirección inicial y final y velocidad del movimiento al momento de la vuelta.
 - Periodos de movimientos rectos, curvos y circulares.
 - Características de estos periodos: tiempo de inicio y fin, posición inicial y final, direcciones, dirección mayor, ángulos y radio de las curvas.
 - Orden de los periodos y vueltas en tiempo.

Además de examinar una simple trayectoria, un analista está típicamente interesado en la comparación de dos o más trayectorias. Estas pueden ser trayectorias de diferentes entidades, trayectorias de la misma entidad realizadas en distintos tiempos o días, o diferentes fragmentos de una misma trayectoria. Generalmente la meta de la comparación es establecer relaciones entre objetos que son comparados. Estos son ejemplos de posibles relaciones:

- Igualdad o desigualdad.
- Orden (mayor o menor, anterior o posterior).
- Distancia (en espacio, tiempo o cualquier escala numérica).
- Relaciones topológicas (inclusión, superposición, cruce, toque).

Muchos otros tipos de relaciones pueden ser interesantes, dependiendo de la naturaleza de las cosas a comparar. Comparando trayectorias, los estudios tradicionales están centrados frecuentemente en establecer los siguientes tipos de relaciones:

- Similitud o diferencia del total de las características de las trayectorias que se listaron anteriormente.
- Relaciones espacio-temporales.
 - Co localización en espacio, total o parcial (la trayectoria consiste de las mismas posiciones o tienen algunas posiciones en común).
 - *Co localización ordenada, las posiciones comunes fueron alcanzadas en el mismo orden.*
 - *Co localizaciones no ordenadas, las posiciones fueron alcanzadas en distinto orden.*
 - Co existencia en tiempo, total o parcial (las trayectorias fueron hechas durante el mismo periodo de tiempo).
 - Co incidencia en espacio y tiempo total o parcial (algunas posiciones fueron alcanzadas al mismo tiempo).
 - Co incidencia rezagada (si una entidad alcanza las mismas posiciones que otra pero después de un retraso de tiempo).
 - Distancias en espacio y tiempo.

Muchos investigadores tratando con datos de movimiento afirman la necesidad de considerar no solo trayectorias con sus características espaciales y temporales sino también la estructura y las propiedades de espacio y tiempo donde el desplazamiento es llevado a cabo. Para esto a continuación se explican dichos conceptos.

2.3 Espacio.

Espacio puede ser visto como un conjunto de localizaciones o lugares, una propiedad importante del espacio es la existencia de distancias entre elementos. Al mismo tiempo, el espacio no tiene un punto de origen natural y un orden natural entre elementos. Por tanto, para comparar y ordenar posiciones en espacio uno necesita introducir en el algún sistema de referencia, por ejemplo un sistema de coordenadas.

Dependiendo de las necesidades practicas, uno puede tratar el espacio como de dos dimensiones (cada posición está dada por un par de coordenadas) o tridimensional (cada posición es dada por tres coordenadas). En casos específicos el espacio puede verse con una sola dimensión, por ejemplo cuando el movimiento sobre una ruta estándar es analizado. Teóricamente uno también podría tratar con espacios que tengan más de tres dimensiones, dichos espacios son abstractos en vez de físicos.

El espacio físico es continuo, lo que significa que se compone de un número infinito de localidades y, además, para cualquiera de dos lugares diferentes hay infinitos lugares «en medio». Además se cumple la propiedad de la desigualdad triangular, que establece que la suma de distancias a un punto intermedio debe ser menor o igual a la distancia directa entre los lugares. Aunque se trate de un espacio continuo, también puede ser útil tratarlo como un espacio discreto o incluso conjunto finito de lugares. Por ejemplo, en el estudio de la circulación de los turistas más de un país o una ciudad, puede "reducir" el espacio en un conjunto de puntos de interés visitados por los turistas. Discretizar el espacio puede ser incluso indispensable, en particular, cuando posiciones de las entidades no pueden ser medidas con precisión y que se especifican en términos de áreas tales como células de una red de telefonía móvil, barrios, o países.

Los ejemplos antes citados demuestran que el espacio puede ser estructurado por zonas. La división puede ser jerárquica, por ejemplo, un país es dividido en provincias, las provincias en municipios y los municipios en distritos. Zonas también pueden ser derivadas de una descomposición geométrica (por ejemplo, celdas de 1km^2), sin semántica asociada a la descomposición. Una red de calles (carreteras) es otra forma común de estructurar el espacio físico. Al igual que los sistemas de coordenadas, la estructuración del espacio puede proveer de un sistema de referencia, que puede ser utilizado para distinguir las posiciones, por ejemplo, al referirse a las calles o a fragmentos de carreteras y las posiciones relativas de ellos (los números de las casas o las distancias de los extremos). Las posibles maneras de especificar posiciones en el espacio pueden resumirse como lo siguiente:

- Referencias basadas en coordenadas: las posiciones son especificadas en duplas de números representando distancias lineares o angulares de ciertos ejes o ángulos seleccionados.
- Referencias basadas en divisiones: se refieren al compartimiento de una geometría aceptada o una base semántica de la división del espacio posiblemente jerárquica.

- Referencia lineal: refiriéndose a las posiciones relativas a lo largo de objetos lineales tales como calles, caminos, ríos, tuberías, por ejemplo los nombres de las calles más números de casas o códigos de caminos más distancias desde uno de los extremos.

Dado que a menudo se da el caso de que las posiciones de las entidades no pueden determinarse con exactitud, pueden ser representadas en los datos como incertidumbre, por ejemplo, como zonas en lugar de puntos.

A veces, un analista no está tan interesado en posiciones absolutas en el espacio como en las posiciones relativas con respecto a un lugar determinado. Por ejemplo, el analista puede estudiar cuando una persona viaja con respecto a su casa o los movimientos de los espectadores de un cine o un estadio. En tales casos, es conveniente definir las posiciones en términos de distancias y direcciones desde el lugar de referencia (o, en otras palabras, por medio de coordenadas polares). Las direcciones se pueden definir como la base de unos ángulos de dirección o en situación geográfica: norte, noroeste y así sucesivamente.

Un análisis comprensivo puede requerir consideración de los mismos datos en diferentes sistemas de referencia espacial y, por tanto, la transformación de un sistema de referencia a otro: de coordenadas geográficas a polar (con diversos orígenes), la coordinación basada en referencias a la división basada en red, etc.

También puede ser útil hacer caso omiso de las posiciones espaciales de los lugares y considerarlos desde la perspectiva de su ámbito específico de la semántica, por ejemplo, casa, lugar de trabajo, lugar de compras. En este caso se suele hablar de secuencias de eventos.

Cabe señalar que el espacio (en particular, el espacio físico) no es uniforme, sino heterogéneo, y sus propiedades varían de un lugar a otro. Estas propiedades pueden tener un gran impacto en los comportamientos de circulación y, por tanto, debe tenerse en cuenta en el análisis. Las características relevantes de la ubicación de individuos son las siguientes:

- Altitud, pendiente, aspecto y otras características del terreno.
- Accesibilidad en relación con diversas restricciones (obstáculos, la disponibilidad de caminos, etc)
- Carácter y propiedades de la superficie: la tierra o el agua, el hormigón o el suelo, los bosques o campo, etc
- Los objetos presentes en una ubicación: edificios, árboles, monumentos, etc
- Función o modo de uso, por ejemplo, la vivienda, las compras, la industria, la agricultura o el transporte
- Actividad de base semántica, por ejemplo, hogar, trabajo, compras, ocio

Cuando los lugares se definen como compartimentos de espacio (es decir, áreas en las dos dimensiones del espacio o volumen en el espacio tridimensional), o los elementos de la red en lugar de puntos, las características pertinentes también incluyen lo siguiente:

- Medida y forma espacial
- Capacidad, es decir, el número de entidades que la ubicación puede contener simultáneamente
- La homogeneidad o heterogeneidad de las propiedades

Cabe señalar que las propiedades de los lugares pueden cambiar con el tiempo. Por ejemplo, puede ser una ubicación accesible entre semana y de difícil acceso los fines de semana, una plaza puede ser utilizada como un mercado en horas de la mañana, un segmento de carretera puede ser bloqueada o reducida su capacidad a causa de un accidente o reparación de obras.

2.4 Tiempo.

Similar al espacio, hay diferentes maneras de definir posiciones en el tiempo, y el tiempo también puede ser heterogéneo en términos de propiedades de los momentos y los intervalos de tiempo.

Matemáticamente, el tiempo es un conjunto continuo con un orden lineal y distancia entre elementos, donde los elementos son momentos o posiciones en el tiempo. Análogo a las posiciones en el espacio, algún sistema de referencia es necesario para la especificación de los momentos en los datos. En la mayoría de los casos, las referencias temporales se realizan sobre la base de los estándares del calendario Gregoriano y el nivel de división del día en horas, las horas en minutos y así sucesivamente. La hora del día pueden ser especificadas según la zona horaria del lugar donde se recogen los datos o como la hora del Meridiano de Greenwich (GMT). Hay casos, sin embargo, cuando los datos se refieren a momentos de tiempo relativos, por ejemplo, el tiempo transcurrido desde el comienzo de un proceso o una observación, o el resumen sellos de tiempo especificado en forma de números 1, 2 y así sucesivamente. A diferencia del tiempo físico, el tiempo abstracto no es necesariamente continuo.

Al igual que las posiciones en el espacio, se puede trabajar con momentos de forma imprecisa mediante intervalos en lugar de instantes en el tiempo. Por otro lado, aunque el tiempo es continuo, los datos no pueden hacer referencia a cada punto posible. Para dos momentos sucesivos t_1 y t_2 referidos en los datos, hay momentos en que no hay datos. Por lo tanto, no se puede saber lo sucedido entre t_1 y t_2 , pero esto sólo se puede estimar por medio de interpolación.

El tiempo físico no es sólo una secuencia lineal de momentos, sino que incluye los ciclos inherentes de la rotación diaria de la tierra y su revolución anual. Estos ciclos naturales se reflejan en el método estándar de tiempo de referencia: las fechas se repiten cada año y en los tiempos de cada día. Además de estos ciclos naturales, también hay ciclos relacionados con las actividades de la gente, por ejemplo, el ciclo semanal.

Ciclos temporales pueden ser anidados, en particular, el ciclo diario está anidado dentro del ciclo anual. Por lo tanto, el tiempo puede verse como una jerarquía anidada de los ciclos. Varias jerarquías alternativas pueden existir, por ejemplo, el año / mes / día en el mes y el año/ semanas del año / días de la semana.

Es muy importante saber que los ciclos temporales son relevantes para los estudios de movimiento y que se tengan debidamente en cuenta en el análisis. Por este propósito, es necesario que los ciclos se reflejen en las referencias temporales de los elementos de datos. Normalmente, esto se hace a través de especificar el número de ciclo y la posición desde el inicio del ciclo. De hecho, el estándar de las referencias a fechas y horas del día, se construyen de acuerdo con este principio. Sin embargo, además de las referencias a la norma anual y los ciclos diarios, las referencias a otros (potencialmente) ciclos relevantes, por ejemplo, el ciclo semanal de las actividades de las personas o los ciclos de los movimientos de los satélites,

puede ser necesario o útil. Por lo tanto, un analista puede tener que transformar las referencias estándar en referencias jerárquicas de tiempo alternativo.

Los ciclos pueden tener períodos variables. Para que los datos relativos de los diferentes ciclos sea comparable, se necesita de alguna manera normalizar las referencias de tiempo, por ejemplo, dividir el tiempo absoluto que cuenta desde el comienzo de un ciclo entre la longitud de este ciclo.

La transformación de las referencias de tiempo absoluto a relativas es útil cuando se necesita comparar los movimientos que se inician en diferentes momentos y / o proceder con diferentes velocidades. La referencia de tiempo relativa sería, en este caso, el tiempo contado desde el comienzo de cada movimiento, posiblemente, normalizado en la manera de dividir entre la duración del movimiento.

Como se ha señalado, las propiedades de los momentos y los intervalos de tiempo pueden variar, y esta variación puede tener una influencia significativa sobre los movimientos. Por ejemplo, los movimientos de personas en días de la semana pueden diferir de los movimientos en fines de semana y, además, los movimientos de los viernes distintos a los lunes y los sábados distintos de los domingos. En este ejemplo, tenemos un caso de una diferencia entre las posiciones dentro de un ciclo. Otro ejemplo es la diferencia entre los momentos de un día: mañana, mediodía, tarde y noche. Tales cambios irregulares también deberían tenerse en cuenta en el análisis de fenómenos tiempo-dependientes, en particular, los movimientos (por ejemplo, los días festivos entre semana).

Normalmente, la heterogeneidad de las propiedades de tiempo no está explícitamente reflejada en los datos y, por tanto, no puede ser tenido en cuenta automáticamente en el tratamiento de los datos, sino que son factores introducidos en el análisis posterior.

2.5 Entidades en movimiento y sus actividades.

Además de las propiedades de localización y tiempo, las entidades que se mueven tienen sus propias características, que pueden influir en el movimiento y, por tanto, hay que tener en cuenta en el análisis. Por ejemplo, los movimientos de las personas pueden depender en gran medida a su ocupación, edad, condición de salud, estado civil, y otras propiedades. También es pertinente si una entidad se mueve por sí misma o por medio de algún vehículo. La forma y los medios del movimiento plantean sus limitaciones en las posibles vías y otras características del movimiento.

Las personas son un ejemplo de las entidades que suelen moverse según un propósito. Los objetivos determinan las rutas y también pueden influir en otras características, en particular, la velocidad. Para otro tipo de entidades, por ejemplo, los tornados o las partículas elementales, uno necesita atender a las causas del movimiento en lugar de los fines.

Las características del movimiento también pueden depender de las actividades realizadas por las entidades durante su movimiento. Por ejemplo, el movimiento de una persona en una tienda difiere de la circulación en una calle o en un parque. Las características de la circulación pueden cambiar cuando la persona comienza a hablar por un teléfono móvil. En el caso del turismo las actividades pueden variar en función de las rutas seguidas o inclusive de los horarios de visitas de los sitios turísticos ya que si un museo está próximo a cerrar el turista puede cambiar su ruta para desplazarse hasta el sitio próximo a cerrar y esto generaría un

cambio total en las actividades a realizar e inclusive podría eliminar ciertos lugares que en un principio estaban tomados en cuenta como lugares de interés.

Cualquier movimiento ocurre en un ambiente y está sujeto a la influencia de varios eventos y fenómenos que puedan ocurrir en este ambiente. Los movimientos de la gente son influenciados por el clima y la temperatura en la que se encuentren, por deportes y eventos culturales y por muchos aspectos más. Para detectar dichas influencias o para timarlas en cuenta en los datos de movimiento, el análisis requiere incluir datos adicionales y conocimientos previos.

3. Anonimización de Trayectorias.

3.1 Introducción.

En esta sección se da una descripción del estado del arte en la protección de la localización desde una perspectiva dual de privacidad de los datos en consultas para obtener servicios basados en la localización y anonimización de trayectorias.

Como se ha explicado en la introducción, la anonimización de trayectorias es cada vez más importante, puesto que es más fácil obtener este tipo de datos. Sin embargo la privacidad no debe tener un impacto negativo en el sentido de que la utilidad de los datos publicados debe ser preservada.

Así pues podemos formular el problema de anonimización como encontrar una modificación la información de un conjunto de datos T dondequiera que exista fuga de privacidad, para convertirla en la base de datos segura T' . Las dos metas conflictivas en la transformación de estos datos son, la satisfacción de ciertos requerimientos de privacidad y maximizar la utilidad de los datos publicados medidos por la similitud de las trayectorias originales en T con sus respectivas transformaciones en T' .

id	trajectory	id	trajectory
t_1	$a_1 \rightarrow b_1 \rightarrow a_2$	t'_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t_2	$a_1 \rightarrow b_1 \rightarrow a_2 \rightarrow b_3$	t'_2	$a_1 \rightarrow b_1 \rightarrow a_2$
t_3	$a_1 \rightarrow b_2 \rightarrow a_2$	t'_3	$a_1 \rightarrow b_2 \rightarrow a_2$
t_4	$a_1 \rightarrow a_2 \rightarrow b_2$	t'_4	$a_1 \rightarrow a_2 \rightarrow b_2$
t_5	$a_1 \rightarrow a_3 \rightarrow b_1$	t'_5	$a_3 \rightarrow b_1$
t_6	$a_3 \rightarrow b_1$	t'_6	$a_3 \rightarrow b_1$
t_7	$a_3 \rightarrow b_2$	t'_7	$a_3 \rightarrow b_2$
t_8	$a_3 \rightarrow b_2 \rightarrow b_3$	t'_8	$a_3 \rightarrow b_2$

Ilustración 4. Base de datos de Trayectorias

Originalmente la propiedad de k -anonimidad previene a un adversario de identificar información de calidad dada en un conjunto con menos de k elementos en el conjunto de datos anónimo. Asumiendo que los adversarios conocen acerca de todos o algunos de los puntos espacio-temporales de un individuo, el conjunto de todos los puntos correspondientes a la trayectoria pueden ser usados como casi identificadores en esta aplicación. En este contexto, la propiedad de k -anonimidad requiere que una trayectoria dada en el conjunto de datos original pueda en el mejor de los casos ser vinculada por lo menos en k trayectorias del conjunto de datos anonimizados. Puede ser demostrado que la siguiente definición de k -anonimidad satisface los requerimientos y además preserva la verdad del conjunto de datos original:

Definición 1. (Nergiz et al., 2009) *K-anonimidad de trayectorias.* Una base de datos de trayectorias T^* es una k -anonimización de un conjunto de datos de trayectorias T si.

- Por cada trayectoria en T^* , hay al menos $k-1$ otras trayectorias con exactamente el mismo conjunto de puntos.
- Las trayectorias en T y T^* pueden ser clasificadas de tal manera que las i th trayectorias $tr_i^* \in T^*$, $tr_i \in T$ satisfaga $tr_i \subset tr_i^*$ para todas las i .

Una solución sencilla para proteger la privacidad es remover todos los identificadores del usuario de los datos, o reemplazarlos con pseudo-identificadores. Sin embargo, esta anonimización no evita que el individuo pueda ser re-identificado a partir de los datos propios de la secuencia de eventos.

Las técnicas de preservación de la privacidad en la publicación de trayectorias en términos generales pueden clasificarse en dos categorías (Ghinita, 2009):

- Métodos que publican muestras de localización independiente. El objetivo de estas técnicas es prevenir a un atacante reconstruir trayectorias basadas en localizaciones independientes. Publicar muestras de localizaciones independientes es usual en aplicaciones que requieren solamente información agregada, tales como monitoreo de tráfico. En este caso solo datos de localización global son requeridos.
- Métodos que publican trayectorias individuales. En otras clases de aplicaciones, los patrones de movimiento y la relación de causalidad entre ciertas localizaciones de origen y destino pueden ser de interés, en estos casos es necesario publicar trayectorias completas, en vez de muestras independientes de localización. El objetivo está en perturbar datos de trayectorias para prevenir la asociación de individuos a trayectorias. Estos métodos distorsionan la localización de las muestras en cada muestra de tiempo, estos métodos son adecuados para aplicaciones donde la relación de causalidad entre la localización del origen y el destino son importantes.

3.2 Métodos de Anonimización de Trayectorias.

Los métodos de perturbación de trayectorias analizados formulan el problema de confundir al atacante con un problema de optimización no lineal con restricciones. El objetivo es maximizar la función de privacidad E bajo la limitante de que la máxima distorsión para cada localización publicada no exceda el umbral R , que es dependiente de la aplicación. La fase de perturbación necesita considerar todas las permutaciones de la ubicación de las muestras para la asignación de pistas, por lo tanto el costo computacional es muy alto. Por ejemplo para realizar la perturbación para N trayectorias de usuarios de M muestras cada uno, la complejidad es $O(N!)^M$, que no es factible en la práctica. Para disminuir la sobrecarga, una fase de “Segmentación de la secuencia” es realizada antes de la perturbación.

La idea es reducir el espacio de búsqueda para el problema de limitación de la optimización, descartando algunas de las hipótesis poco probables de asignación de muestras a las secuencias.

El algoritmo de perturbación de secuencias maximiza la privacidad métrica teniendo en cuenta las limitaciones de la precisión R . Sin embargo, esto no es suficiente para proteger la privacidad de los usuarios en áreas pocos densas. Si las trayectorias de los usuarios están situadas lejos una de otra, inclusive la mejor expectativa de error de distancia puede no ser suficiente para prevenir la re identificación. En la práctica, la amenaza a la privacidad ocurre cuando las trayectorias individuales pueden ser asociadas a usuarios en particular. Además dicha asociación no puede ser realizada en áreas muy densas, solamente en áreas poco densas, y la probabilidad de éxito del atacante se incrementa con la longitud de la trayectoria divulgada. Basado en estas observaciones existe un método de medición de privacidad propuesto, que calcula para cada usuario “ u ” la entropía.

$$H(u) = - \sum_i p_i \log p_i$$

Donde p_i es la probabilidad de asociar u a la muestra de localización i que es una muestra particular.

A continuación se van a detallar los algoritmos concretos estudiados.

3.2.1 Protegiendo la privacidad de la localización a través de confusión de trayectorias.

En (Hoh&Gruteser,2005) se estudian mecanismos para prevenir a un adversario rastrear completamente una trayectoria individual. Los algoritmos de perturbación imponen errores tolerables en las muestras de la localización original para mantener niveles específicos de calidad en el servicio. Se distinguen los siguientes pasos:

1. Definir el modelo de privacidad de la localización en términos de confianza y distancia espacial.
2. Definir la calidad del servicio (QoS) en términos del error que el algoritmo impone a las muestras de localización.
3. Derivar un algoritmo de una limitada optimización del problema formulado que maximiza la medida de la privacidad de la localización dado cierto requerimiento de QoS.

Privacidad de la localización y medición de la calidad en el servicio.

Definimos el grado de privacidad de la localización como la exactitud con la que una fuente no fiable puede localizar a un usuario en específico.

Para confundir al atacante, la ubicación de las muestras individuales es distorsionada, con el fin de minimizar la probabilidad de obtener una acertada ubicación de las trayectorias.

Distorsionando las muestras de localización inherentemente introduce inexactitud de los datos, y puede tener repercusiones en la búsqueda correcta en los datos. Un compromiso entre privacidad e inexactitud debe conseguirse. En (Hoh&Gruteser, 2005) proponen una medida para cuantificar ambas, privacidad e inexactitud.

Específicamente la privacidad es medida a través de la “Expectativa del error de distancia”, que captura que tan exacto un adversario puede igualar las localizaciones con las secuencias. Dados N usuarios (por consiguiente N muestras de la localización en cada instante de tiempo) y un tiempo de observación de M de muestras de tiempo, la expectativa del error de distancia para la trayectoria del usuario u se mide de la siguiente manera.

$$E[u] = \frac{1}{NM} \sum_{i=1}^M \sum_{j=1}^{I_i} p_j(i) d_j(i)$$

Ecuación 1

Donde I_i es el número total de hipótesis (posibles localizaciones) asignadas por al usuario u en el instante de tiempo i , $p_j(i)$ es la probabilidad asociada con la hipótesis j en la muestra de tiempo i , y $d_j(i)$ es la distancia entre la posición actual y la estimada de u en la muestra de tiempo i .

La inexactitud de los datos se mide de acuerdo a la calidad del servicio (QoS).

$$QoS = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \sqrt{(x_{u_i}(j) - x'_{u_i}(j))^2 + (y_{u_i}(j) - y'_{u_i}(j))^2}$$

Ecuación 2

Donde (X_{ui}, Y_{ui}) y (X'_{ui}, Y'_{ui}) son las coordenadas actuales y las perturbadas de U_i respectivamente.

Algoritmo de anonimización por confusión de trayectorias (Path confusion)

La idea principal del siguiente algoritmo de privacidad es el concepto de confusión de trayectos. Cada vez que el trayecto de dos usuarios se encuentra (se define encontrar como estar cerca uno de otro) hay una probabilidad de que el adversario confunda los trayectos y siga al usuario equivocado. Un algoritmo de privacidad puede explotar esto perturbando la información de la localización en ciertas áreas de encuentro para incrementar las probabilidades de confusión.

El problema de confundir al atacante se puede formular como un problema de optimización no lineal limitada. El objetivo es maximizar la expectativa de error E , bajo la limitante de que la máxima distorsión para cada localización publicada no exceda el umbral R , que es dependiente de cada aplicación concreta.

$$(\widetilde{x}_n(k) - x_n(k))^2 + (\widetilde{y}_n(k) - y_n(k))^2 \leq R^2$$

Ecuación 3

Donde $(x_n(k), y_n(k))$ son las coordenadas reales, $(\widetilde{x}_n(k), \widetilde{y}_n(k))$ son las coordenadas distorsionadas, k es un cierto instante de tiempo y R es un parámetro que define la máxima perturbación permitida. El objetivo es maximizar:

$$\max_{\forall \widetilde{x}_n(k), \forall \widetilde{y}_n(k)} \frac{1}{N} \sum_{i=1}^I p_i(k) d_i(k)$$

Ecuación 4

Donde la distancia total de error $d_i(k)$ y la probabilidad del adversario $p_i(k)$ es descrita por las siguientes ecuaciones.

$$d_i(k) = \sum_{n=1}^N \sqrt{(\widetilde{x}_{m_i(n)}(k) - x_n(k))^2 + (\widetilde{y}_{m_i(n)}(k) - y_n(k))^2}$$

Ecuación 5

$$p_i(k) \equiv P(\Omega_i^k | Z^k) \approx \prod_{n=1}^N f_n(\widetilde{x_{m_i(n)}}(k), \widetilde{y_{m_i(n)}}(k))$$

Ecuación 6

Las formulas para estimar la probabilidad de asignación del adversario son derivadas del “Algoritmo de Rastreo de Multi Hipótesis de Reid” (Reid D., 1979). La probabilidad, $p_i(k)$ denota la probabilidad de la hipótesis Ω_i en el tiempo k , dada una serie de observaciones Z^k . (denota el conjunto acumulativo de muestras de localización a través del tiempo k mientras que $Z(k)$ indica el conjunto de muestras de localización solamente en tiempo k . En la ecuación anterior m_i es una tarea de un vector para la i hipótesis. Cada campo j en el vector contiene el índice del usuario que la hipótesis le asigna a la muestra de localización j . La función f_n es una densidad gaussiana multi variable, descrita en la siguiente ecuación.

$$f(z^k | \bar{x}^k) = N(z^k - H\bar{x}^k, B),$$

Ecuación 7

Donde x^k es el vector de estado consistente de la posición estimada y la velocidad en el paso k y z^k son un nuevo vector de observación. El vector de estado, x^k puede ser predicho desde su estado anterior x^{k-1} de acuerdo al modelo de proceso z^k se relaciona a su estado actual a través de un modelo de observación.

$$x^k = Fx^{k-1} + w \quad y \quad z^k = Hx^k + v$$

Ecuación 8

Donde w representa el vector de ruido procesado y la matriz F describe una predicción linear del siguiente estado dado el anterior. La matriz H convierte un vector de estado en medidas de dominio y v representa el vector de medida de ruido.

El algoritmo de perturbación de secuencias entrega secuencias perturbadas de la serie de dos secuencias de usuarios. Maximiza instantáneamente la privacidad de la localización en cada paso modificando la serie original de muestras de localización dentro del radio de perturbación R . Mayor R resulta un mayor grado de privacidad, menor R limita el efecto de la perturbación, lo que conduce a mayor calidad de servicio y menor privacidad.

Algoritmo 1, Perturbación de secuencias calculadas en una serie de muestras para dos usuarios usando una matriz de 2 por B por 2.

Entradas: Secuencias originales, el conjunto de localizaciones continuas de dos usuarios; R , radio del círculo perturbado como entrada de usuario; B , el tamaño del segmento; proceso (movimiento del usuario) y observación (error de tracking) modelo para seguimiento de objetivos.

for $k = 1$ to B **do**

for all hypothesis i **do**

 1. (Predicción de estado): Calcular el estado de predicción de cada usuario basados en árboles padre.

end for

2. (Generación de hipótesis I): con la predicción del estado obtenida del paso uno, establecer ecuación (6) para cada hipótesis i donde secuencias perturbadas fueron dadas.

3. (Generación de hipótesis II): Establecer ecuación (5) para cada hipótesis i .

4. (Limitantes de calidad en el servicio): Establecer ecuación (3) para cada usuario n con R .

5. (Resolver el problema de optimización limitado): Construir la función costo en la ecuación (4) con el resultado de los pasos 2 y 3. Establecer limitaciones de la desigualdad con el resultado del paso 4.

for all i th hypothesis do

6. (Paso de corrección de estado): Calcula la actualización del estado de cada usuario basado en las muestras perturbadas.

7. (Guardar probabilidad de padres): Las probabilidades obtenidas en el paso 6 sin guardadas para el próximo árbol de probabilidad.

end for

end for

El uso de este algoritmo de perturbación de secuencias se ilustra en un escenario simple donde dos usuarios viajan por trayectos aproximadamente paralelos.

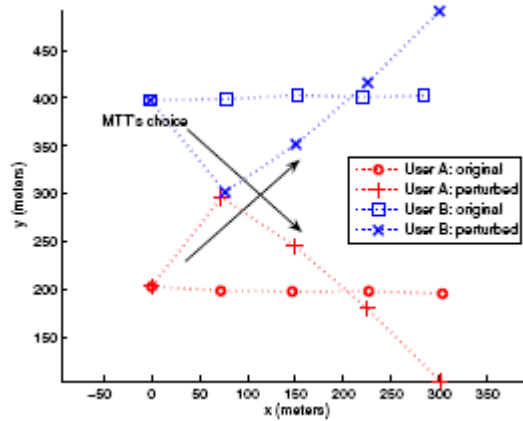


Ilustración 5. Dos usuarios moviéndose en paralelo.

La imagen anterior representa las confusiones que el algoritmo de perturbación de secuencias crea. Los círculos rojos son localizaciones periódicas de un usuario A mientras que los azules son de un usuario B. Las cruces y las marcas x son muestras perturbadas del usuario A y B respectivamente. Ambos usuarios se movieron de izquierda a derecha empezando con cerca de 200 metros de separación con una velocidad horizontal de 15m/s. R está establecido a 100 metros. El algoritmo asume una asignación correcta con probabilidad 1 en el primer paso. Después de eso el algoritmo genera dos hipótesis para cada hipótesis padre, que son generados durante el paso previo. Para el caso de dos usuarios el algoritmo debe mantener una hipótesis 2^{k-1} en el paso k . Comenzando en el segundo paso, trata de maximizar la expectación del error de distancia, lo que conduce a la conversión de trayectos paralelos en

trayectos que se cruzan. Las flechas en la imagen muestran el resultado de aplicar el algoritmo de multi tracking a los datos perturbados: el algoritmo confunde a los dos usuarios y siguen el camino incorrecto.

Segmentación de trayectorias

La fase de perturbación necesita considerar todas las permutaciones de la ubicación de las muestras para la asignación de pistas, por lo tanto el costo computacional es muy alto. Por ejemplo para realizar la perturbación para N trayectorias de usuarios de M muestras cada uno, la complejidad es $O(N!)^M$, que no es factible en la práctica. Para disminuir la sobrecarga, una fase de “Segmentación de la secuencia” es realizada antes de la perturbación.

La idea es reducir el espacio de búsqueda para el problema de limitación de la optimización, descartando algunas de las hipótesis poco probables de asignación de muestras a las secuencias.

El algoritmo de perturbación de trayectorias se desempeña mejor en trayectos paralelos cortos, de forma que el algoritmo de segmentación se diseñó de la siguiente manera.

En cada paso N los usuarios reportan sus muestras de localización, el algoritmo toma en cuenta de la distancia entre las muestras de localización en cada paso y después filtra $\frac{N(N-1)}{2}$ combinaciones en una lista de candidatos que permanece lo suficientemente cerca para los siguientes B pasos. Estos segmentos candidatos deben contener segmentos con usuarios comunes.

La segmentación de trayectos toma una matriz $In [2][K][N]$ como entrada la cual es el conjunto de muestras originales de N usuarios para K muestras de tiempo. Además de eso toma α (un factor escalable) y R del usuario. Después de la segmentación la salida es $Out[2][K][N]$ que es el conjunto de las muestras de localización perturbadas (en dos dimensiones) de N usuarios. El algoritmo usa las siguientes estructuras de datos.

El algoritmo de perturbación de secuencias maximiza la privacidad métrica teniendo en cuenta las limitaciones de la precisión R . Sin embargo, esto no es suficiente para proteger la privacidad de los usuarios en áreas pocos densas. Si las trayectorias de los usuarios están situadas lejos una de otra, inclusive la mejor expectativa de error de distancia puede no ser suficiente para prevenir la re identificación. En la práctica, la amenaza a la privacidad ocurre cuando las trayectorias individuales pueden ser asociadas a usuarios en particular. Además dicha asociación no puede ser realizada en áreas muy densas, solamente en áreas poco densas, y la probabilidad de éxito del atacante se incrementa con la longitud de la trayectoria divulgada. Basado en estas observaciones existe un método de medición de privacidad propuesto, que calcula para cada usuario “ u ” la entropía.

$$H(u) = - \sum_i p_i \log p_i$$

Donde p_i es la probabilidad de asociar u a la muestra de localización i que es una muestra particular.

3.2.2 Ofuscación de localizaciones.

En (Ruppel et al., 2006) se presenta un método de Anonimización adecuado para detección de proximidad y separación. Puede ser clasificado como Anonimización basada en *ofuscación de datos*, de forma que la anonimidad se define como “el estado de ser no identificable dentro de un conjunto de objetos, el conjunto anónimo”. La meta es proteger la identidad de los objetivos revelados por un proveedor LBS y el proveedor de localización (LP) respectivamente. Esta aproximación está basada en comunicación pseudónima.

Se pueden definir los servicios basados en localización como servicios que permiten ofrecer a los usuarios un valor agregado, basados en la información de localización de los dispositivos móviles. (Castañeda et al, 2006)

A continuación se define la nomenclatura de este método.

$E = \{e_1, e_2, \dots, e_n\}, 1 < i \leq n$ sean un conjunto de objetivos que forman una comunidad.

$P(e, t): E \times R \rightarrow R^2$ la posición actual del objetivo $e \in E$ en tiempo t .

$S_G \in N$ una clave secreta específica de E .

$P^*(e, t, S_G): E \times R \times N \rightarrow R^2$ la posición ofuscada de e en tiempo t .

Esta aproximación sigue una estrategia de dos pasos: primero todas las coordenadas son transformadas por una transformación global independiente de tiempo, consiste en una rotación con ángulo α alrededor de $(j, k) \in R^2$ seguido del traslado $(x_{global}, y_{global}) \in R^2$.

En el segundo paso, el movimiento local de los objetivos es empañado añadiendo un vector de traslado dependiente del tiempo $v := (x_{local}, y_{local}) \in R^2$.

La motivación para el primer paso es la ofuscación de la referencia global de las coordenadas en fin de evitar ataques basados en saber el paradero de un objetivo. Para el segundo paso, los ataques basados en el conocimiento del paradero del objetivo, patrones de movilidad y patrones de camino deben ser evitados.

Es importante observar que todos los parámetros de este algoritmo $(\alpha, j, k, x_{global}, y_{global})$ dependen de S_G y son por lo tanto secretos dentro de una comunidad. Igualmente, la ofuscación local del vector $v := (x_{local}, y_{local})$ está limitado a la longitud $|v| \leq r_{max_local}$ y depende de S_G así como del tiempo actual.

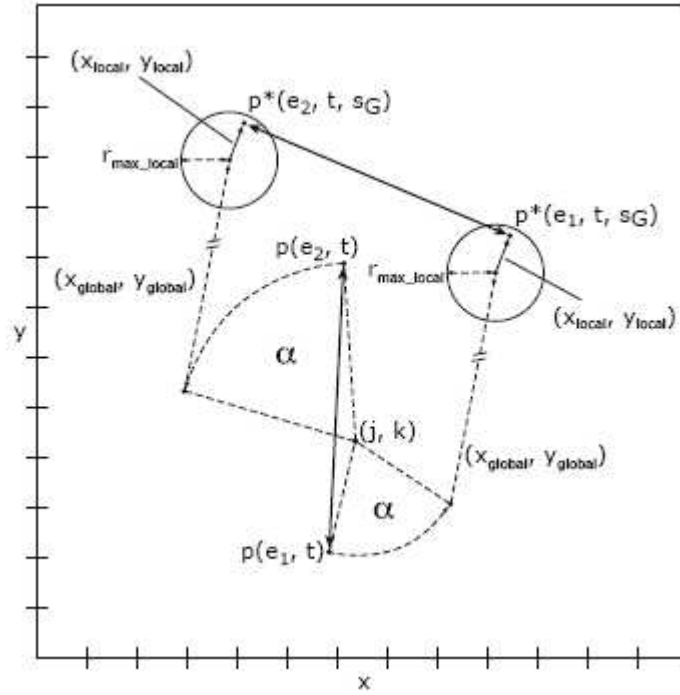


Ilustración 6. Distancia preservando la ofuscación de dos posiciones $p(e_1, t)$ y $p(e_2, t)$.

3.2.3 Modificación de proyecciones.

El trabajo “Privacy Preservation in the Publication of Trajectories” (Terrovitis&Mamoulis, 2008) considera un escenario donde las muestras de localización son señaladas en un conjunto discreto (por ejemplo, puntos al por menor, atracciones turísticas etc.) y asumen un modelo de ataque con un claro conocimiento de los antecedentes. Específicamente el atacante ya conoce algunos fragmentos de la trayectoria, y la identidad de los usuarios correspondientes a esos fragmentos.

Consideremos el ejemplo de una compañía P (publicador) que comercializa tarjetas como formas de pago convenientes. Dichas tarjetas pueden ser usadas para pagar el transporte, así como compras del día a día. En un tiempo P (que es confiable para todos los usuarios de tarjetas) reunirá grandes cantidades de datos de trayectorias que pueden ser usados para diferentes propósitos (por ejemplo, inferir en los viajes del consumidor y sus patrones de gasto). Sin embargo P está obligado por ley a no comprometer la privacidad de sus clientes. Además, las compañías asociadas de P no son confiables. Por ejemplo una compañía de distribución A tiene acceso a todas las compras del usuario u , y además aprende la identidad de u a través de un servicio de fidelidad de cliente, así A tiene acceso a un sub conjunto de trayectorias seguidas por u , y desea saber cuáles fueron los lugares que u ha visitado. P debe prevenir este tipo de amenazas a la privacidad.

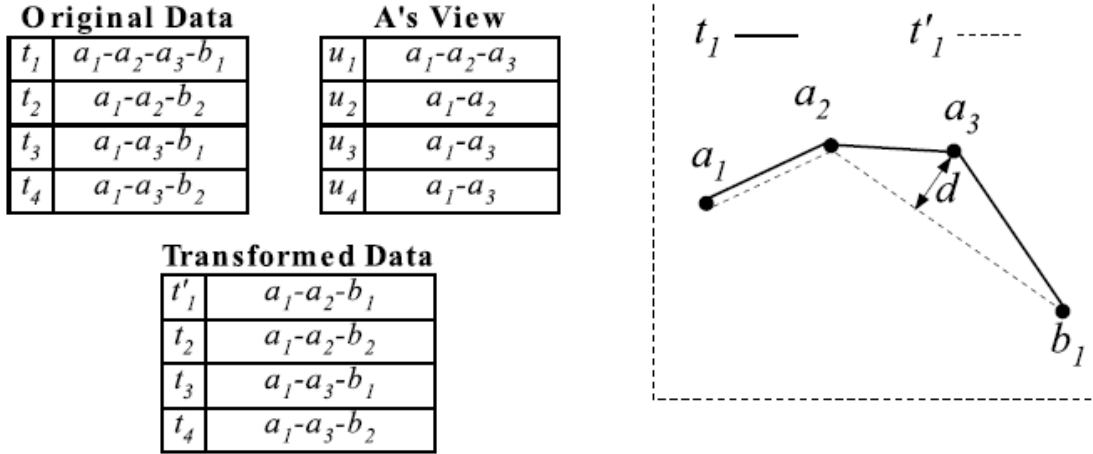


Ilustración 7. Anonimización de dos trayectorias.

Considerando la figura anterior donde P publica los datos de localización original. Las compañías asociadas A y B tienen conocimiento acerca de sub conjuntos de trayectorias correspondientes a sus puntos de operación. Estas ubicaciones están denotadas por a_i y b_j respectivamente. A puede tratar de inferir las otras ubicaciones que sus clientes han visitado inspeccionando los datos originales. Por ejemplo A puede identificar que u_1 corresponde a la trayectoria t_1 ya que solo t_1 corresponde al patrón de movimiento $a_1 - a_2 - a_3$ conocido por A . Por lo tanto A puede inferir con certeza que u ha visitado b_1 que puede corresponder a un club nocturno. Una asociación tan sensitiva es claramente una violación a la privacidad.

Definición del problema

Se modelan las trayectorias originales en la base de datos original como secuencias de localizaciones, que son puntos precisos en un mapa. A diferencia de la hipótesis común en datos espacio temporales, aquí se consideran un dominio espacial discreto, por ejemplo, la información espacial puede ser considerada como las direcciones en un mapa de una ciudad. Tal caso es muy cercano a la realidad para datos que refrenan a transacciones comerciales. Más formalmente se define como trayectoria lo siguiente.

Definición 1 Una trayectoria t con longitud n es una secuencia $t=[p_1, \dots, p_n]$ de direcciones tomadas de un conjunto P .

En el ejemplo mencionado anteriormente P incluye todas las direcciones de tiendas que aceptan tarjetas. Para simplificar se asume que hay una relación 1 a 1 entre las tiendas y sus direcciones, dado que las compañías suelen tener varias sucursales P puede ser particionada en m conjuntos no vacíos de direcciones disjuntas $P_1, P_2, \dots, P_m$, tal que cada conjunto contenga todas las direcciones de las sucursales de la compañía. Por lo tanto, asumiremos que queremos publicar una base de datos T de trayectorias, donde elementos secuenciales toman valores de P , y hay m adversarios, (por ejemplo compañías). Cada adversario $i \in V$ controla direcciones P_i de manera que $\forall i, j \in V, P_i \cap P_j = \emptyset$ y $\bigcup_{i \in V} P_i = P$. Para cada trayectoria $t \in T$, cada adversario $i \in V$ tiene una proyección t^i , definida a continuación.

Definición 2 Una proyección de una trayectoria $t=[p_1, \dots, p_n]$ con respecto a un adversario A , que tiene $P_A \subset P$ (potencialmente vacío) la trayectoria $t^A=[P_1^A, \dots, P_k^A]$, $\{P_j^A \mid P_j^A \in t \wedge P_j^A \in P_A\}$ el orden de elementos $P_j^A \in t$ esta preservada en t^A .

La proyección t^A de un $t \in T$ es la sub trayectoria de t que contiene solamente todos los puntos de P_A en t . Por lo tanto cada adversario tiene una base de datos local con las proyecciones de t

$\in T$ con respecto a P_A . El adversario no tiene conocimiento de las trayectorias teniendo proyecciones vacías, por lo tanto T_A puede ser menor que la base de datos del publicador. Una trayectoria puede aparecer en múltiples ocasiones en T y una o más trayectorias pueden tener la misma proyección con respecto a P_A . La propiedad más importante de t^A es que el adversario A puede vincular directamente a las personas que van pasando en su base de datos local (programa de lealtad).

Definición 3 Dada una base de datos T de trayectorias, donde las ubicaciones toman valores de P , se construye una base de datos transformada T' , de modo que si T' es pública para todos los $t \in T$, cada adversario A no pueda inferir correctamente cualquier ubicación $\{P_j \mid P_j \in t \wedge P_j \notin t^A\}$ con probabilidad más larga que P_{br} . En simples palabras no se desea que un adversario A que tiene una proyección t^A de una trayectoria original t en la base de datos original pueda inferir correctamente cualquier otra ubicación de las trayectorias después de acceder a T' .

El algoritmo

La idea principal detrás de este algoritmo de anonimización es transformar proyecciones largas y detalladas en proyecciones pequeñas y simples. Haciendo esto se es capaz de aumentar el apoyo de las proyecciones, diversificar las ubicaciones que están siendo monitorizadas por los adversarios haciendo así imposible para ellos inferir con alta certeza si las trayectorias incluyen ciertos puntos.

El mecanismo usado para la generalización de las trayectorias, es suprimir la existencia de ciertos puntos en ellas, tomando en consideración el beneficio en términos de privacidad y la desviación de la dirección de la trayectoria principal.

Encontrando el conjunto de puntos optimo para borrar de T en orden para obtener un seguro T' y conseguir la mínima perdida de la información es más complicado que la simple *k-anonimidad* en bases de datos relacionales, de la que se hablara más adelante.

En la primera etapa la base de datos proyectada T_i de cada atacante $i \in V$ es extraída. Después el algoritmo identifica las proyecciones que conducen a una violación de la privacidad escaneando T una vez (algoritmo 3), para cada dupla $t \in T$ y para cada adversario i , tal que la proyección t^i no este vacia, un contador $sup(p_i, t^i, T)$ es incrementado para cada $P_j \in t \mid P_j \notin t^i$. Después de escanear T los pares (p_i, t^i) para cada $P(p_i, t^i, T) = sup(p_i, t^i, T) / S(t^i, T) > P_{br}$ son identificados y t^i esta marcado como una proyección problemática para el adversario i . Después el Algoritmo 2 corre un ciclo; mientras que las violaciones a la seguridad son identificadas.

Algoritmo 2. Algoritmo de anonimización

Anonimización (T, P_{br})

Construir proyección DB T_i para cada atacante $i \in V$

Identificación de violación (T, P_{br}). Llama al Algoritmo. 3

$T' := T$. Inicializar la base de datos de salida

while todavía existan violaciones a la privacidad **do**

$U(t_x^i, t_y^i) := \text{unificación de bajo costo } (t_y^i \subset t_x^i \text{ and/or } t_y^i \text{ son o es problemático})$

for all $t' \in S(t_x^i, T')$ **do** . traj. En T' soportando t_x^i

Borrar todas $p \in t', p \in t_x^i \mid p \notin t_y^i$

Borrar t_x^i de T_i t_x^i no se soporta más en T'

Identificación de violación (T', P_{br})

Algoritmo 3. Identificación de violaciones a la privacidad

Identificación de violación (T, P_{br})

```

for all  $i$  in  $V$  do . Inicialización
    for all  $p_j \notin P_i$  do
        for all  $t_i \in t^i$ , do
             $\text{sup}(p_j, t^i, T) := 0$ ;
for all  $t$  in  $T$  do . Escaneo de base de datos para actualizar contadores
    for all  $i$  in  $V$  do
         $t^i := \text{proyección de } t \text{ en } T_i$ 
        if  $t^i \neq 0$  ; then
            for all  $p_j, p_j \in t, p_j \notin P_i$  do
                 $\text{sup}(p_j, t^i, T)++$ ;
for all  $i$  in  $V$  do . Inicialización
    for all  $p_j \notin P_i$  do
        for all  $t^i \in T_i$  do
            if  $\frac{\text{sup}(p_j, t^i, T)}{S(t^i, T)} > P_{br}$  then
                Marcar  $(p_j, t^i)$  como problemático

```

3.2.4 Anonimización por generalización.

En el trabajo (Nergiz et al., 2009) se propone que la privacidad de las trayectorias puede ser conseguida aplicando el concepto de k-anonimidad, donde las trayectorias de distintas líneas de tiempo pueden ser anonimizados al mismo tiempo y su diferencia de tiempo cuenta en la métrica usada para la inexactitud de los datos. Un algoritmo para la agrupación de la trayectoria en un dominio tridimensional de espacio tiempo es propuesto, el cual crea grupos de al menos k transacciones cada uno. Las trayectorias pertenecientes al mismo grupo son generalizadas de manera que sean indistintas unas de otras. El agrupamiento implica la generalización de ambas coordenadas, temporales y espaciales.

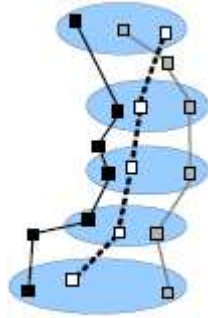
Aunque existen numerosos algoritmos de k anonimidad propuestos para tablas sencillas de datos, una aproximación de agrupación se muestra más adecuada para la anonimización de estructuras complejas, debido a la identificación directa de entidades privadas (trayectorias en nuestro caso) siendo anonimizadas. Encontrar la agrupación óptima de dos trayectorias es lo mismo a encontrar el punto de emparejamiento entre las dos trayectorias de manera que anonimizando las trayectorias a través de emparejamiento minimiza el costo de sesión.

K anonimidad para bases de datos de trayectorias

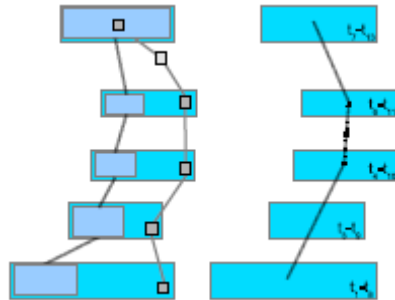
En (Nergiz et al., 2009) se extiende la definición de k-anonimidad a las trayectorias tradicional, que no incluye en la distorsión de los datos y la incertidumbre. En vez de esto se remueve información de los datos marcando el uso de generalizaciones espacio temporales, alineación de puntos en espacio y en tiempo, supresión de puntos y trayectorias. La medición usada es derivada estadísticamente y captura la sensibilidad de tiempo y espacio para hacer frente a varias aplicaciones. Además los trabajos previos parecen no medir el nivel de distorsión debido a la anonimización en el contexto de aplicaciones de minería de trayectorias, lo cual se considera una de las metas principales en la publicación de trayectorias.

A continuación se presentan algunas definiciones necesarias para entender el nuevo concepto de anonimización de un conjunto de trayectorias.

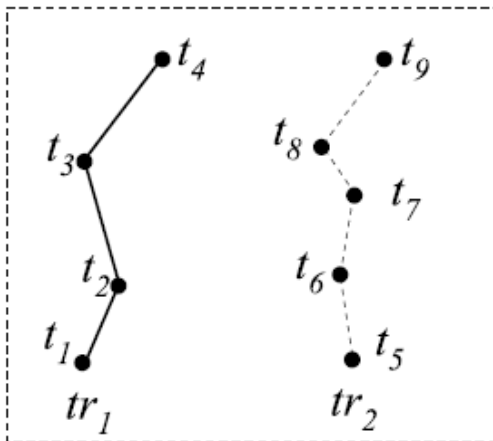
- *Punto de enlace y emparejamiento*, un punto entre un conjunto de trayectorias $TR = \{tr_1, \dots, tr_n\}$ es un conjunto de puntos ordenado $PL = \{p_1, \dots, p_n\}$ dado que $p_i \in tr_i$. Un conjunto de puntos ordenados y enlazados entre trayectorias en TR , $PM = \{PL_1, \dots, PL_M\}$, es un emparejamiento de puntos entre trayectorias si para todas $i < j$ y todos los posibles k , $PL_i t_k < PL_j t_k$. La figura muestra un emparejamiento de puntos entre trayectorias tr_1, tr_2 y tr_3 .



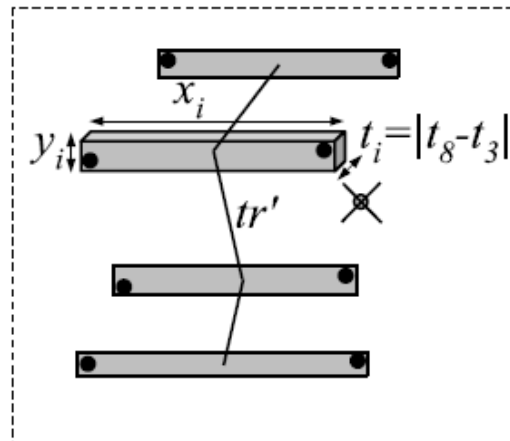
- Sea $TR = \{tr_1, \dots, tr_n\}$ un conjunto de trayectorias y sea $PM = \{PL_1, \dots, PL_m\}$ un válido emparejamiento de puntos entre ellos. Sea $TR^* = \{tr_1^*, \dots, tr_n^*\}$ otro conjunto dado que $tr_1^*, p_i = \dots = tr_n^*, p_i = BB_{PL_i}$. TR^* es una n anonimización de TR .



Un ejemplo de Anonimización de dos trayectorias tr_1 y tr_2 se muestra a continuación.



Original Trajectories



2-Anonymous Group

Cada muestra de localización esta etiquetada con un sello del instante de tiempo en la que fue recolectada. Los cuadros resultantes con sus diferencias de tiempo (ejemplo, el periodo de tiempo cubierto por cada cuadro) provocan una pérdida de información, que es medida de acuerdo a la ecuación LCM (log cost metric). LCM cuantifica la inexactitud de las trayectorias en ambos espacio y tiempo, y es calculado sumando la ampliación requerida sobre cada uno de los M extractos de localización publicados. Un factor de ponderación para espacio (w_s), así como tiempo (w_t) puede ser especificada, dependiendo de la aplicación que use los datos.

$$LCM = \sum_{i=1}^M [w_s (\log |x_i| + \log |y_i|) + w_t \log |t_i|]$$

La anonimización de trayectorias se realiza en dos etapas. Primero, el algoritmo selecciona las trayectorias que pertenecen a cada grupo. Esta fase es realizada a través de una heurística que es similar a la comparación de cadenas. Después una fase de anonimización es realizada, donde se decide que muestras de que trayectorias serán anonimizadas con muestras de otras trayectorias.

Agrupamiento de Trayectorias

El problema de alineamiento para dos trayectorias es poli nómico y puede ser resuelto usando una aproximación de programación dinámica (proceso de optimización). La ecuación que resuelve el problema de alineamiento para optimización contra la función incremental σ se menciona abajo. La medición del costo de sesión (LCM) es además es incremental y define σ de la siguiente forma:

$$\sigma_{LCM}(p_1, p_2) = \begin{cases} \log U, \\ \log BB_{\{p_1, p_2\}}, \end{cases}$$

Así que la distancia entre las dos trayectorias tr_1 y tr_2 es dada por.

$$DST(tr_1, tr_2) = OPT_{\sigma_{LCM}}(tr_1, tr_2)$$

Ecuación que resuelve el problema de alineamiento para optimización contra la función incremental σ

$$OPT_{\sigma}(tr_1, tr_2) = \begin{cases} \sum_{p_i \in tr_1} \sigma(p_i, \perp), & |tr_2| = 0; \\ \sum_{p_i \in tr_2} \sigma(p_i, \perp), & |tr_1| = 0; \\ \min \{ OPT_{\sigma}(tr_1 - tr_1.p_1, tr_2 - tr_2.p_1) + \sigma(tr_1.p_1, tr_2.p_1), \\ OPT_{\sigma}(tr_1, tr_2 - tr_2.p_1) + \sigma(tr_2.p_1, \perp), \\ OPT_{\sigma}(tr_1 - tr_1.p_1, tr_2) + \sigma(tr_1.p_1, \perp) \}, & |tr_1|, |tr_2| > 0. \end{cases}$$

El pseudocódigo para calcular el costo de sesión entre dos trayectorias esta dado en el algoritmo 4, la salida de este algoritmo es la distancia dada entre dos trayectorias y el punto optimo de emparejamiento que minimiza el costo de sesión.

Algoritmo 4

Requiere: Trayectorias $tr1 = \{p1, \dots, pm\}$, $tr2 = \{p1, \dots, pn\}$

Asegura: regresa la distancia entre $tr1$ and $tr2$ y el emparejamiento de puntos asociados PM.

```
1: PM = {}
2: Let M be a  $(m + 1) \times (n + 1)$  matrix.
3:  $M[i][0] = i \cdot \log U$  for all  $i \in [0 - m]$ 
4:  $M[0][j] = j \cdot \log U$  for all  $j \in [0 - n]$ 
5:  $i = 1, j = 1$ 
6: while  $i \leq m$  do
7:   while  $j \leq n$  do
8:      $M[i][j] = \min\{ M[i - 1][j - 1] + \log BB_{tr1.pi, tr2.pj}, M[i][j - 1] + \log U, \)$   

 $M[i - 1][j] + \log U \}$ 
9:     if  $M[i][j] = M[i - 1][j - 1] + \log BB_{tr1.pi, tr2.pj}$  then
10:       $PM += \{tr1.p_i, tr2.p_j\}$  //link  $tr_i.p_i$  and  $tr_j.p_j$ 
11:    end if
12:     $j++$ 
13:  end while
14:   $i++$ 
15: end while
16: Regresa la distancia  $M[m][n]$  y el emparejamiento de puntos PM.
```

Algoritmo de anonimización

Ya que los grupos están formados, las trayectorias dentro de cada grupo necesitan ser anonimizadas, como se menciona anteriormente el proceso de anonimización necesita especificar el emparejamiento de puntos óptimo que minimizara el costo de sesión. Encontrar el punto óptimo de emparejamiento entre dos trayectorias es fácil. El verdadero reto es encontrar el punto óptimo de alineamiento entre trayectorias $n > 2$.

Dado un conjunto de trayectorias $TR = \{tr_1, \dots, tr_n\}$ para $n > 2$ arbitrario, existe un emparejamiento de puntos PM entre las trayectorias en TR de manera que el costo de sesión (con pesos arbitrarios w_s y w_t) de anonimización TR a través de PM sea cuando mucho c .

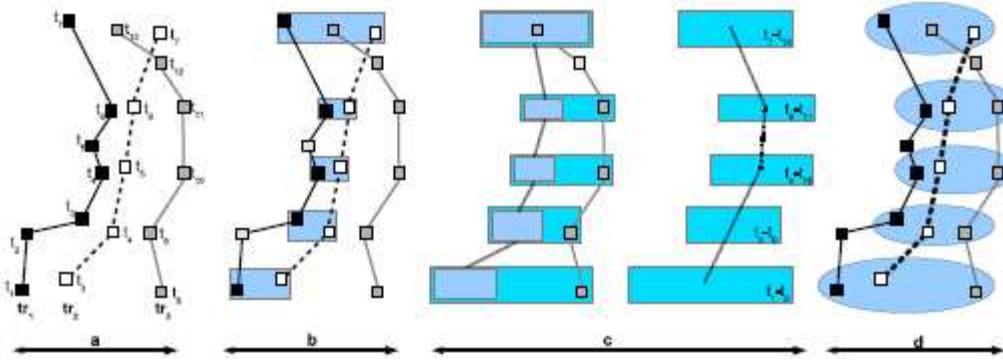
(por ejemplo es $DTA(TR) \leq c$). (DTA, Alineamiento de Decisión de Trayectorias)

Primero se asume que la función de costo de sesión tiene parámetros $w_s = 1, w_t = 0$. Se prueba que DTA es *NP complejo* (En teoría de la complejidad computacional, la clase de complejidad NP-hard es el conjunto de los problemas de decisión que contiene los problemas H tales que todo problema L en NP puede ser transformado polinómicamente en H .) reduciendo el problema de sub secuencia común más larga (LCS).

La Sub secuencia Común Más Larga (LCS) consiste, como su nombre sugiere, en encontrar cuál es el largo máximo que puede tener una palabra que sea sub secuencia de X e Y simultáneamente. Por ejemplo, si consideramos $X = \text{matemáticas}$ e $Y = \text{astronomía}$, una LCS (y en este caso la única) es *atmia*. El largo de una LCS se usa comúnmente como criterio de comparación de palabras, pues está relacionada con la cantidad de "pasos" necesarios para ir de una palabra a la otra mediante operaciones de *inserción*, *eliminación* y *reemplazo* de caracteres.

Problema LCS, dado un entero l y el conjunto de secuencias $SQ = \{sq_1...sq_n\}$ donde cada $sq_i = \{s_1...s_m\}$ es un conjunto de cadenas del alfabeto ordenadas $\Sigma = \{0,1\}$.

El algoritmo *anonTraj* usa la heurística para tratar de obtener una posible alineación de puntos, primero identifica la trayectoria tr_m cuyo costo de sesión de distancia con otras trayectorias es mínimo y marca tr_m como hecho. En cada pasa $OPT_{\sigma_{LCM}}$ encuentra el óptimo emparejamiento entre puntos de una trayectoria no marcada tr_{new} y la actual anonimización de las trayectorias marcadas, y marca tr_{new} . Cada marcaje crea vínculos entre los puntos. Las supresiones de puntos y las generalizaciones son aplicadas de acuerdo al emparejamiento.



- Trayectorias tr_1 , tr_2 y tr_3
- Anonimización tr^* de tr_1 y tr_2
- Anonimización de tr^* y tr_3
- Emparejamiento de puntos usados en la anonimización.

Algoritmo num. anonTraj(G)

Requiere: un (conjunto) grupo de trayectorias G .

Asegura: anonimizar las trayectorias dentro de G .

let $tr_m \in G$ sea la trayectoria cuyo emparejamiento total de distancia con otras trayectorias es mínimo.

let set of trajectories M contains initially tr_m .

repeat

let tr^* be the anonymization of trajectories in M through linked points.

let $tr_{new} \in G - M$ sea una trayectoria seleccionada aleatoriamente

run $OPT_{\sigma_{LCM}}$ to find a min cost matching between the points in tr_{new} and tr^*

create links between the points matched by $OPT_{\sigma_{LCM}}$.

Suprime todos los puntos sin emparejamiento y los puntos directa o indirectamente relacionados con los puntos no emparejados.

$M = M + tr_{new}$

until $M = G$

for all unsuppressed point p of each $tr \in M$ **do**

let PL be the point link containing p .

$p = BB_{PL}$

end for

3.3 Aplicaciones y usos en Turismo.

Después de un estudio de la literatura, no se han encontrado apenas ejemplos prácticos y aplicados del uso de los métodos de anonimización de trayectorias en el campo turístico. A continuación se exponen los dos casos de aplicación de los que se ha podido obtener información.

En (Nergiz et al., 2009) hacen una implementación de su técnica propuesta de Anonimización usando datos sintéticos y datos reales.

Se usan las huellas reales de los dispositivos GPS de taxis en la ciudad de Milán, Italia. Los datos obtenidos comprenden un mes completo de información, compilando así más de 1000 trayectorias de datos reales obteniendo un conjunto de 98544 puntos.

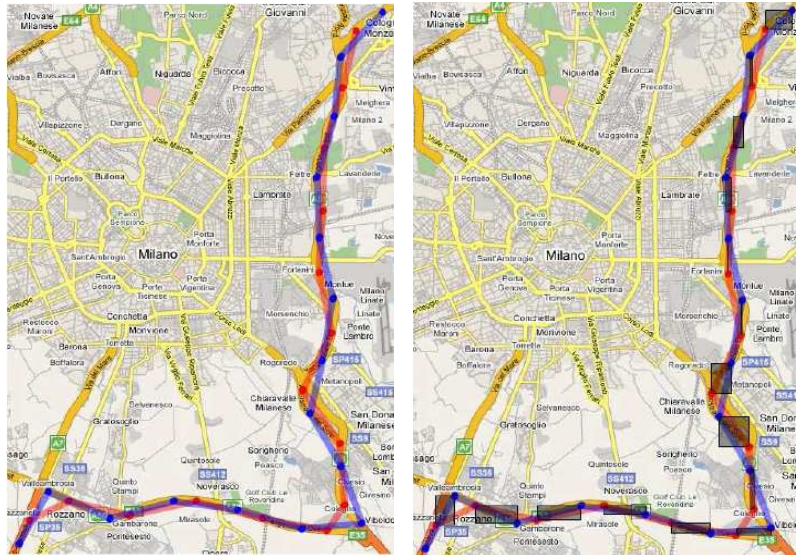




Ilustración 8. Trayectorias originales, anonimizadas y reconstruidas respectivamente.

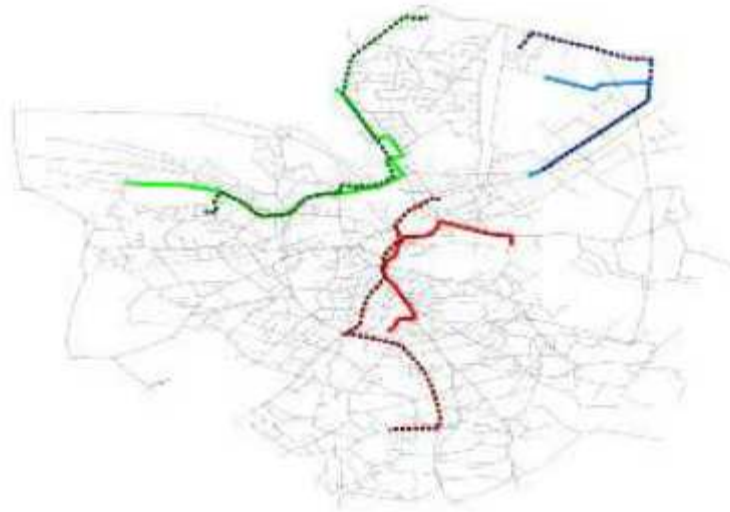


Ilustración 9. Mapa de la ciudad con 3 grupos, cada uno con 2 trayectorias.

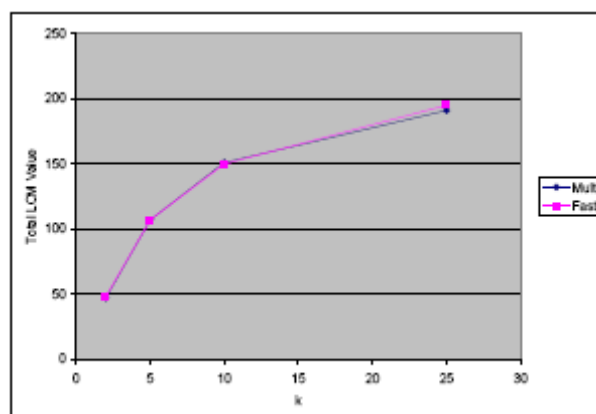


Ilustración 10. LCM para anonimización - Datos Milán.

También analizaron la utilidad del conjunto de datos anonimizados para propósitos de minería de datos. Midieron la desviación de la agrupación original de los resultados, es decir, se

comparan los grupos obtenidos a partir de la trayectoria de datos (la partición de referencia) contra de los grupos obtenidos a partir de la partición de respuesta. Para la evaluación, se utilizó un algoritmo ascendente de agrupamiento, junto con el ERP, que ha sido desarrollado específicamente para trayectorias.

Como el algoritmo requiere especificar el número de agrupaciones como entrada, se experimentó con un rango de 2 a 60 agrupaciones. Debido al gran número de experimentos y la complejidad del algoritmo de agrupamiento se ha utilizado todo el proceso de comparación, se ha recorrido de días de trabajo. Se utilizó un método estándar para evaluar las agrupaciones. Se consideraron las trayectorias de cada par de y se comprobó si ambos están en el mismo grupo de referencia en la partición y si se encuentran en la partición respuesta.

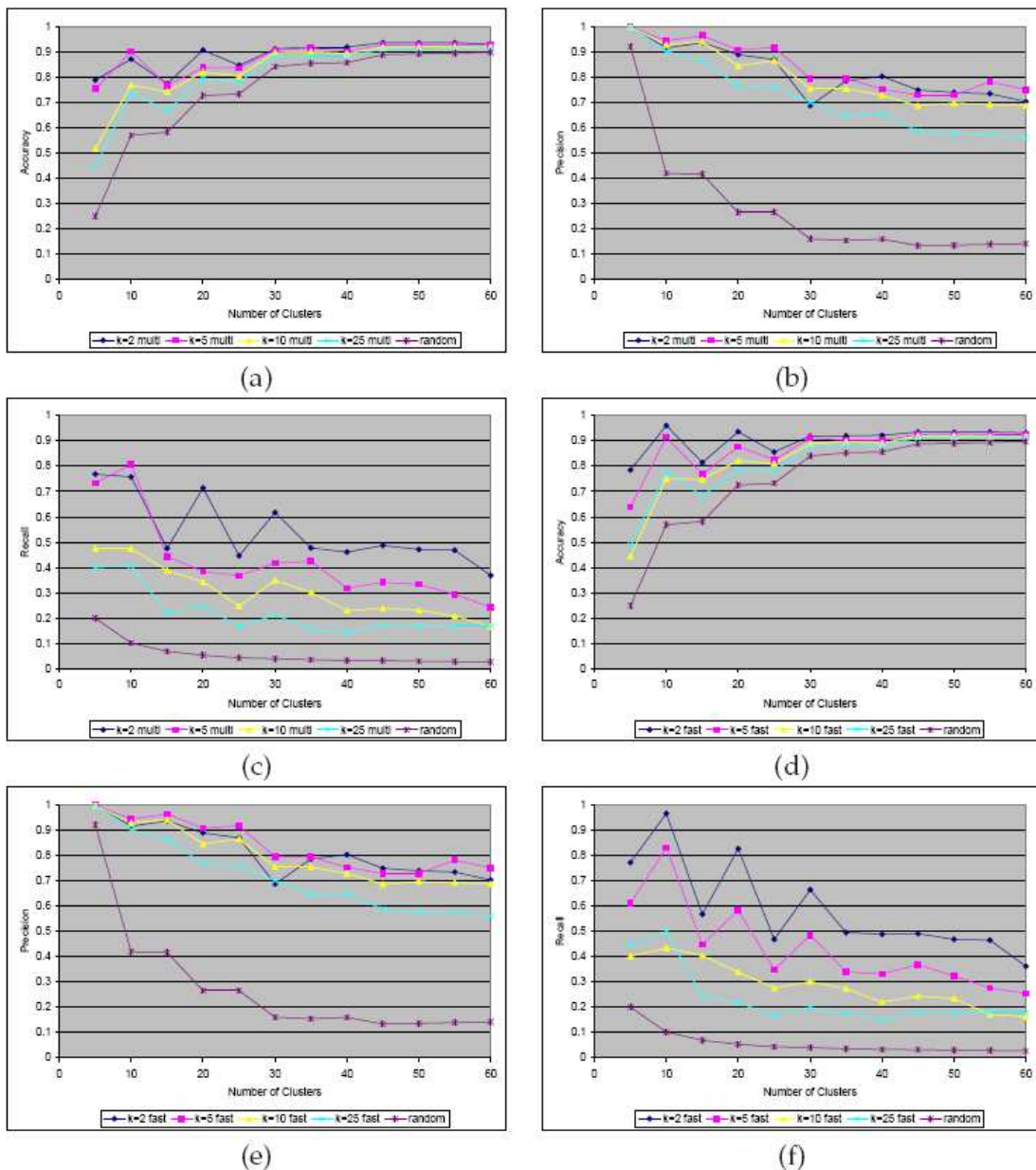


Ilustración 11. Resultados del Agrupamiento - Datos Milán.

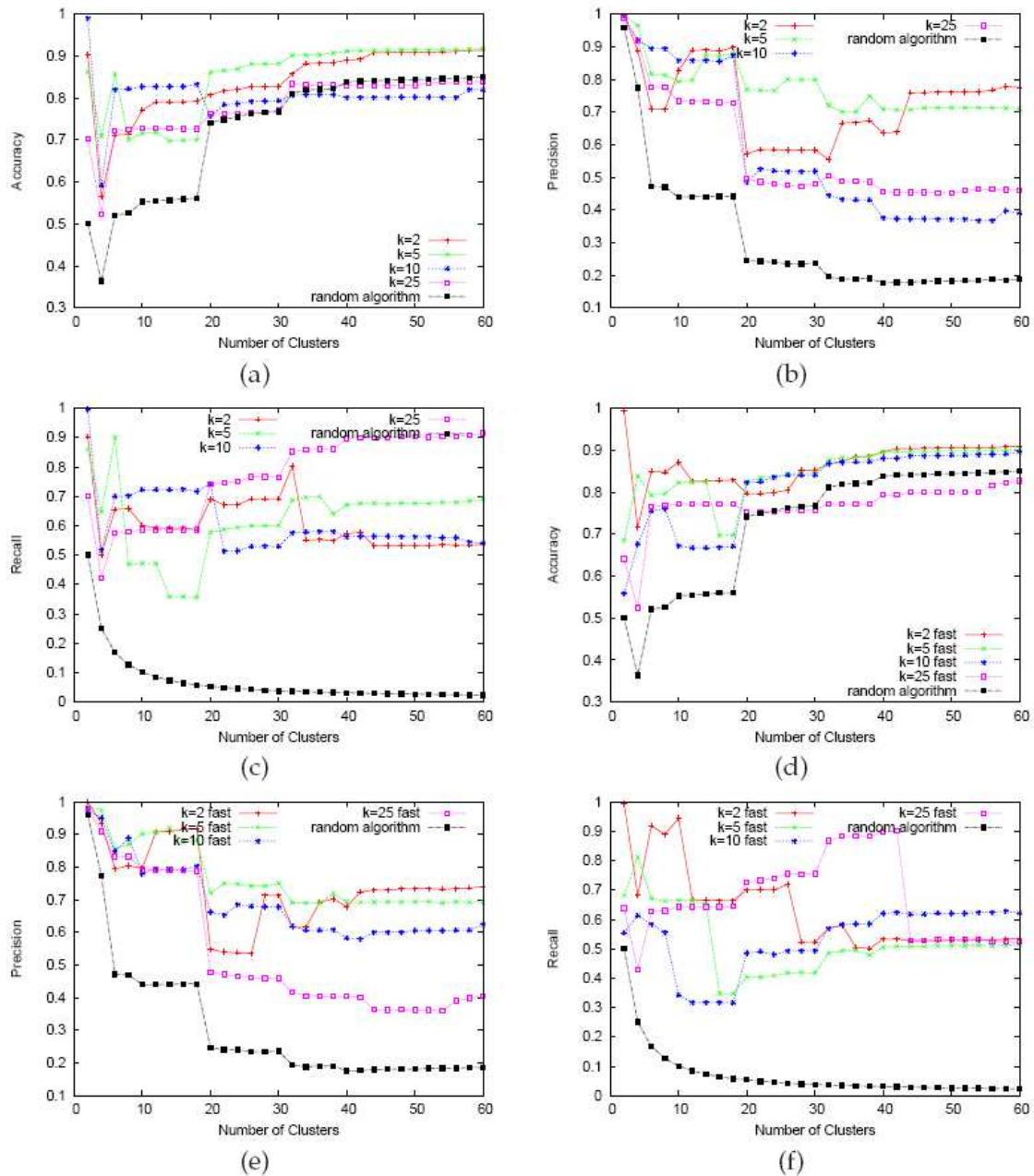


Ilustración 12. Resultados del Agrupamiento - Datos sintéticos.

En las Ilustraciones 10 y 11, las gráficas (a, b, c) muestran el comportamiento de la multi heurística. Mientras que las gráficas (d, e, f) muestran un comportamiento similar para la heurística rápida.

En (Terrovitis&Mamoulis,2008) encontramos otro campo de aplicación para el sector turístico. En este trabajo se considera un ejemplo muy práctico en el cual se habla de tarjetas inteligentes Octopus , que es un sistema de pago usando una tarjeta, en el principio se pensó simplemente en un método sencillo para pagar las tarifas del transporte público en la ciudad de Hong Kong pero después se extendió a simples micro pagos para comprar en tiendas así como usarse para obtener accesos a edificios y escuelas además de servir como identificación, esto podría extenderse como un servicio a turistas, imaginemos que a un turista le podrían vender una tarjeta con prepago de servicios de transporte público, monedero para compras en almacenes e incluso descuentos en sitios turísticos, los turistas podrían evitarse cosas como el

tipo de cambio de moneda y evitarse problemas de tener que moverse con grandes cantidades de dinero, simplemente se abona la cantidad deseada a gastar en las vacaciones dentro de una tarjeta de estas y no hay necesidad de cargar con distintos de divisas, esto solo es un idea de algo viable para el turismo, pero por supuesto que como esto pueden surgir distintos tipos de usos y aplicaciones que pueden ser de uso práctico en el turismo. Básicamente lo que se trata es de buscar maneras de proveer al turismo de más servicios pero claro no se puede dejar de lado la parte de la privacidad.

3.4 Grupos de investigación

Uno de los objetivos del proyecto financiado de turismo en el que se enmarca este trabajo, consiste en identificar los centros de investigación en las líneas de interés del proyecto, con el fin de poder iniciar acciones que den lugar a futuras colaboraciones.

Este aspecto también ha sido abordado en este trabajo de máster y a continuación se detallan los centros que están trabajando en anonimización de datos espacio-temporales. Identificando quienes son los integrantes del grupo especialmente dedicados a esta línea de investigación.

Integrantes	Centro de Investigación	Proyectos	Web
B aik Hoh , Marco Gruteser.	ECE Department, Rutgers, The State University of New Jersey, USA.	Los principales intereses de investigación se encuentran generalizados en los sistemas inalámbricos, con especial atención a la situación consciente de la creación de redes, medición, y la privacidad de la ubicación. Muchos de estos problemas están motivados por la futura comunicación inter vehicular y aplicaciones de transporte.	http://www.winlab.rutgers.edu http://www.winlab.rutgers.edu/~gruteser/
Gabriel Ghinita.	Dept. of Computer Science Purdue University, USA.	Contexto consiente Geo-espacial de control de acceso, Micro datos anónimos, Consultas privadas basadas en la localización, P2P y redes de gestión de datos.	http://www.cs.purdue.edu/ http://www.cs.purdue.edu/homes/ghinita/
Peter Rupel, George Treu, Axel Küpper, Claudia Linnhoff – Popien.	Mobile and Distributed Systems Group, Institute for Informatics, Ludwig - Maximilian University Munich, Germany.	Servicios basados en redes celulares como GSM o UMTS ,Tecnologías para LBS pro activos, , Middleware para los servicios sensibles al contexto, Modelado de información de contexto , Política de gestión basada en redes celulares , Comunicación entre autos.	http://www.mobile.ifi.lmu.de/
Manolis Terrovitis, Nikos Mamoulis.	Department of Computer Science, University of Hong Kong.	Bases de datos espaciales, multimedia y espacio-temporales , minería de datos, agrupamiento de datos en espacios de alta dimensión, preservación de la privacidad en bases de datos y redes sensoriales.	http://www.cs.hku.hk/ http://www.cs.hku.hk/people/profile.jsp?teacher=nikos

Integrantes	Centro de Investigación	Proyectos	Web
Mehmet Nergiz, Yücel Saygin.	Sabanci University, Istanbul Turkey.	Seguridad y Privacidad de las cuestiones planteadas por la integración de datos y técnicas de minería de datos distribuidas, cuestiones de seguridad en las bases de datos; Criptografía Aplicada y Seguridad de múltiples técnicas de cálculo, K-anonimidad en bases de datos.	http://www.sabanciuniv.edu/eng/anasayfa/anasayfa.php http://people.sabanciuniv.edu/ysaygin/go.php?page=welcome http://www.cs.purdue.edu/homes/mnergiz/research.xml
Maurizio Atzori	KDD Lab., ISTI-CNR, Pisa, Italia.	Razonamiento y descubrimiento de conocimiento geográfico, descubrimiento de conocimiento y minería de la red, descubrimiento de patrones y preservación de la privacidad en la minería de datos.	http://www-kdd.isti.cnr.it/ http://www.di.unipi.it/~atzori/
Baris Güc	Swiss Federal Institute of Technology, Zurich, Switzerland.	Towards Trajectory Anonymization: a Generalization-Based approach.	http://www.inf.ethz.ch/
Fosca Giannotti	KDD Laboratory. ISTI-CNR, Istituto di Scienze e Tecnologie dell'Informazione "A. Faedo", Pisa, Italia.	Mobility, Data Mining and Privacy. Libro, Minería de datos espacio-temporales.	http://www.isti.cnr.it/index.html http://www.kdd.isti.cnr.it/userpage.php?user=7
Dino Pedreschi	KDD Laboratory Dipartimento di Informatica Università di Pisa, Pisa, Italia	Mobility, Data Mining and Privacy. Libro	http://www-kdd.isti.cnr.it/ http://www.di.unipi.it/~pedre/

Tabla 1 Grupos de Investigación.

4. Generación de perfiles en datos secuenciales.

4.1 Data Mining espacio-temporal.

Durante los últimos cinco años, ha habido intentos para extender muchas técnicas para extender los métodos de descubrimiento de conocimiento al contexto de datos espaciales, como los métodos de aprendizaje de reglas de asociación, de descubrimiento de patrones, agrupación o clasificación, predicción y análisis de series de tiempo. En gran parte de esta investigación se analizan algunas de las clases simples de modelos y se centra principalmente en los aspectos algorítmicos, dando a menudo, técnica de aproximación que permiten adaptar los algoritmos clásicos a este tipo de datos secuenciales. Sin embargo, la investigación en este campo aún no ha producido un marco teórico para la minería de datos espaciales. (Giannotti & Pedreschi, 2008)

Esto hace que la investigación en minería de datos en el contexto de objetos en movimiento sea más difícil. Y los objetivos en este ámbito son múltiples. En primer lugar, tenemos que descubrir los patrones pertinentes a analizar. En segundo lugar, una taxonomía de estos patrones debe quedar clara para que las nuevas tareas en técnicas mineras puedan ser desarrolladas. En tercer lugar, adecuadas soluciones algorítmicas han de ser propuestas para la aplicación de estas tareas de la minería. Por último, este nuevo campo de investigación podría beneficiarse de un marco teórico unificado.

Cuando se habla de objeto en movimiento o datos de trayectorias, el ejemplo más usado es el que representa el tráfico en algunas ciudades o las provincias. Una de las tareas que obviamente gustaría estudiar son los atascos de tráfico, predecir los atascos de tráfico y descubrir las relaciones entre los atascos de tráfico. Un ejemplo típico es:

Buscar todos los atascos de tráfico en Pisa entre 7 y 9 de la mañana. **(1)**

Los atascos de tráfico pueden ser definidos en términos de la densidad y la velocidad del tráfico y hay una clara relación con los métodos de agrupación o clasificación (*clustering*). También es evidente que la detección de atascos de tráfico se suele hacer sobre determinados fragmentos de los datos. Puesto que hay muchas formas en que las distancias o medidas de similitud entre las trayectorias se pueden definir, muchas variaciones de la agrupación son posibles. Típico de los objetos que se mueven es que tienen velocidad, y la agrupación se puede dirigir de manera similar para detectar objetos en movimiento rápido. En el siguiente ejemplo se pide la agrupación de los coches, bicicletas y peatones:

Encontrar tres grupos de objetos que tienen velocidad similar (lento, medio y rápido). **(2)**

Las propiedades físicas de las trayectorias, como la velocidad, la aceleración y la longitud, se puede esperar que desempeñen un papel importante en el descubrimiento de conocimiento en objetos en movimiento.

En muchos casos, los diferentes atascos de tráfico están relacionados temporalmente. Relaciones espacio-temporales entre los fenómenos pueden expresarse mediante reglas de asociación, como por ejemplo.

Atasco de tráfico (Pisa, 7.30) $\Rightarrow$ atasco (Lucca, 8.30 AM), **(3)**

Lo que significa que cada vez que el primer evento (un atasco de tráfico en Pisa a las 7.30 AM) se produce, por lo general es seguido por el segundo (un tráfico atasco en Lucca a las 8.30 AM). Una versión más general de esta regla podría ser:

Atasco (Pisa, t) $\Rightarrow$ atasco (Lucca, $t + 1$ h), **(4)**

En estos casos suele ser interesante descubrir reglas más generales, por ejemplo una generalización de los ejemplos (3) y (4) sería la siguiente:

Atasco (Pisa) $\Rightarrow$ atasco (Lucca). **(5)**

En el mismo estilo de estos ejemplos, se puede descubrir cuáles son los patrones de comportamiento más frecuentes en trayectoria de datos. Encontrar ejemplos de la clasificación relativa a la trayectoria de datos parece ser más difícil. El problema (2) puede considerarse como la tarea de clasificación de las trayectorias en tres grupos que se definen en términos de la duración, la velocidad y otras particularidades de las trayectorias. Otras tareas de clasificación pueden implicar el reconocimiento de una situación, como distintivo de los atascos de tráfico normales de circulación, o el objetivo de las personas, como si él / ella va a trabajar, ir de compras, hacer un paseo, etc.

Hay muchas más oportunidades para descubrir patrones en las trayectorias de datos. Supongamos que algunos eventos asociados a las características o trayectorias como pueden pasar en el lugar A, B, C, . . . y, a continuación, se pueden encontrar patrones de transiciones habituales como:

$A \rightarrow B \rightarrow C$ **(6)**

Posiblemente asociados con un soporte indica que A, B y C aparecen en ese orden temporal. Además, podemos tener el patrón

$A \rightarrow {}_3B \rightarrow {}_7C$ **(7)**

Donde $_i$ indica demora temporal de $_i$ minutos. Otra clase prometedora de modelos espacio-temporales es la tendencia espacio-temporal. Un ejemplo de una tendencia es:

La velocidad de los objetos aumenta a medida que se alejan de Pisa. **(8)**

Algunos modelos son también en gran medida pre-definidos y pueden ser vistos como una consulta. El siguiente ejemplo, que aborda un aspecto típico de la conducta humana, a saber, la periodicidad, puede considerarse como un ejemplo de ello:

Buscar todos los patrones periódicos (por un período determinado). **(9)**

Un patrón puede ser definido a ser periódico si los mismos objetos se repiten con algunos intervalos fijos de tiempo. También otras pautas de comportamiento, como los atascos de tráfico pertenecen a esta categoría. Una última categoría de las tareas se refiere a la extrapolación de trayectorias de datos. Un ejemplo es la pregunta

¿Cuántas trayectorias cruzaran Pisa mañana a las 5:00 pm? **(10)**

Las categorías discutidas anteriormente de los problemas de la minería se discutirán en detalle en las siguientes secciones de este capítulo, dando una visión general de los enfoques espacio-temporales de la minería de datos.

4.2 Clustering

Una necesidad común en el análisis de grandes cantidades de datos consiste en dividir el conjunto de datos, lógicamente, en grupos distintos, de tal manera que los objetos en cada uno de los grupos que sean más similares entre ellos que respecto a los objetos de otros grupos. Como tal, la agrupación busca un modelo global de datos, por lo general con el principal objetivo de asociar a cada objeto con un grupo (es decir, un clúster), aunque en algunos casos interesa (también) saber donde están localizadas las agrupaciones en los datos espaciales.

En esta sección, nos centramos en el contexto de objetos en movimiento y, por tanto, sobre las trayectorias que describen sus movimientos. En este contexto, la agrupación se compone esencialmente en el intento de encontrar y definir grupos de personas que muestran comportamientos similares.

En cuanto a otras formas de datos complejos, podemos tener dos enfoques principales del problema:

1. La aplicación de métodos genéricos de agrupación mediante la definición de una medida de distancia entre las trayectorias en función de la única información conocida sobre las trayectorias. En este caso, la semántica de la trayectoria de datos está completamente encapsulada en función de la distancia.
2. Se definen nociones y algoritmos ad hoc adaptados alrededor de los tipos de datos específicos.

4.2.1 Clustering basado en distancias de trayectorias.

Definir una distancia entre los objetos implícitamente determina, en gran medida, los objetos que deben ser parte del mismo grupo y, a continuación qué tipo de grupos vamos a descubrir. Saber cómo debe componerse cada grupo depende del método de agrupación que se elija. Por ejemplo, algoritmos basados en el centro como el *k-means* producirá un conjunto de planos esféricos y compactos; métodos jerárquicos organizarán grupos en una estructura multi-nivel de las agrupaciones y sub-grupos; los métodos de clasificación por densidad, forman grupos según la densidad de los objetos, por tanto, pueden no limitar el tamaño del grupo.

Un criterio básico para definir una distancia es considerar los pares similares de objetos que siguen aproximadamente la misma trayectoria espacio-temporal, es decir, en cada momento instantáneo están aproximadamente en el mismo lugar. Esencialmente, por la agrupación de objetos con esa distancia, podemos responder a preguntas del siguiente tipo:

¿Qué individuos de una población se mueven juntos? **(11)**

Cada grupo representará a encontrar, en función del contexto, un grupo de amigos viajen juntos, un rebaño de animales, etc.

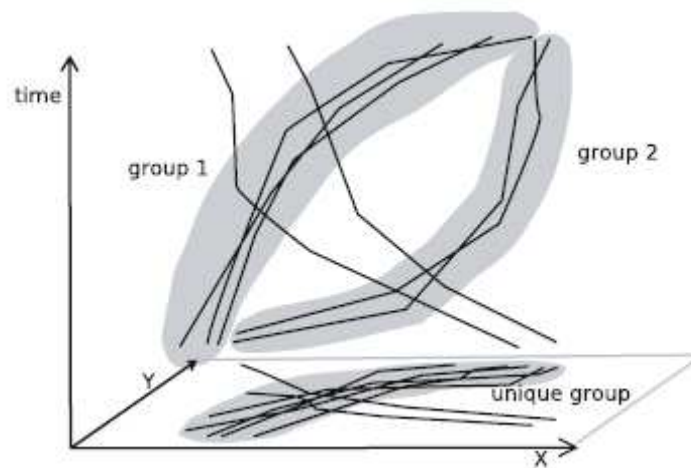


Ilustración 13. Trayectorias en coordenadas espacio temporales.

Tomamos nota, en particular, que todas las trayectorias agrupadas siguen similares caminos, como puede verse en su proyección espacial representado en la parte inferior de la ilustración 12, pero con diferentes velocidades y, por tanto, diferentes tiempos: los del primer grupo comienzan a avanzar a una velocidad rápida y van disminuyendo, mientras que los del segundo grupo comienzan lento y luego van acelerando. Una manera sencilla de modelar esta comparación es representar las trayectorias como vectores de longitud fija de las coordenadas y luego comparar esos vectores por medio de alguna medida de distancia estándar usada en series de tiempo, como la Distancia Euclidiana (la más común).

Una solución alternativa se da en (Nanni, 2002), donde la distancia espacial entre dos objetos es computada para cada instante de tiempo, y los resultados se agregan a obtener la distancia total, por ejemplo, calculando el valor medio, mínimo o el máximo. Avanzando en la misma forma, al mismo tiempo, a veces es demasiado restrictivo para descubrir información útil y, por tanto, la limitación temporal se podría eliminar. En estos casos, podríamos ver a los grupos de objetos que siguen el mismo camino (es decir, la proyección temporalmente orientada al espacio de una trayectoria), pero en cualquier momento, por lo tanto, la formulación de las solicitudes del tipo.

Encontrar grupos de personas que se desplazan a lo largo de las mismas carreteras, **(12)**

Por ejemplo, barcos siguiendo el mismo itinerario para cruzar un mar, o coches siguiendo el mismo camino desde el domicilio hasta el lugar de trabajo y de vuelta, etc.

En la parte inferior de la ilustración anterior, donde las trayectorias son espacialmente proyectadas en el plano X-Y se muestra un ejemplo de esto, y el resultado es un único grupo de objetos que siguen el mismo camino, aunque en diferentes momentos y velocidades. Podemos encontrar algunos métodos generales que producen unos resultados similares. Uno de ellos es la comparación de parejas de series de tiempo, permitiendo (dinámicamente) pandeo de tiempo, es decir, una transformación no lineal del tiempo, de modo que el orden de aparición de los lugares de la serie se mantiene, pero posiblemente comprimiendo/expandiendo los tiempos de movimiento. Otro método, propuesto en (Agrawal et al, 1995), consiste en calcular la distancia como la “longitud de la menor sub-secuencia común” (LCSS) de las dos series, esencialmente formulados como un problema de edición-distancia.

Un último paso en la flexibilización de las restricciones impuestas a las agrupaciones consiste en no exigir una estricta co-ubicación de trayectorias y rutas, pero sólo preguntar al grupo de

objetos que realicen movimientos similares, como ir en la misma dirección o la realización de la misma vueltas (es decir, se convierte de la misma amplitud, sea cual sea la dirección absoluta). El primer ejemplo, puede ser simplemente modelado definiendo como similar cualquier pareja de objetos que siguen aproximadamente el mismo camino, pero que permite el traslado espacial, tal como se propone en (Vlachos et al, 2002) a través de un traslado-invariante, extensión no métrica del mencionado LCSS. Un paso más se logra donde una distancia que es también rotación-invariante se propone, por tanto, nos permite responder a las solicitudes del tipo:

Encontrar grupos de objetos que realizan secuencias similares de los cambios (o no cambios) en su dirección. **(13)**

Por otro lado, debemos mencionar la existencia de otros métodos de agrupación basados en definir nuevos criterios que describan a los datos espacio-temporales a partir de características extraídas de la serie, en lugar de comparar la serie sí mismos (Liao, 2005). Por ejemplo, podríamos extraer todos los pares de valores consecutivos en cada serie (en nuestro contexto, lugares consecutivos dentro de cada trayectoria), y luego simplemente contar el número de parejas que comparten las dos series en comparación, como se propone en (Agrawal et al, 1995).

4.2.2 Clustering de Trayectorias específicas.

Los métodos basados en pura distancia imponen algunas limitaciones a dos niveles, el nivel de expresividad y el nivel de desempeño. Esto es principalmente debido a la fuerte separación entre los criterios de similitud y el esquema de agrupamiento, que se usa sin conocimientos exactos de su semántica, usualmente solo asumiendo la distancia como medida. Por ejemplo, cualquier método basado en una idea de centro o, en general, representante de un grupo, tiene que calcular los grupos de una manera que sea coherente con la función de la distancia adoptada.

El ejemplo más destacado es el algoritmo K-means, donde el representante es generalmente calculado como el objeto (posiblemente nuevo, que aún no está presentes en los datos de entrada) que minimiza la distancia media entre él y todos los objetos de la agrupación. En otros casos, el modelo de clúster solicitado no se basa en ninguna idea de distancia en absoluto o, al menos, no toda la distancia entre trayectorias. Un ejemplo de modelo de clúster que no requiere ninguna noción de distancia explícita entre trayectorias es la generalización del ejemplo **(11)**, donde la desviación entre las trayectorias en un clúster es expresada como ruido en una formulación probabilística, por ejemplo el siguiente problema.

Encontrar grupos de objetos que sigan una trayectoria común, permitiendo una cantidad limitada de ruido aleatorio. **(14)**

En (Gaffney & Smith, 1999) se propone una mezcla de un modelo basado en métodos de agrupación continua de trayectorias, que agrupa a los objetos que son probablemente generados desde un núcleo común de trayectoria añadiéndoles ruido Gaussiano. Otro modelo es cuando el representante de un grupo no es una trayectoria, sino un modelo de Markov que intenta explicar la transición entre una posición y la siguiente, generando posiciones discretas a priori. Más específicamente, los modelos ocultos de Markov (HMMs) se utilizan para modelar las agrupaciones (clústeres) y una mezcla de modelo de enfoque, y el algoritmo EM, en particular, es adoptada por el parámetro de estimación de tareas.

Un enfoque alternativo se basa en la búsqueda de sub-sectores de trayectorias que coinciden lo suficientemente. En (Hwang et al, 2005), las trayectorias se representan como una pieza de manera lineal, posiblemente con los segmentos que faltan (por ejemplo, debido a la desconexión de un teléfono de su red celular). A continuación, un estrecho intervalo de tiempo para un grupo de trayectorias se define como el intervalo máximo, tal que todos los objetos son pares cercanos uno del otro. Los grupos de trayectorias están asociados con un peso que expresa la proporción del tiempo en que las trayectorias están cerca y, a continuación, el problema de la minería es encontrar todos los grupos de trayectoria con un peso más allá de un determinado umbral. A continuación, el método intenta descubrir el tamaño máximo, la máxima extensión temporal de los grupos de segmentos de cerca de la trayectoria. Desde este punto de vista, un similar pero simplificado objetivo se persigue en (Li et al, 2004). Aquí, una extensión de una micro-agrupación para objetos en movimiento se propone, que agrupa a los segmentos rectilíneos de las trayectorias de los que se establecen dentro de un rectángulo de tamaño determinado en un intervalo de tiempo. Incluso en este caso, la proximidad espacial se decide a través de umbrales (el tamaño del rectángulo), mientras que el tamaño del grupo y la extensión temporal es maximizada, en este caso la restricción a considerar sólo son los intervalos de tiempo. Por último, un enfoque diferente a un problema similar se ha propuesto recientemente en (Lee et al, 2007). Las trayectorias se representan como secuencias de puntos sin información explícita temporal y una simplificación heurística se aplica a cada partición de la trayectoria en un conjunto cuasi-lineal de segmentos. Después, todos los segmentos se agrupan por medio de la densidad de un método basado en la agrupación, y al final un representante de la trayectoria se calcula para cada grupo resultante.

Un equilibrio entre los enfoques basados en la distancia y la trayectoria específica se produce cuando la distancia no tiene que comparar las trayectorias de la totalidad, sino sólo algunas partes de ellas, centrándose en un intervalo de tiempo que se da como un parámetro de la función de la distancia. Que lleva a abordar los problemas del siguiente tipo:

Encontrar objetos que se mueven juntos en algunos (desconocidos) intervalos de tiempo de tamaño mínimo **(15)**

Aquí, las trayectorias se agrupan por medio de un algoritmo genérico basado en la densidad, donde la distancia adoptada es el promedio de distancia espacial entre las trayectorias dentro de un determinado intervalo de tiempo, que es un parámetro de la distancia. Entonces, para cada intervalo de tiempo T , el algoritmo se puede ejecutar centrándose en segmentos de la trayectoria por la que se establecen dentro de T . El objetivo final es descubrir que intervalo de tiempo T resulta en los grupos de mejor calidad y, a continuación, volver estas agrupaciones, junto con una muestra de T .

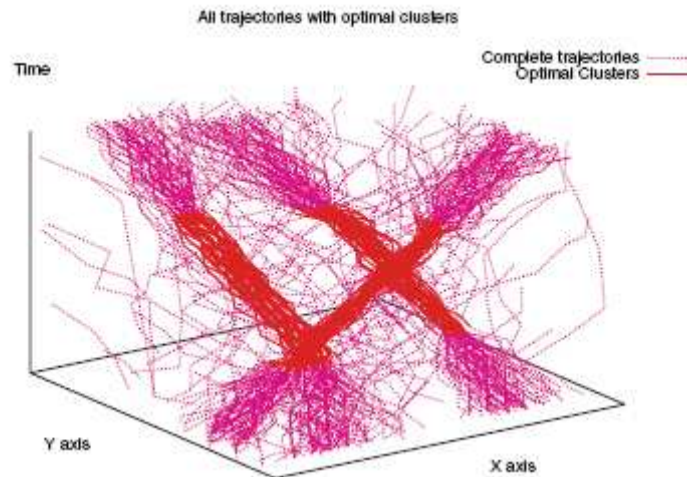


Ilustración 14. Agrupaciones sobre un intervalo de tiempo.

La ilustración anterior representa un conjunto de trayectorias que forman tres grupos (además de algo de ruido) y muestra el intervalo de tiempo óptimo (donde los grupos estén claros) como segmentos de la trayectoria más oscuros. Un objetivo similar se persigue en (Kalnis et al, 2005), pero desde una perspectiva diferente. Aquí, los autores consideran como objetos en movimiento como una asociación con una posición espacial de un conjunto de líneas de tiempo y afrontar el problema de descubrir agrupaciones basadas en la densidad espacial que aproximen persistentemente a lo largo de severas líneas de tiempo contiguas, como se muestra en la siguiente ilustración donde una agrupación que es persistente en tres unidades de tiempo es encontrada.

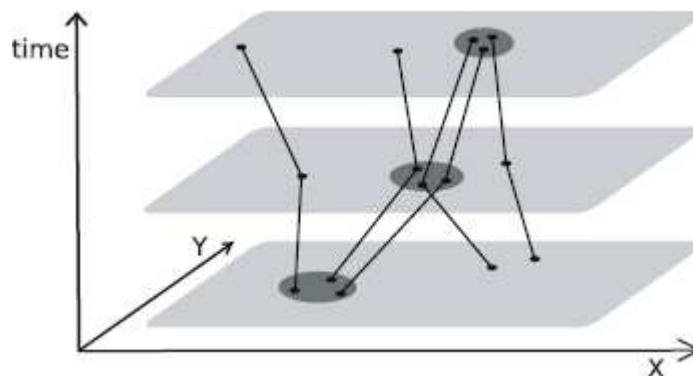


Ilustración 15. Una agrupación en movimiento.

La persistencia de una agrupación significa que los individuos contenidos en un grupo o agrupación en algunas líneas de tiempo son aproximadamente los mismos que aparecen en una agrupación en la siguiente línea de tiempo. Tómese en cuenta que la primera línea de una agrupación en movimiento y su última podrían compartir sólo algunos objetos (en nuestro ejemplo, sólo uno), o incluso ningún objeto en absoluto, ya que gradualmente y por etapas, las migraciones en los grupos son permitidas, y ninguna verificación global se lleva a cabo, es decir, aquí el enfoque está en el espacio de las regiones cubiertas por grupos y no en las trayectorias que contienen. Por esta razón, este enfoque puede considerarse un caso límite entre la agrupación (con miras a encontrar una compartimentación de los objetos) y la frecuencia de los patrones (el fin de encontrar regularidades que implican a muchas personas),

las agrupaciones siendo (segmentos de) las trayectorias implicadas y los patrones siendo las localizaciones espacio temporales donde las agrupaciones se encuentran.

Todos los enfoques descritos anteriormente comparten una propiedad común: intentan buscar grupos de objetos que se mueven fundamentalmente en conjunto con distintas variantes. Relajando requisitos temporales, podemos buscar grupos de objetos sólo buscando en los caminos que siguen.

4.3 Patrones Locales Espacio-Temporales.

La minería espacio-temporal se entiende como buscar patrones de las representaciones concisas de comportamientos interesantes de simples objetos en movimiento o grupos de objetos. En particular, en esta sección, estamos interesados en la minería de patrones locales, es decir, los patrones que tienen por objeto caracterizar potencialmente pequeñas porciones de los datos espaciales, tales como sub-conjuntos de individuos, pequeños intervalos de tiempo limitado o de regiones del espacio.

El tipo de comportamientos interesantes que queremos descubrir depende estrictamente de el contexto y por lo general es especificado por la selección de un sub-conjunto de todos los patrones posibles y, posiblemente, la aplicación de ciertas restricciones sobre cómo se producen los patrones en los datos, es decir, en los comportamientos que resumir.

Aparte de la especificación de los patrones, el patrón de la minería depende de si el enfoque específico de la tarea es interesante en la búsqueda de patrones o en la búsqueda de sucesos de los patrones (es decir, donde y cuando se producen y que se trata). En una búsqueda directa, podemos especificar las hipótesis de espacio H , el espacio de todos los modelos considerados en nuestra búsqueda, que suele ser muy grande, y tratar de identificar todos los patrones frecuentes $h \in H$. Alternativamente, se podría especificar un conjunto de patrones interesantes (o hipótesis) H de antelación, H por lo general es relativamente pequeño, y pedir que todas las apariciones coincidan con estos patrones en los datos.

Una característica adicional de los dos tipos de problemas es que los patrones en búsquedas directas son generalmente (pero no necesariamente siempre) muy simples e incluyen individuos simples, después, incluyen algunas restricciones sobre el número de sucesos. Por el contrario, las búsquedas inversas en los patrones son generalmente bastante complejas e implican un conjunto de individuos y las limitaciones se encuentran en el tamaño o la composición de dicha serie. Los algoritmos aplicados en el modelo de minería siempre dependerán de los datos objeto de la búsqueda, la información derivada puede incluir conjuntos o secuencias de eventos como; los lugares visitados por la trayectoria, o eventos espacio-temporales tales como las maniobras realizadas: Vueltas en U, detenerse, aceleraciones extremas, etc. En algunos casos, podemos derivar información que describe a la población, por ejemplo atascos de tráfico o la última semana en la ciudad, así enfocándose más en el comportamiento de grupos. En algunos patrones de descubrimiento de tareas, ambos tipos están contemplados, por ejemplo queremos descubrir si un atasco de tráfico aquí está usualmente seguido de un accidente de tráfico más adelante, o si alguien esta atorado en el tráfico, usualmente después tiene un accidente.

4.3.1 Extrayendo patrones frecuentes.

Los patrones frecuentes son un elemento básico de la minería de datos, una aproximación simple y muy común en la minería de patrones frecuentes espacio-temporales consiste en una solución basada en extracción de características relevantes. En primer lugar, conjuntos de características se derivan de los datos, produciendo eventos, valores de atributos o, más en general, predicados espacio-temporales que describen cada trayectoria. A continuación, los algoritmos de minería genéricos se aplican sobre la nueva representación de los datos, los conjuntos de extracción frecuente, asociación normas o frecuencia de las secuencias de características. Siguiendo este enfoque, la semántica de los datos espacio temporales tomados en consideración durante la etapa de tratamiento previo está implicada en modo alguno la fase de la minería. Con esta aproximación, la variedad de patrones de frecuencia que podemos obtener con esta simplificación del problema es todavía amplia, como reglas de la forma:

Longitud (trayectoria) > 50 kilómetros $\Rightarrow$ velocidad media (trayectoria) > 60 kilómetros **(16)**

Obviamente la selección de los atributos (características) a extraer es un aspecto crucial en el proceso de aprendizaje, dado que define de una vez por todas, el patrón espacial a buscar. Una familia básica de características para trayectorias de objetos en movimiento consiste en características basadas individualmente, por ejemplo esas que describen el comportamiento de cada objeto separado de los demás, por ejemplo podemos tener:

- Agregados espaciales y/o temporales (la longitud del trayecto cubierto, la cantidad de tiempo gastado en el centro de la ciudad, el mínimo y el máximo promedio de velocidad, la dirección más frecuente seguida, etc.)
- Eventos espaciales (visitando algunas regiones espaciales predefinidas o visitando dos veces el mismo lugar).
- Eventos espacio-temporales (maniobras localizadas temporalmente como realizar una vuelta en U, paradas repentinas, aceleraciones inesperadas o comportamientos de larga duración como cubrir cierto segmento del camino en algún momento y cubrirlo nuevamente mas tarde en la dirección opuesta) serian secuencias de la forma:

Visitas (x , Plaza del mercado) $\rightarrow$ parada repentina (x) $\rightarrow$ vuelta en U (x) **(17)**

Esto combina eventos espaciales (visitar una región dada) con comportamientos temporales simples (paradas y vueltas en U).

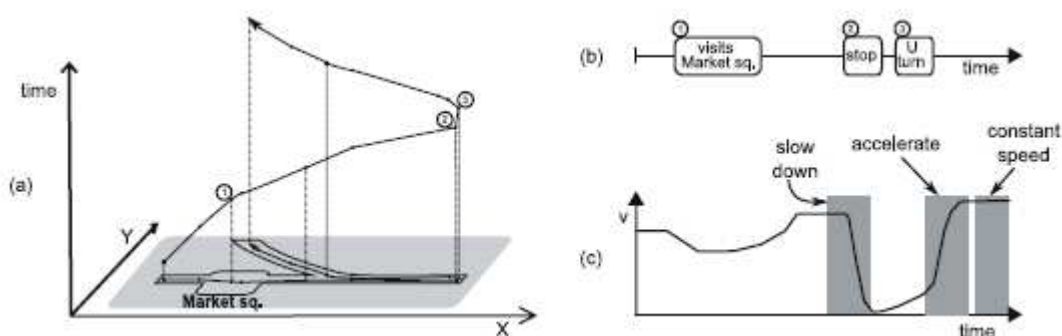


Ilustración 16. Simple trayectoria (a), Eventos derivados espacio temporales (b), Variación de la velocidad (c).

Otras características posibles pueden también incluir predicados en el estilo espacio-temporal, que permiten expresar alguna forma de topología espacio temporal entre referencias de

regiones espaciales y trayectorias teniendo una incertidumbre espacial (por ejemplo, las localizaciones no son puntos sino límites circulares que contienen la posición real). Una muestra de predicado espacio-temporal es *A veces Definitivamente Adentro* (x, A), que significa que al menos hay un instante de tiempo (*A veces*) tal que el objeto x esta seguramente en la región A (*Definitivamente Adentro*), tomando en cuenta la incertidumbre. Así entonces podemos obtener reglas de la forma;

A veces Definitivamente Adentro (x , *hospital*) $\rightarrow$ *siempre_posiblemente_adentro* (x , *centro de la ciudad*). **(18)**

En algunos contextos, características más complejas se requieren para describir el comportamiento de grupos en movimiento, ejemplos de esto ya han sido mencionados como los atascamientos de tráfico, interacciones entre objetos (por ejemplo, al menos n individuos se encuentran en alguna parte) o más general la aparición de cualquier patrón que describa algunos comportamientos predefinidos.

Del mismo modo, una característica basada en grupos puede ser definida como la co-ocurrencia de varios eventos simples en el mismo lugar al mismo tiempo, por lo tanto, basándose en herramientas de análisis que la búsqueda de regiones concurridas en el espacio-tiempo, tales como la exploración espacial estadística de (Kulldorff, 1997), que busca cilindros espacio-temporales (es decir, circular regiones dentro de un intervalo de tiempo), donde la densidad de eventos es más alta que fuera, y sus extensiones .

Un cierto grado de flexibilidad puede ser añadido a la función de enfoque, mediante la adición de información temporal a las características extraídas, 1ue corresponde a fecha y hora de los eventos espacio-temporales, y la extracción de los atributos dinámicos, es decir, los atributos que tengan valores dependientes del tiempo. Los sellos de tiempo permiten extraer patrones más detallados que describen igualmente las relaciones temporales entre eventos. Por ejemplo, secuencias con los tiempos de transición entre las características, eventos consecutivos, o más series de eventos en general con las limitaciones temporales entre ellos. Similarmente, atributos dinámicos proveen esencialmente series de tiempo que pueden ser obtenidas por medio de reglas para extraer secuencias de ellos, por ejemplo las asociaciones entre formas típicas que ocurren en las series, como la siguiente:

Intenso_frenado $\rightarrow$ *intensa_aceleración* $\rightarrow$ *velocidad_constante* **(19)**

La alternativa opuesta frente al enfoque basado en patrones de frecuencia consiste en el análisis directo de trayectorias, por ejemplo, para descubrir los caminos frecuentemente seguidos por los automóviles en el centro de la ciudad, las frecuentes maniobras realizadas por los animales depredadores o presas de caza, etc. Eso significa, en particular, que no hay una discretización prioritaria u otra forma de tratamiento previo de la ordenación de la información espacio-temporal, y por lo tanto, la semántica de datos espacio-temporales pueden desempeñar un papel en la fase de minado. Una primera consecuencia de este escenario es que el concepto de patrón frecuente tomado de minería de datos transaccionales, es decir, un patrón que exactamente ocurre varias veces en los datos, por lo general no se puede aplicar. De hecho, la continuidad del espacio y del tiempo por lo general hace que sea casi imposible ver a una configuración ocurriendo más de una vez a la perfección de la misma manera, y, por tanto, algún tipo de tolerancia a las pequeñas perturbaciones es necesaria.

El problema de la continuidad antes mencionado puede ser abordado en al menos dos formas complementarias, primero teniendo en cuenta las pautas que se presentan en forma de segmentos de trayectoria y la búsqueda de la aproximación de los casos en los datos y

segundo, teniendo en cuenta los patrones que se encuentran en la forma de mover las regiones de los intervalos de tiempo, como los cilindros o tubos espacio-temporales que, en cierto sentido, representan un segmento de una trayectoria más una limitada aproximación o incertidumbre y contando todos los sucesos de la trayectoria como segmentos plenamente contenidos en el movimiento de regiones. El trabajo en (Cao et al, 2005) proporciona un ejemplo del primer enfoque: una trayectoria de aproximación por medio de una sucesión de segmentos espaciales obtenidos a través de una simplificación y luego los patrones son extraídos fundamentalmente en forma de secuencias de segmentos contiguos espaciales, en particular, cada elemento de la secuencia tiene que ser similar a varios segmentos de la trayectoria de entrada, se define la similitud en tres parámetros básicos: la proximidad espacial, la longitud y ángulo.

Las secuencias frecuentes son extraídas como una sucesión de rectángulos de tal manera que su anchura cuantifica la distancia media entre cada segmento y los puntos de la trayectoria que abarca. La ilustración siguiente muestra un modelo simple de este tipo, formado por dos segmentos correspondientes y rectángulos.

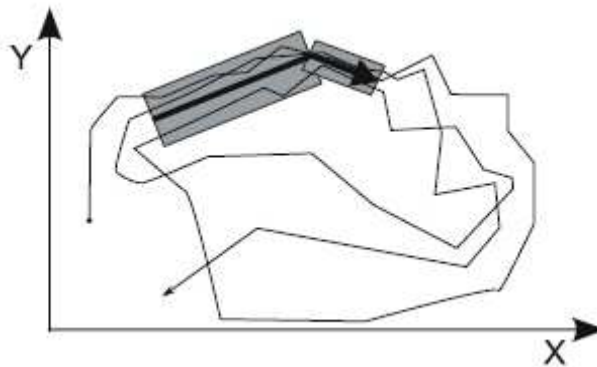


Ilustración 17. Ejemplo básico de patrones espacio-temporales.

El segundo enfoque, basado en las regiones en movimiento, es seguido por (Kalnis et al, 2005), se refiere al descubrimiento de la densidad espacial basado en las agrupaciones que persisten a lo largo de varias fracciones de tiempo. Una meta similar, pero centrada en modelos cíclicos, define el patrón espacio-temporal periódico del problema de la minería (es decir, la búsqueda de patrones de secuencia cíclica en un período determinado) y proponer una forma eficaz y un algoritmo de minado rápido para la recuperación máxima de patrones periódicos. Mientras que el tiempo simplemente se supone que es discreto, las localizaciones espaciales son discretas dinámicamente a través de agrupamientos basados en la densidad cada vez que un patrón periódico, en forma de una secuencia espacial de las regiones, realiza una comprobación para asegurarse de que todas las regiones en el patrón son densas y, por ende significativas.

Algunas variantes de los problemas mencionados se han propuesto y estudiado en los últimos años. Un ejemplo es, donde los patrones en forma de las secuencias de localizaciones son obtenidos, y también la incertidumbre de la localización de los objetos es considerada desde un punto de vista probabilístico (es decir, la posición de cada uno de los objetos sigue una probabilidad dada de distribución sobre espacio alrededor del punto de localización dado). Aquí, los patrones candidatos son construidos sobre una pre-definida discretización del espacio y tiempo, y el apoyo de un patrón se calcula como su apoyo esperado de ubicación de los objetos de entrada.

4.3.2 Recuperación de ocurrencia

Contrario a la extracción de patrones frecuentes de los datos, un usuario puede ya tener algunos patrones específicos en mente y preguntar por todas las ocurrencias. Existen dos tipos de consultas que pueden distinguirse: elementales y sinópticas, mientras que las consultas elementales representan el comportamiento del movimiento de entidades individuales, las consultas sinópticas representan patrones del comportamiento de movimientos colectivos.

Las consultas elementales inversas involucran patrones que pueden ser resueltos desde una trayectoria simple, por ejemplo la consulta:

Encontrar todas las trayectorias que pasen por la ubicación A en los tiempos t_1 y t_2 **(20)**

Esto puede generar muchas trayectorias, sin embargo cada trayectoria por sí misma es suficiente para decidir si el patrón se ha cumplido o no. En la consulta anterior, la localización es especificada explícitamente mientras que la constante temporal corresponde a un rango de búsqueda. Nótese que el modelo no implica la información secuencial. Para incluir información secuencial en nuestra consulta, se podría pedir que después de una ubicación A , una segunda B deba pasarse. En (Hadjieleftheriou et al, 2005), este tipo de consulta que se denomina patrón espacio-temporal de consulta (STP) y se define como una secuencia de la ordenación del territorio, ya sea con predicados de exacto o relativo orden temporal. Alternativamente, se podría incluir información acerca de un derivado del objeto, la velocidad o la dirección de viaje, o solicitar alguna periodicidad en el movimiento de un objeto. En la literatura de bases de datos, las consultas que se concentran en una sola parte de una trayectoria son conocidas como consultas basadas en coordenadas, mientras que las preguntas que se basan en la información secuencial se les llama trayectoria a base de consultas. Las consultas sinópticamente inversas identifican objetos que conforman un específico comportamiento colectivo. Estos patrones tienen como objetivo movimientos simultáneos y la interacción entre los objetos. También son referidos como un grupo de patrones y puede incluir información derivada relativa a todo el grupo de objetos (por ejemplo, la velocidad media). Intuitivamente, un grupo es formado por un número de objetos que están cercanos en espacio en un periodo de tiempo significativo. El algoritmo de (Wang et al, 2003) descubre patrones de grupos móviles en datos de trayectorias donde la ubicación es registrada como puntos arreglados de espacio en tiempo.

Aparte de la definición general de la proximidad espacio-temporal, un grupo puede ser especificado por algunas características de su estructura interna. Por ejemplo, un grupo podría estar encabezado por algún individuo que prevé el movimiento del grupo. Este patrón se llama liderazgo (Ilustración 18, izquierda) y se introdujo por primera vez por (Laube & Imfeld, 2002) en el marco del concepto general de movimiento relativo (REMO). Otro grupo de modelos espacio-temporales básicas de REMO son manada, la convergencia y divergencia. Una manada corresponde a un grupo móvil de patrones que se ha definido anteriormente, mientras que la convergencia y la divergencia simultánea describen el movimiento de objetos hacia o desde algún punto en el espacio (ver ilustración 18, medio y derecha). Es fácil pensar en una extensión de estos patrones a fin de incluir nuevas características de movimiento. Por ejemplo, la velocidad de los coches podría ser necesario que aumente a medida que se alejan de Pisa como en el ejemplo **(8)**, con lo cual se cubrirá en cierta medida el problema de la detección de tendencias. Sin embargo, es importante señalar que no todos los patrones, según lo especificado por REMO y relacionados en la literatura, son derivados directamente de las trayectorias. La recuperación de los sucesos de convergencia o de encuentro de patrones se basa usualmente en un solo instante en el tiempo y asume una velocidad constante y la orientación de los objetos.

Hasta ahora, todos los patrones se basan en un grupo estable de los objetos. Sin embargo, un patrón puede continuar con el paso del tiempo, aunque cambien los miembros del grupo. Por ejemplo, un atasco de tráfico puede prevalecer durante varias horas, mientras que coches nuevos continuamente llegan a un extremo y escapan en el otro extremo. Este fenómeno es llamado **grupo en movimiento** y se refiere a un grupo que mantiene su densidad (y otras propiedades similares, como el tamaño de clúster o el diámetro), aunque diferentes objetos participan en el grupo durante su duración.

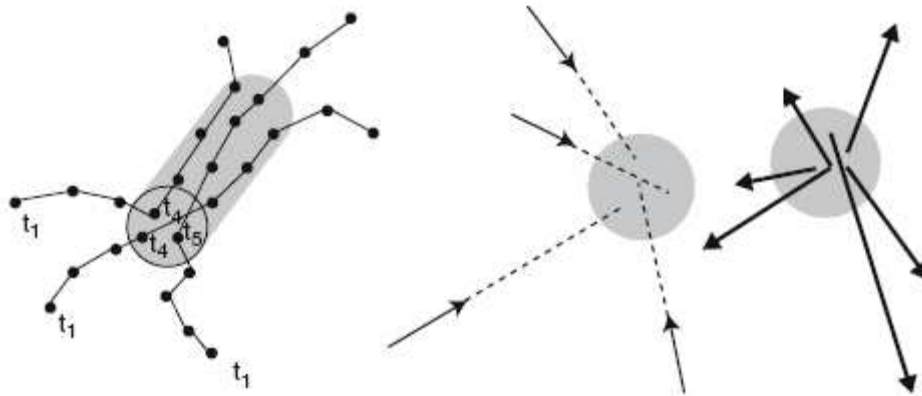


Ilustración 18. Liderazgo de patrones de movimiento.

4.4 Predicción

En la era de la logística “*just-in-time*”, la gestión del tráfico en tiempo real, los servicios basados en la localización y de navegación GPS, la predicción sobre el futuro de la posición o el destino de un individuo u objeto tiene un papel central. Anticipar el movimiento de personas o grupos de objetos de estos sistemas permite tomar las acciones preventivas en caso de retraso, incluso más favorable a fin de evitar aglomeraciones, o para entregar información útil en el tiempo deseado. Por lo tanto, los datos espacio-temporales ofrecen una amplia perspectiva de las tareas de predicción, que incluyen la predicción de las trayectorias y los lugares, la predicción de la densidad, el alcance y eventos, así como la clasificación de las trayectorias.

4.4.1 Predicción de Ubicaciones y Trayectorias.

Durante los últimos años, la predicción confiable de la futura ubicación de objetos en movimiento ha sido de interés principalmente en dos áreas de investigación, sistemas de bases de datos y redes de comunicación inalámbrica. Las bases de datos de objetos en movimiento emplean ubicaciones futuras de objetos, por ejemplo, en búsquedas al vecino más cercano. Estas preguntas requieren sofisticadas estructuras de indexación de las futuras posiciones de los objetos en movimiento. En las redes inalámbricas, la previsión de movimiento futuro es importante para permitir una asignación eficiente de los recursos de la red.

En la literatura de bases de datos, las preguntas se basan en la previsión de la indexación de las posiciones actuales de estructuras y vectores de movimiento. Dada la ubicación actual l_c y el vector de velocidad v_c de un objeto, la futura posición después de tiempo Δt puede ser calculado como $l_f = l_c + v_c \Delta t$. Con este fin, se ha propuesto una estructura de datos en árbol, llamada árbol TPR (Tao et al, 2003), que se ha desarrollado para manejar las preguntas de predicción de rango, consultas de vecino más cercano o el retroceso de las consultas al vecino más cercano sobre el futuro de las posiciones de los objetos en movimiento (Benetis et al, 2006).

La hipótesis subyacente de todas las técnicas es que los objetos que participan continúan su movimiento con el vector de velocidad dado hasta la hora de finalización del intervalo de la consulta. Este supuesto se aplica para el movimiento lineal en los espacios libres de obstáculos, como por ejemplo, los buques, aviones o fenómenos meteorológicos. Sin embargo, no es razonable para redes de calles donde los objetos cambian de dirección y de velocidad en intervalos cortos de tiempo. Tales condiciones inestables se cumplen en las redes de comunicaciones inalámbricas, donde la gestión de la movilidad sirve principalmente dos tareas.

En primer lugar, los recursos deben ser asignados para garantizar una transición de servicios suave si un usuario se mueve de una móvil a otro. Segundo, cuando una llamada entrante llegue, la red debe llamar la menor cantidad de móviles posible en una determinada área. Ambas tareas requieren anticipar el movimiento de los usuarios en el futuro cercano. Bastantes algoritmos han sido investigados para cumplir con esta tarea. Además de la predicción de la ubicación en el futuro cercano, una importante tarea de investigación consiste en anticipar la ruta más probable y el destino de un objeto en movimiento. Por ejemplo, servicios basados en la localización pueden ofrecer servicios más sofisticados, cuando a sabiendas de cuales lugares pasará el usuario y si esta en el camino al trabajo o al supermercado. La premisa general detrás de la predicción de las rutas y destinos es que la gente siga las rutinas diarias o semanales. Normalmente, la gente visita sólo unos pocos lugares con frecuencia, como por ejemplo su casa, lugar de trabajo o restaurante favorito. Además, las personas son criaturas de hábito y seleccionan su ruta actual de un pequeño conjunto de rutas candidatas. En (Karimi & Liu, 2003) adaptan una matriz de transición a las preferencias personales y, por tanto, son capaces de predecir la ruta más probable y el destino de una sola persona con un segmento de tiempo dado. Mientras que Karimi y Liu, basan sus predicciones únicamente en información de enrutamiento, (Laasonen, 2005) incorpora los tiempos de residencia en su modelo. El autor detecta los lugares en que un usuario gasta una cantidad comparativamente grande de tiempo. Estos lugares forman el conjunto de todos los destinos posibles y delimitan las rutas. Laasonen agrupa rutas históricas y compara los tipos obtenidos con la trayectoria actual. Las previsiones de destino pertenecen al tipo de trayectoria más similar y opcionalmente, puede ser condicionada a la hora del día y día de la semana.

4.4.2 Predicción de Densidad.

La densidad de una zona se define como el número de objetos dentro de la zona en proporción al tamaño del área en un punto dado en el tiempo. Es una característica que emerge a través de la interacción de una serie de objetos y cambios en el tiempo. La predicción de densidades promete muchos beneficios, especialmente en el dominio del tráfico. Por ejemplo, un sistema de gestión del tráfico que es capaz de identificar las regiones densas y los cuellos de botella debe contrarrestar esos efectos en el tiempo. El concepto de densidad se extiende a un

intervalo de tiempo donde el enumerador contiene el número mínimo de objetos que están concurrentemente en el área dada. Para calcular la densidad, un cubo espacio-temporal es creado. Cada celda contiene la densidad de un área dada (eje x y eje y) durante algún momento o intervalo de tiempo (eje z). Para la predicción de la densidad (Hadjielefteriou et al, 2003) asume un movimiento lineal de los objetos y computa futuras densidades por extrapolación. Una aproximación diferente considera el desarrollo espacial de cada celda por separado y calcula la densidad más próxima en el tiempo como suma ponderada de las densidades anteriores.

4.4.3 Extrapolación y Predicción de Alcance.

El alcance es una medida dependiente del tiempo, acerca de la publicidad de una localización dentro de una población. Imaginemos que un nuevo restaurante abre en el centro de la ciudad. Después de 1 día el 20% de los habitantes habrán notado la nueva ubicación, después de una semana el alcance se habrá incrementado a un 60%, si el restaurante hubiese abierto en las afueras de la ciudad, solo el 40% de los habitantes habrían pasado después de una semana. El alcance no está limitado a una sola ubicación sino que puede abarcar una red de lugares. Se define como la proporción de la población que pasa al menos por una de las ubicaciones de la red en un cierto periodo de tiempo dado.

Dadas las trayectorias de un GPS de un grupo de personas sobre varios días, el número de contactos con una red dada puede ser fácilmente calculado. El reto está en la extrapolación de la des balanceada e incompleta muestra de las trayectorias. Si las personas en la encuesta no son representativas de toda la población, por ejemplo si viven la mayoría en una sola parte de la ciudad, las muestras de datos necesitan ser estratificados para proveer un alcance imparcial. Además, las trayectorias incompletas que se originan por defectos de los dispositivos GPS, olvido o abandonos de la encuesta, constituyen un grave problema en las mediciones de las secuencias de días consecutivos.

4.4.4 Predicción de Eventos.

El trabajo en (Brown et al, 2001) estudia el problema de predecir eventos espacio-temporales que están asociados con otras características por ejemplo, existe la probabilidad que algún crimen sea cometido dada cierta región e intervalo de tiempo basado en las localizaciones, los tiempos y las características socio económicas de incidentes pasados. La justificación subyacente es la definición del modelo de densidad de transición, que predice la probabilidad de densidad en espacio-tiempo dados los datos históricos, junto con el proceso de estimación de densidad para descubrir importantes conjuntos de características y ubicaciones de espacio y tiempo.

4.4.5 Predicción en Series de Tiempo Geo Referenciadas.

El cubo de espacio-tiempo se divide en un número de series de tiempo Geo referenciadas, una para cada ubicación del área. En contraste con las series de tiempo normales, las series geo-referenciadas temporales no son independientes una de otra, pero son espacialmente

correlacionadas. El objetivo general en la predicción de series temporales, es, dados los datos históricos en los tiempos $t_1, \dots, t_n$, obtener el valor de algunas variables en el tiempo $t_n + 1$. Otros enfoques aplican modelos de regresión no lineal y explotan la correlación espacio-temporal de los residuos de la regresión. La correlación espacial puede también ser utilizada para reducir los costes computacionales. El resto de la sección demuestra la complejidad de las series de tiempo Geo-referenciadas donde se analiza la migración de clientes y la interrelación entre las tiendas que venden periódicos.

Considere el número de periódicos que se entregan a varias tiendas de reventa. El número debe ser cuidadosamente calculado, también si muy pocos periódicos son entregados puede resultar una pérdida de beneficios, demasiados periódicos implica un despilfarro de recursos. Dado el número de periódicos vendidos en el pasado en cada tienda, ¿cuántos periódicos se deben entregar al día siguiente? Las cifras de ventas de cada uno de los lugares forman series de tiempo que son espacialmente correlacionadas a todos los demás puntos de venta. Echando un vistazo más de cerca en la variación y dependencias de periódicos en cifras. En primer lugar, la serie temporal se rige por una tendencia mundial. Si un titular interesante plantea la atención de la población, más periódicos se venden en todo el país. En segundo lugar, la variación espacial y temporal a mediana escala se plantea, por ejemplo debido a la meteorología o las vacaciones. En tercer lugar, las tendencias locales existen en la serie de tiempo que dependen del entorno inmediato del punto de venta. Un lugar dentro de una zona habitacional se distingue de la estación central, así como los antecedentes sociales de la zona juegan un papel importante. Considerando el tiempo, diferencias (periódicas) se llevan a cabo durante días laborables y los fines de semana.

Por último, la influencia de los demás lugares, que pueden suponer el mayor desafío de todos. Si en un lugar se vende todo, los clientes pueden obtener un periódico en un lugar cercano. Sin embargo, si en una tienda se vende todo con frecuencia, la gente adapta sus rutas y en consecuencia puede cambiar su comportamiento a largo plazo. También pueden comprar copias en su camino hacia el trabajo o en un punto central de venta y no en la vecindad inmediata. Estos pocos ejemplos muestran ya el complejo comportamiento dinámico de los objetos en movimiento que requiere ser capturado y predicho en series de tiempo espacio-temporales.

4.4.6 Clasificación de Trayectorias.

En el ámbito de la predicción, también los métodos de clasificación de (partes de) trayectorias en unos grupos ya definidos ofrecen un amplio rango de meta-datos que son derivados y se adjuntan a las trayectorias. Estos meta-datos pueden ser usados en la tarea de predicción. Imagina por ejemplo, la forma en que la ruta de un turista difiere de la ruta de un obrero local. Tal información puede ser utilizada por los servicios basados en localización para adaptar sus ofertas a la necesidad actual de un cliente potencial. Otra tarea es la clasificación para inferir los medios de transporte de una trayectoria que permite responder a preguntas como las siguientes: ¿Qué parte de un movimiento diario de la persona se puede atribuir a los vehículos particulares?, ¿Que calles ubicadas fuera del centro de la ciudad son en su mayoría utilizados por peatones?

A partir de nuestro estudio de la literatura, no hemos encontrado ningún método para clasificar trayectorias en grupos ya definidos. Sin embargo, existen problemas similares en el área de análisis de series temporales. (Keogh & Pazzani, 1998) utilizan de manera lineal la representación de series de tiempo y el peso cada uno de los segmentos en función de su

importancia. En (Geurts, 2001), las series de tiempo se clasifican por la aplicación de unas pautas como criterios de prueba en árboles de decisión. Cada patrón de lo que corresponde a un modelo temporal lleva asociado una constante de la señal, lo que puede, por ejemplo, representar la velocidad de un objeto. En general, las trayectorias se pueden clasificar usando al vecino más cercano utilizando algoritmos siempre que una adecuada función de la distancia sea dada. Sin embargo, la definición de una función de distancia depende de la tarea de clasificación, como se ha visto también en los métodos de agrupación. Además, no es fácil determinar la escala, el traslado y los efectos de ruido, elementos que han de tenerse en cuenta.

4.5 El papel de la incertidumbre en la Minería de Datos Espacio-Temporales.

La incertidumbre es una característica inherente de datos espacio-temporales. Surge debido a las limitaciones físicas y técnicas durante la recolección de datos y almacenamiento. Mientras que puede ser ampliamente asumido que el tiempo es entregado con gran precisión, la incertidumbre de la ubicación varía en función de la tecnología aplicada entre unos pocos metros (GPS) y kilómetros (GSM).

Además, la tasa de muestreo posee una gran influencia en la precisión. Entre más rápido se mueve un objeto, más frecuentemente la ubicación de un objeto debe ser reportada para sostener un nivel de dado de incertidumbre espacial. El conocimiento general del dominio, así como ciertos supuestos sobre el comportamiento del movimiento ayudan a reducir la incertidumbre en los datos. Por ejemplo, en cuanto el seguimiento de un vehículo, uno puede estar seguro de que todos los movimientos están restringidos a la red de calles; es poco probable que los coches puedan desplazarse a través de los edificios. Otro supuesto es el hecho de movimiento lineal entre dos posiciones, en general, tomando en cuenta dos posiciones P_1 y P_2 en momentos t_1 y t_2 y una velocidad máxima, la posición de un objeto en cada momento en el tiempo $t \in [t_1, t_2]$ se limita a algunas áreas, si no se da más información, una distribución uniforme de los objetos dentro de esta zona puede ser asumida.

5. Análisis de los estudios realizados.

En esta sección se van analizar los datos adquiridos con los estudios realizados en este trabajo de máster. En primer lugar se analizará el estado actual de los métodos de anonimización de trayectorias. Sobre este tema, se va a presentar un estudio comparativo de los métodos de anonimización presentados en la sección 3. A continuación, se van a presentar las conclusiones del estudio preliminar de los métodos de minería de datos aplicados a secuencias espacio-temporales. Finalmente, se presentan otras consideraciones interesantes que deben tenerse en cuenta si se quiere realizar un trabajo de investigación o aplicación en estos temas.

5.1 Análisis sobre los métodos de anonimización

En la esta sección se resumen brevemente los métodos de anonimización estudiados y a continuación se hace un análisis de ellos.

En el primer método, que hemos denominado “Confusión de Trayectorias” se estudian mecanismos para prevenir a un adversario rastrear completamente una trayectoria individual, genera una confusión de las trayectorias, cada vez que el trayecto de dos usuarios se encuentra, hay una posibilidad de que un atacante confunda los trayectos y pueda seguir al usuario equivocado, por lo que un algoritmo de privacidad explota esta posibilidad perturbando la información de la localización en algunas de estas áreas que se encuentran aumentando así las probabilidades de esta confusión,.

El siguiente trabajo analizado, basado en “Ofuscación de localizaciones”, presenta un método de anonimización adecuado para detección de proximidad y separación, ofusca los datos individuales de manera que no pueda ser identificado un individuo en cada uno de los datos del conjunto, la meta es proteger la identidad.

El método de “Modificación de proyecciones” considera un escenario donde las muestras de la localización son señaladas en un conjunto, la idea principal es transformar proyecciones largas y detalladas en proyecciones pequeñas y simples, con esto se es capaz de diversificar las ubicaciones que están siendo monitorizadas por los adversarios haciendo así imposible inferir en los datos con certeza si las trayectorias incluyen ciertos puntos, para lograrlo es necesario suprimir ciertos puntos en las trayectorias, logrando con esto aumentar la privacidad.

Finalmente, la “Anonimización por generalización” propone que la privacidad de las trayectorias puede ser obtenida aplicando la k-anonimidad, donde las trayectorias de distintas líneas de tiempo pueden ser anonimizadas al mismo tiempo, se remueve información de los datos marcando el uso de generalizaciones espacio-temporales, alineación de puntos en espacio y tiempo y la supresión de puntos y trayectorias, genera agrupaciones de trayectorias que después son anonimizadas.

La tabla siguiente muestra los métodos analizados y los relaciona con las características de los datos explicadas en el capítulo 2.

Método	Localización Independiente	Trayectorias Individuales	Espacio	Tiempo
Confusión de Trayectorias		X	Coordenadas (x,y)	Cada muestra corresponde a un instante de tiempo.
Ofuscación de Localizaciones	X		Coordenadas (x,y)	Cada muestra corresponde a un instante de tiempo.
Modificación de Proyecciones		X	Eventos o localizaciones concretos (datos simbólicos)	No toma en cuenta momentos de tiempo.
Anonimización por Generalización		X	Coordenadas (x,y)	Cada muestra de localización esta etiquetada con un sello del instante de tiempo en la que fue recolectada.

Tabla 2. Métodos de Anonimización vs Información Espacio-Temporal

En esta tabla primero se ha clasificado cada método según si se trata de un método de protección de las localizaciones de la trayectoria de forma independiente o si la protección se hace a nivel de trayectoria de forma integrada. Podemos ver que 3 de los métodos abordan el problema desde la perspectiva de anonimizar la secuencia de datos completa, y solo el método de **Ofuscación de Localizaciones** realiza una protección a nivel de cada localización.

También podemos observar que tres de los métodos al capturar el espacio utilizan coordenadas (x, y) mientras que el método de **Modificación de proyecciones** define eventos o localizaciones concretas utilizando datos simbólicos (por ejemplo, el nombre de los sitios visitados: tiendas, museos, ...). Por otra parte en los datos que refieren al tiempo este mismo método no toma en cuenta los momentos de tiempo, mientras que los otros tres métodos presentados si toman esto en cuenta, en los métodos **Confusión de trayectorias y Ofuscación de localizaciones** cada muestra corresponde a un instante de tiempo, y en el método de **Anonimización por generalización** cada muestra de localización esta etiquetada con un sello del instante de tiempo en la que fue recolectada donde las trayectorias de distintas líneas de tiempo pueden ser anonimizadas a la vez y su diferencia de tiempo cuenta en la métrica usada para la inexactitud de los datos.

La tabla siguiente muestra los métodos analizados y los relaciona con características del ámbito de la protección de la privacidad.

Método	Medida de la calidad del Resultado.(error)	Medida de la calidad de la Anonimización.
Confusión de Trayectorias	La inexactitud de los datos se mide de acuerdo a la calidad del servicio (QoS). (Ecuación 2)	La privacidad es medida a través de la "Expectativa del error de distancia". (Ecuación 1)
Ofuscación de Localizaciones		
Modificación de Proyecciones		El algoritmo de identificación de violaciones a la privacidad (Algoritmo 3) identifica las proyecciones que conducen a una violación de la privacidad.
Anonimización por Generalización	La pérdida de información es medida de acuerdo a la ecuación LCM (log cost metric)	

Tabla 3. Características de la protección de la privacidad.

En esta tabla se analizan las dos características más comunes en la protección de la privacidad, la medida de la calidad del resultado que nos permite saber que tanto ha sido perturbada la información y la medida de la calidad de la anonimización que nos permite conocer que tanto ha sido realmente anonimizada la información. En el método de **confusión de trayectorias** se mide la calidad de los resultados definiendo la calidad en el servicio (QoS) en términos del error que el algoritmo impone a las muestras de la localización, y la privacidad es medida a través de la "expectativa del error de la distancia" que captura que tan exacto un adversario puede igualar las localizaciones en las secuencias. Como se puede observar en la tabla 3, el método de **ofuscación de localizaciones** no aplica ninguna métrica para medir la calidad de los resultados o de la anonimización obtenida, mientras que en el método de **modificación de proyecciones** no se usa una métrica para la calidad de los resultados pero en lo que respecta a la medida de la calidad de la anonimización, mientras que el mecanismo usado para la generalización de las trayectorias, suprime la existencia de ciertos puntos en ellas, tomando en consideración el beneficio en términos de privacidad y utiliza un algoritmo que identifica las proyecciones que conducen a una violación de la privacidad.

Por último en el método de **Anonimización por generalización** la pérdida de información es medida de acuerdo a la ecuación LCM (log cost metric) que cuantifica la inexactitud de las trayectorias en ambos espacio y tiempo y se calcula sumando la ampliación requerida sobre cada uno de los extractos de localización publicados y en cuanto a la calidad de la anonimización tampoco aplica ningún método para cuantificarla.

De este estudio podemos definir un conjunto de cuestiones abiertas:

- ¿Cómo se puede incluir información geográfica sobre el sitio donde se realizan las trayectorias? En la sección 2 se ha argumentado que hay muchas características espaciales que influyen en las trayectorias, como altitud, pendiente, aspecto y otras características del terreno, accesibilidad en relación con diversas restricciones (obstáculos, la disponibilidad de caminos, etc) ,carácter y propiedades de la superficie: la tierra o el agua, el hormigón o el suelo, los bosques o campo, etc ,los objetos presentes en una ubicación: edificios, árboles, monumentos, etc ,la función o modo de uso, por ejemplo, la vivienda, las compras, la industria, la agricultura o el transporte.

- ¿Cómo incluir información demográfica de los individuos? Esto ayudaría a contextualizar la trayectoria y reconocer algunos patrones de movimiento en espacio y tiempo.
- ¿Cómo tratar otra información temporal sobre los datos? En los métodos presentados solo uno trata la cuestión temporal dentro del algoritmo, es decir, el instante de tiempo en que se tomaron los datos; los demás comparan las localizaciones por instante relativo de tiempo (primer dato, segundo dato, etc.). Sin embargo, ninguno de ellos incluye intervalos de tiempo, ni considera periodicidad en los datos, u otras de las características temporales mencionadas en la sección 2 durante el proceso de anonimización.
- ¿Cuál será la mejor manera de medir el grado de anonimización y disminución de calidad en los datos? En los artículos analizados se proponen solo dos formas de medir el nivel de protección de los datos y también el grado de error introducido. Sería interesante abordar este tema en más detalle y encontrar unas medidas más estándar para este tipo de datos, posiblemente en función de la utilidad que deban tener.

5.2 Análisis sobre los métodos de minería de datos

La Minería de datos espacio-temporales y, en particular, los datos sobre las trayectorias, son una gran zona que aún permanece casi inexplorada.

Los problemas presentados en la sección 4 se han organizado a lo largo de una taxonomía clásica de minería de datos, que incluyen el agrupamiento (o clustering), extracción de patrones y predicción (incluyendo clasificación).

Después de analizar los datos recopilados, queremos destacar las siguientes cuestiones o retos que aún están por resolver:

- ¿Qué nociones de similitud y distancia son las más adecuadas para una determinada agrupación de tareas basadas en distancia? En particular, las diferentes configuraciones pueden requerir diferentes niveles de rigor en la comparación de las trayectorias: de checar las coincidencias espaciales y temporales (las trayectorias son similares si visitan los mismos lugares en las mismas horas) a sólo una coincidencia espacial (el orden de visita puede ser importante, pero no precisa los tiempos), la similitud de mociones relativas (teniendo en cuenta la velocidad, dirección, etc.) o simplemente la similitud de características generales (velocidad media, la duración, etc.)
- ¿Cómo definir el mejor modelo de grupo? La compleja naturaleza de las trayectorias pueden dar lugar a modelos de grupos relacionados con la estructura interna de los datos secuenciales, tales como la circulación de información relativa a determinados sub-intervalos de tiempo o sub-regiones del espacio o, en otros casos, pueden requerir el desarrollo de modelos generales de la circulación general, tales como modelos probabilísticos.
- ¿Qué características modelan mejor los tipos de eventos o características de los que desea extraer patrones clásicos locales, como el establecimiento de puntos frecuentes,

patrones secuenciales de normas y de asociación? Una amplia gama de alternativas es posible, en principio, que van desde la simple información agregada (por ejemplo, la longitud de la trayectoria) para descripciones espaciales o espacio-temporales de los movimientos (por ejemplo, conjunto de lugares visitados, o maniobras como giros en U).

- ¿Qué nociones de patrones locales pueden encajar mejor en es estudio de trayectorias en un dominio concreto? Hasta la fecha, los enfoques disponibles se centran principalmente en el componente espacial o tratan de adaptar modelos de patrones locales clásicos a este otro tipo de datos más complejo.
- ¿Cuáles son los mejores métodos para la predicción de diferentes tipos de fenómenos, tales como futura posición del individuo, densidad o región futura, eventos de varios tipos, valores de variables asociadas a ubicaciones espaciales?
- ¿Qué características y métodos son los más adecuados para la clasificación de los objetos de las trayectorias que describen sus movimientos? Este es un gran campo por explorar, y hay al parecer, ninguna aproximación de ningún tipo específicamente enfocada en trayectorias de datos.
- ¿Cómo lidiar con la incertidumbre de la mejor manera? Además de los métodos de tratamiento previo, los algoritmos de minería de datos podrían ser guiados por sus conocimientos sobre la aproximación que afecta a los datos de entrada.
- ¿Cómo predecir en un entorno inestable? Normalmente, las predicciones asumen un entorno espacial fijo. Sin embargo, las carreteras pueden estar cerradas debido a trabajos de reconstrucción. Además, el comportamiento de desplazamiento de los individuos cambia con el tiempo. Un nuevo lugar de trabajo, apertura o cierre de locales comerciales o cambios en los medios de transporte naturalmente influyen en el comportamiento de los desplazamientos. Por esto es importante que los algoritmos puedan fácilmente incorporar cambios estructurales y adaptarse a nuevos patrones de conducta de movimientos.
- Por último, ¿cómo puede ser usado el conocimiento sobre el contexto del estudio y, en particular las ontologías especialmente diseñadas para el espacio geográfico, para extraer información más significativa y más útil? La integración espacio-temporal de la extracción de patrones y los métodos avanzados de representación del conocimiento es una línea de investigación muy interesante, pero está muy poco desarrollada.

5.3 Otras consideraciones

Hoy en día, las dos fuentes predominantes de trayectorias de datos para objetos en movimiento son las redes inalámbricas y el GPS. Por un lado, las empresas de telecomunicaciones acumulan masas de datos de movimiento basados en móviles. Por otro lado, las tecnologías como el GPS proveen una considerablemente más precisa información de la posición. Sin embargo, el intercambio para datos de alta calidad radica en una reducción sustancial en la cantidad de datos GPS ya que no estos no están disponibles fácilmente. Un reto en la tarea de investigación, por lo tanto, se refiere a la combinación de ambas fuentes de datos y la explotación de los efectos sinérgicos para aumentar la exactitud de predicción.

6. Realización del trabajo dentro del proyecto de investigación.

Este trabajo está realizado dentro de un proyecto inter-departamental entre la Universidad Rovira i Virgili y el Parque Científico y Tecnológico de Turismo y Ocio, en el que participan tres grupos: ITAKA (Intelligent Technologies for Advanced Knowledge Acquisition), CRISES (Criptografia i Secret Estadístic) y GRATET (Grup de Recerca d'Anàlisi Territorial i Estudis Turístics).

La realización de este trabajo de máster en el contexto de este proyecto me ha dado la oportunidad de integrarme no solo a un grupo de investigación, sino también a un equipo más amplio que trabaja en un proyecto, con objetivos, tareas y organización bien definidos.

A continuación se van a presentar algunos aspectos del funcionamiento del proyecto que han sido relevantes en la realización de este trabajo de máster:

- Se han realizado reuniones periódicas donde he presentado los resultados que se iban obteniendo de la investigación de nuestro grupo ITAKA.
- Para facilitar la colaboración e intercambio de información entre grupos se establecieron mecanismos para poder integrar toda la información así como poder compartir la información obtenida por los otros integrantes. Concretamente, se hizo uso de herramientas online como Refworks para poder compartir las fuentes investigadas y los trabajos revisados.
- Para facilitar el análisis de los artículos que cada grupo encuentra, se definieron unos conjuntos de variables interesantes para cada grupo, así como una lista de palabras clave para cada variable. Yo fui el encargado de realizar esta tarea en el grupo ITAKA. Estas plantillas se distribuyeron entre los 3 grupos y se consensuaron las variables, identificando un conjunto de variables comunes para los 3 grupos y otras específicas, de forma, que al leer un artículo se pudiera clasificar fácilmente usando las palabras clave, y detectar si contiene información sobre las variables relevantes para otro de los grupos.
- Se estableció también el formato del (los) documento(s) que se han de redactar durante el proyecto.

7. Conclusiones y contribuciones del trabajo.

En este trabajo de investigación se ha profundizado en el estudio de como explotar datos de trayectorias de individuos preservando su privacidad.

El estudio se ha enfocado en base a tres líneas principales: (1) definir todos los componentes de las secuencias espacio-temporales, (2) encontrar y analizar los métodos de anonimización de trayectorias y (3) revisar qué métodos de análisis inteligente de datos soportan este tipo de información de secuencias.

Después de este estudio se ha observado que existen extensiones de los métodos clásicos de análisis de los datos para secuencias espacio-temporales y hay un gran interés por la recopilación y estudio de estos datos, el problema reside cuando hablamos de la protección de la privacidad de estos datos, aunque es mucha la investigación que existe sobre la seguridad todavía se están desarrollando día a día investigaciones en este tema ya que la privacidad de los datos es un tema que aunque sea muy importante, la gente en general no está preocupada en este aspecto. Muchas veces hemos podido ser partícipes de servicios que atentan directamente con la privacidad de nuestros datos y no nos hemos dado cuenta, es sin embargo hasta que nos enfrentamos a un problema mayor, donde información personal es divulgada cuando nos damos cuenta que nuestra privacidad ha sido violada y ya es muy tarde para actuar, es por eso que los investigadores se están preocupando por este tema y se están generando métodos que pueden resolver estos problemas.

En particular, hemos constatado que existen pocos estudios de la protección de la privacidad para datos de trayectorias. Se han estudiado los métodos encontrados y se pueden observar distintos enfoques: algunos proponen publicar independientemente la localización de las muestras (Ghinita, 2009), proteger la privacidad de la localización a través de una confusión de las trayectorias (Hoh & Gruteser, 2005), ofuscar las localizaciones (Ruppel et al, 2006), modificar las proyecciones (Terrovitis & Mamoulis, 2008) y anonimización por agrupamientos de trayectorias (Nergiz et al, 2009). Sin embargo este es un área que recién se ha iniciado y por tanto se espera que haya más métodos que aborden esta cuestión y que pueden presentar alternativas ya sean similares o innovadoras o derivaciones de estas mismas.

Respecto a los métodos de análisis de datos podemos concluir que los diferentes paradigmas de minería de datos (agregación, clasificación, predicción) han abordado el problema de los datos secuenciales, y por tanto, existen algunas aproximaciones, aunque bastante sencillas, que permiten extraer conocimiento a partir de trayectorias de individuos. Sin embargo, quedan aún muchas cuestiones importantes abiertas que permitirían mejorar los resultados.

En resumen, las contribuciones principales de esta tesis de máster son las siguientes:

- Se analiza una prácticamente nueva línea de investigación en lo que respecta a la privacidad de la información de usuarios en trayectorias y se generan una serie de preguntas abiertas que pueden servir como base para futuras líneas de investigación.
- Se ha identificado a algunos grupos que están realizando investigación en esta misma línea lo que puede significar una posible cooperación en el futuro para proyectos venideros.
- Se hace una comparación y un análisis de los métodos encontrados lo que permite definir la forma en que estos mismos trabajan con la información.

- Se identifican componentes importantes en el estudio de secuencias de datos, lo que permite desarrollar análisis más detallados de información de secuencias.

Algo que queda muy claro es que el estudio de la movilidad de las personas es un área de investigación que tiene mucho campo sobre todo en lo que respecta a la seguridad de la información (anonimidad de los datos). Sin duda alguna el sector turístico se podría beneficiar de estas herramientas, desarrollando servicios y aplicaciones avanzadas para los turistas basándose en su ubicación que permitan conservar la anonimidad de los usuarios y la protección de su información.

Bibliografía.

(AGNU, 1948) Asamblea General de la Naciones Unidas (1948). *Declaración Universal de los Derechos Humanos*. Artículo 12

(Agrawal et al, 1995) Agrawal R., Lin K., Sawhney H., Shim K., *Fast similarity search in the presence of noise, scaling and translation in time series databases*. In Proceedings of 21st International Conference on Very Large Data Bases (VLDB'95), pp. 490-501. Morgan Kaufmann, Los Altos, California, 1995.

(Benetis et al, 2006) Benetis R., Jensen C., Karciauskas G. and Saltenis S. *Nearest and reverse nearest neighbor queries for moving objects*. The Very Large Database Journal, 15(3): 229-249, 2006.

(Brown et al, 2001) Brown D., Liu H., Xue Y., *Mining preferences from spatial-temporal data*. In Proceedings of the 1st International Conference on Data Mining (SDM'01), 2001.

(Cao et al, 2005) Cao H., Mamoulis N. and Cheung D.W., *Mining Frequent spatio-temporal sequential patterns*. In Proceedings of the 5th international Conference on Data Mining (ICDM'05), pp. 82-89 IEEE, New Orleans, LA, 2005.

(Castañeda et al, 2006) Castañeda Hernán, Gómez Juan, Leal Alexander, *Proveedor de Servicios Basados en Localización para Dispositivos Móviles*. Universidad Santo Tomas, Medellín Colombia, GICOMI Grupo de Investigación en Comunicaciones Inalámbricas.

(Gaffney & Smith, 1999) Gaffney S., Smyth P., *Trajectory clustering with mixture of regression models*, In Proceedings of the 5th International conference of Knowledge Discovery and Data Mining (KDD'99), pp. 63-72. ACM, New York, 1999.

(Geurts, 2001) Geurts P., *Pattern Extraction for time series classification*. In Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'01), pp. 115-127, Springer, Berlin Heidelberg New York, 2001.

(Ghinita , 2009) Ghinita Gabriel (2009), *Private Queries and Trajectory Anonymization: a Dual Perspective on Location Privacy*. Transactions on Data Privacy Vol.2, No. 1 3-19.

(Giannoti & Pedreschi, 2008) Giannoti F. and Pedreschi D. *Mobility, Data Mining and Privacy*. Pp. 270- 295 Springer, Berlin Heidelberg, 2008.

(Guttman, 1984) Guttman, R. (1984). *R-trees: A dynamic index structure for spatial searching*. Proceedings of the International Conference ACM SIGMOD.

(Hadjielefteriou et al, 2003) Hadjielefteriou M., Kollios G., Gunopulos D., and Tsotras V.J., *On-line discovery of dense areas in spatio-temporal databases*. In Proceedings of the 31st International Symposium on Advances in Spatial and Temporal Databases (SSTD'03), pp. 306-324. Springer, Berlin Heidelberg New York, 2003.

(Hadjieleftheriou et al, 2005) Hadjieleftheriou M., Kollios G., Bakalov P. and Tsotras V.J., *Complex spatio-temporal pattern queries*. In Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05), pp. 877-888. ACM, New York, 2005.

(Hagerstrand T., 1970) Hagerstrand T. *What about people in regional science?* Papers of the Regional Science Association, 24:7–21, 1970.

(Han&Kamber, 2001) Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

(Hawking et al, 2005) Hawking P., Stein A., Zeleznikow J., Sharma P., Nugent D., Dawson L. and Foster S., *Emerging Issues in Location Based Tourism Systems*, Proceedings of the International Conference on Mobile Business (ICMB'05). IEEE, 2005.

(Hernández et al., 2004) Hernández, J., Ramírez, M. J., & Ferri, C. (2004). *Introducción a la Minería de Datos*. Pearson, Prentice Hall.

(Hoh&Gruteser,2005) Hoh Baik & Gruteser Marco (2005), *Protecting Location Privacy Trough Path Confusion*, Proceedings of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM 2005).

(Hwang et al, 2005) Hwang S.Y., Liu Y.H., Chiu J.K. and Lim E.P., *Mining mobile group patterns: A trajectory based approach*. In Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05), pp. 713-718. Springer, Berlin Heidelberg, New York, 2005.

(Kalnis et al, 2005) Kalnis P., Mamoulis N. and Bakiras S., *Discovering moving clusters in spatio-temporal data*. In Proceedings of 9th International Symposium on spatial and temporal Databases (SSTD'05), pp. 364-381. Springer, Berlin Heidelberg New York, 2005.

(Karimi & Liu, 2003) Karimi H. and Liu X., *A predictive location model for location based services*. In Proceedings of the 11th International Symposium on Geographic Information Systems (GIS'03), pp. 126-133. ACM, New York, 2003.

(Keogh & Pazzani, 1998) Keogh E., and Pazzani M., *An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback*. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98), pp. 239-241. ACM, New York, 1998

(Kulldorff, 1997) Kulldorff M., *A spatial scan statistic*. Communications in statistics: Theory and Methods, 26(6), 1481-1496, 1997.

(Laasonen, 2005) Laasonen K., *Clustering and prediction of mobile user routes from cellular data*. In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), pp. 569-576. Springer, Berlin Heidelberg New York, 2005.

(Lacramioara et al., 2007) Lacramioara L., Müller M., Retz P., Musso V., Carebear E. (2007). *Las nuevas tecnologías en el turismo*. Primer encuentro internacional sobre Turismo y Desarrollo, Universidad de Málaga.

(Laube & Imfeld, 2002) Laube P. and Imfeld S., *Analyzing relative motion within groups of trackable moving point objects*. In Proceedings of 2nd International Conference on Geographic Information Science (GIS'02), pp. 132-144, Springer, Berlin Heidelberg New York, 2002.

(Lee et al, 2007) Lee J.G., Han J. and Whang K.Y., *Trajectory Clustering: A partition and group framework*. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD'07), pp. 593-604. ACM, New York, 2007.

(Li et al, 2004) Li Y., Han J. and Yang J., *Clustering moving objects*. In Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD'04), pp. 617-622. ACM, New York, 2004.

(Liao, 2005) Liao T.W., *Clustering of time series data. A survey* *Pattern Recognition*, 38(11), pp. 1857-1874 (2005)

(Miller H., 2005) Miller H., *A measurement theory for time geography*. *Geographical Analysis*, 37:17–45, 2005.

(Nanni ,2002) Nani M., *Clustering Methods for Spatio-Temporal data*, PHD Thesis, Computer Science Department, University of Pisa 2002.

(Nergiz et al., 2009) Nergiz Mehmet., Atzori Maurizio, Saygin Yucel (2009) *Towards Trajectory Anonymization: a Generalization-Based Approach*. *Transactions on Data Privacy* Vol.2, No.1. 47-75.

(Pérez, 2006) Pérez, C. (2006). *Data Mining*. Ra-Ma.

(RAE ,2005) Real Academia Española (2005). *Diccionario de la Lengua Española*. 23ª edición.

(Reid D., 1979) Reid D. (1979). *An algorithm for tracking multiple targets* .IEEE Transactions on Automatic Control, 24(6):843-854.

(Ruppel et al., 2006) Ruppel Peter, Treu George, Küpper Axel, Linnhoff Claudia (2006), *Anonymous User Tracking for Location Based Community Services*. LoCA 2006, LNCS 3987,116-133, Springer-Verlag Berlin.

(Schmidt-Belz et al., 2003) Schmidt-Belz B., Laamanen H., Poslad S., Zipf A., *Location-based mobile tourist services - first user experiences*. Information and communication technologies in tourism 2003: Proceedings of the International Conference in Helsinki, Finland, 2003.

(Tao et al, 2003) Tao Y., Papadias D., Sun J., *The TPR* tree: An optimized spatio-temporal access method for predictive queries*. In Proceedings of the 29th International Conference on Very Large Data Bases (VLDB'03), pp. 790-801. Morgan Kaufmann, Los Altos, CA, 2003.

(Tchetchik et al., 2009) Tchetchik A., Fleischer A., and Shoval N., *Segmentation of Visitors to a Heritage Site Using High-resolution Time-space Data* *Journal of Travel Research* 2009, doi:10.1177/0047287509332307.

(Terrovitis&Mamoulis,2008) Terrovitis Manolis, Mamoulis Nikos (2008), *Privacy Preservation in the Publication of trajectories*. The Ninth International Conference on Mobile Data Management.

(Vlachos et al, 2002) Vlachos M., Kollios G., Gunopulos D., *Discovering Similar Multidimensional trajectories*. In Proceedings of the 18th International Conference on Data Engineering (ICDE'02), pp. 673-684. IEEE, San Jose, California, 2002.

(Wang et al, 2003) Wang Y., Lim S., and Hwang S., *Mining group patterns of mobile users* .In Proceedings of the 14th International Conference on Database and Expert Systems Applications (DEXA'03), pp. 287-296, Springer, Berlin Heidelberg New York ,2003.