Universitat Oberta de Catalunya

uoc.edu

Toolkit for privacy evaluation of geolocated data

Manel Gil García

Máster Universitario en Seguridad de las Tecnologías de la Información y de las Comunicaciones

Privacidad de datos

Tutor: Julián Salas Piñón

Profesor responsable de la asignatura: Víctor Garcia Font

Diciembre de 2018

Resumen

El auge de los sistemas de geolocalización de los últimos años ha venido acompañado de una nueva problemática, la privacidad de los datos de los usuarios que usan estos sistemas. Este Trabajo de Fin de Máster (TFM) se propone evaluar la privacidad de los datos obtenidos al sanitizar un dataset de puntos espacio-temporales mediante una serie de métodos de sanitización. Para evaluar la privacidad de un dataset geolocalizado, se evalúa cuán difícil para un atacante sería deducir información sensible para un sujeto cuya información se encuentre en el dataset, ya sea mediante la deducción de puntos de interés (POIs) o mediante la reidentificación de un sujeto mediante el linkado de trazas. Se han evaluado la perturbación, la agregación y el swapping o intercambio como métodos de sanitización. Con tal de evaluarlos, se han atacado los datasets sanitizados mediante los ataques de deducción de hogares, de extracción de stays y de extracción de localizaciones inicio-fin (begin-end). Los resultados deben ser tenidos en cuenta desde el punto de vista de la aplicación que se le dará a los datos sanitizados. Si se desea mantener todos los datos agregados intactos y usar el dataset sanitizado para los mismos propósitos que el dataset original, el método swapping es el más potente manteniendo un buen nivel de privacidad. Si la utilidad del dataset sanitizado no tiene en cuenta que los datos extraídos mantengan los datos agregados ni que se pueda usar para, por ejemplo, crear mapas de movilidad -un escenario poco realista-, la perturbación es un método muy sencillo de implementar que sería suficiente, aunque si no se combinan distintos tipos de perturbación, un atacante podría fácilmente deducir las localizaciones originales.

Palabras clave: Privacidad, geolocalización, sanitización, ataques de deducción de POIs

Abstract

The rise of Location Based Systems during the last years has lead to a new problem, the data privacy of system's users. This Master's Final Project (TFM) aims to evaluate the privacy of the data obtained by sanitizing a dataset of spatiotemporal points through a series of sanitization methods. In order to evaluate the privacy of a geolocated dataset, we try to know how difficult it would be for an attacker to deduce sensitive information for a subject whose information is in the dataset, either by the deduction of points of interest (POIs) or by re-identifying a subject by linking traces. Perturbation, aggregation and swapping of traces have been evaluated as sanitization methods. In order to evaluate them, the sanitized datasets have been attacked by the following methods: home inferring, extraction of stays and begin-end location finder. The results must be considered from the point of view of the application that will be given to the sanitized data. If you want to keep all the aggregated data intact and use the sanitized dataset for the same purpose as the source one, swapping is the most powerful method, maintaining a good level of privacy. If the utility of the sanitized dataset does not take into account that the extracted data keeps aggregated data or that it can be used to, for example, create mobility maps -an unrealistic scenario-, perturbation is a very simple method to implement, although if different types of disturbance are not combined, an attacker could easily deduce the original locations.

Keywords: Privacy, geolocation, sanitization, POI inference attacks

Índice

Toolkit for privacy evaluation of geolocated data

1 Introducción

En los últimos años, el uso generalizado de dispositivos móviles que incluyen sensores de geoposicionamiento ha permitido la expansión de sistemas basado en la localización (LBS, Location Based Services). Este avance en los sistemas de localización ha permitido, entre otras cosas, que los usuarios puedan visualizar información en tiempo real acerca de tráfico o transporte público, recomendaciones basadas en la posición, cálculo de rutas en función de la localización y un largo etcétera. Con esta expansión, también se ha introducido una nueva problemática, que no es otra que el riesgo para la privacidad de los usuarios que puede representar la revelación de sus datos de ubicación geográfica. Una de las mayores amenazas para la privacidad de un usuario es que un atacante pueda deducir el lugar en el que se encuentra. En caso de que un atacante pudiese relacionar a un individuo con sus datos de localización, podría deducir los puntos de interés (POIs) del individuo -como podrían ser la localización de su residencia o de su trabajo-, descubrir sus hábitos e incluso sus preferencias personales.

Para tratar de evitar estos peligros en la medida de lo posible, varios autores han desarrollado diversas técnicas de sanitización de datos que abordan el problema desde distintas perspectivas. Durante este trabajo se han implementado varios métodos de sanitización que pretenden que los datos de geolocalización sanitizados obtenidos puedan ser utilizados para el mismo propósito que los datos del dataset original, protegiendo en mayor o menor medida la privacidad de los sujetos del dataset. Por otro lado, en este trabajo también se han implementado una serie de ataques de inferencia o deducción de puntos de interés de los individuos de un dataset geolocalizado. Con tal de poder evaluar los datos sanitizados por los distintos métodos de sanitización implementados, se ha escogido un dataset con datos de geolocalización reales para el que se ha calculado la unicidad de sus trazas y se han deducido puntos de interés de sus sujetos mediante los ataques implementados. Los resultados obtenidos para este dataset original han servido como base para evaluar los distintos métodos de sanitización. Se han aplicado los distintos métodos de sanitización implementados al dataset escogido, obteniendo sendos datasets sanitizados. Contra cada uno de los datasets sanitizados obtenidos, se aplican todos los ataques que se han implementado. Los resultados de los ataques contra cada dataset sanitizado, comparándolos entre ellos y con los resultados sobre el dataset original, nos permiten evaluar el nivel de privacidad o la dificultad añadida para un atacante para conseguir información sensible acerca de los sujetos del dataset sin sanitizar. Además, también se ha calculado la unicidad de las trazas para cada uno de los datasets sanitizados, con tal de comparar la privacidad de los distintos datasets.

2 Objetivos

A continuación, se enumeran brevemente los objetivos que se han tratado de alcanzar mediante la realización de este Trabajo de Fin de Máster.

1. Medir el riesgo de revelación de información relativa a los sujetos que aportan datos de localización geográfica a un sistema geolocalizado.
2. Creación de visualizaciones de datos de localización.
3. Implementación de métodos de sanitización de datos.
4. Implementación de ataques de deducción de puntos de interés sobre los sujetos de un sistema de geolocalización.
5. Sanitizar un dataset real usando los distintos métodos implementados y atacar los datasets sanitizados con los métodos de ataque implementados.
6. Evaluación, mediante las medidas de riesgo y visualizaciones creadas, de la efectividad de los métodos de sanitización de datos.

3 Datos de geolocalización

Los datos que localizan geográficamente a todos los sujetos de un sistema de geolocalización determinado forman el conjunto de datos o dataset D del sistema. En este trabajo, solamente se tendrán en cuenta los datos de los puntos espacio-temporales de cada sujeto, aunque en los sistemas reales se usa mucha información adicional aparte de la estrictamente espacio-temporal. Estos datos a los que hacemos referencia son conocidos como trazas de movilidad, o simplemente, trazas. [1]
En general, para cualquier sistema, y también en el caso de este trabajo, una traza está formada por la siguiente información:

- Un identificador o id del individuo. Éste puede ser el identificador real del sujeto o de su dispositivo, o puede tratarse de un pseudónimo que anonimice al usuario y no permita relacionar sin ninguna dificultad los hábitos y localizaciones del usuario. En el dataset utilizado para la realización de este trabajo se usan pseudónimos para identificar a los sujetos.
- Las coordenadas de la posición geográfica de la traza. En el caso de este trabajo, la posición de los individuos se identifica mediante la latitud y la longitud en grados decimales del punto en el que se encuentre. Esto se denota durante el trabajo como (latitud, longitud), ya sea usando º para indicar los grados o sin él.
- Un timestamp o fecha con hora. Puede tratarse de un intervalo de un momento exacto. Para este trabajo, se trata de fechas con horas exactas.
- Información adicional. En el dataset usado en este trabajo, existe para cada traza información adicional que indica si el taxi está libre u ocupado.

4 Métodos de sanitización

A continuación, se describen los métodos de sanitización que se han implementado y evaluado durante este trabajo. En todos los casos, disponemos del mismo dataset original D y obtenemos un dataset sanitizado D_s con el mismo número de trazas que el dataset D mediante la aplicación de un método de sanitización determinado.

En primer lugar, vemos los métodos que se han implementado:

- Perturbación
  - Traslación
  - Escala
  - Rotación
- Agregación
- Swapping

4.1 Perturbación

Este tipo de método de sanitización consiste en modificar las coordenadas de una traza añadiéndole algún tipo de perturbación arbitraria. Tal como hemos visto, en este trabajo se han implementado 3 tipos distintos de perturbación: la traslación, la escala y la rotación. [2][3]

4.1.1 Traslación

Esta transformación consiste en desplazar un punto en su latitud o longitud.
Un punto (X, Y) se trasladará a una nueva posición (X', Y') al aplicarle Dx y Dy.
X' = Dx + X
Y' = Dy + Y

En forma matricial:
$P' = \begin{bmatrix} X' \\ Y' \end{bmatrix}$
$T = \begin{bmatrix} D_x \\ D_y \end{bmatrix}$
$P = \begin{bmatrix} X \\ Y \end{bmatrix}$
$P' = T + P = \begin{bmatrix} D_x \\ D_y \end{bmatrix} + \begin{bmatrix} X \\ Y \end{bmatrix}$

4.1.2 Escala

Esta transformación consiste en escalar, es decir, acercar o alejar un punto, respecto al origen de coordenadas. En una aplicación geográfica como la de este trabajo, el origen de coordenadas corresponde al punto en el que se cruzan el Ecuador (paralelo en latitud 0º) y el meridiano de Greenwich (longitud 0º), que corresponde por tanto a la posición (0,0).
Un punto (X, Y) se escalará a una nueva posición (X', Y') al aplicarle Sx y Sy.
X' = Sx X
Y' = Sy Y

En forma matricial:

$$P' = \begin{bmatrix} X' \\ Y' \end{bmatrix}$$

$$S = \begin{bmatrix} S_x & 0 \\ 0 & S_y \end{bmatrix}$$

$$P = \begin{bmatrix} X \\ Y \end{bmatrix}$$

$$P' = S\,P = \begin{bmatrix} S_x & 0 \\ 0 & S_y \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$$

Al aplicar una perturbación de escala a un punto geográfico, existen 3 posibilidades:
- Escala > 1 (Sx>1 | Sy>1): El punto se aleja del origen de coordenadas.
- Escala = 1 (Sx=1 | Sy=1): El punto se mantiene inalterado en la misma posición.
- Escala < 1 (Sx<1 | Sy<1): El punto se acerca respecto al origen de coordenadas.

La escala será uniforme cuando Sx=Sy.

4.1.3 Rotación

Esta transformación consiste en rotar un punto respecto al origen de coordenadas en un ángulo concreto.
Un punto (X, Y) rotará un ángulo Θ respecto al origen de coordenadas a una nueva posición (X', Y').
X' = X cos(Θ) - Y sin(Θ)
Y' = X sin(Θ) + Y cos(Θ)
En forma matricial:

$$P' = \begin{bmatrix} X' \\ Y' \end{bmatrix}$$

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$P = \begin{bmatrix} X \\ Y \end{bmatrix}$$

$$P' = R\,P = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$$

Al igual que sucede con la escala, esta transformación se aplica respecto al origen de coordenadas, que en nuestra aplicación corresponde con el punto (0, 0).

4.1.4 Uso de coordenadas homogéneas

Tal como hemos podido ver hasta ahora, la traslación de un punto se obtiene mediante la adición, mientras que tanto la escala como la rotación se obtienen mediante la multiplicación. Con tal de poder integrar la traslación con la escala y la rotación, usaremos un sistema de coordenadas homogéneas. Las coordenadas homogéneas nos permiten tratar todas las transformaciones correspondientes a la perturbación como multiplicaciones de matrices.
En el caso de una representación 2D, como en este caso, esta solución implica añadir una tercera coordenada a cada punto, de manera que P pasará a ser P = (X,Y,W), lo que nos permite manejar las traslaciones mediante multiplicaciones, de igual manera que con la escala y la rotación. Como veremos a continuación, para nuestra aplicación de las coordenadas homogéneas, para las coordinadas X,Y conocidas, W tomará el valor 1, de manera que, en la práctica, P = (X,Y,1). [4]
A continuación, vemos cómo se realizan las transformaciones definidas previamente usando un sistema de coordenadas homogéneas.

4.1.4.1 Traslación

Para trasladar un punto P en una distancia Dx, Dy tenemos que:

$$P' = \begin{bmatrix} X' \\ Y' \\ 1 \end{bmatrix}$$

$$T = \begin{bmatrix} 1 & 0 & D_x \\ 0 & 1 & D_y \\ 0 & 0 & 1 \end{bmatrix}$$

$$P = \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

$$P' = T\,P = \begin{bmatrix} 1 & 0 & D_x \\ 0 & 1 & D_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

Se puede ver que de esta forma podemos trasladar un punto mediante la multiplicación de matrices.

4.1.4.2 Escala

Para escalar un punto P aplicando Sx, Sy tenemos que:

$$P' = \begin{bmatrix} X' \\ Y' \\ 1 \end{bmatrix}$$

$$S = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P = \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

$$P' = S\,P = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

4.1.4.3 Rotación

Para rotar un punto P aplicando un ángulo ϴtenemos que:

$$P' = \begin{bmatrix} X' \\ Y' \\ 1 \end{bmatrix}$$

$$S = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P = \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

$$P' = R\,P = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

4.1.5 Composición de transformaciones

Con tal de poder utilizar de forma efectiva la perturbación, se hace necesario combinar distintos tipos de perturbación para poder obtener transformaciones razonables del dataset.
La primera razón que hace necesaria la composición de transformaciones, también en el caso de este trabajo, es la necesidad de rotar o escalar respecto a un punto arbitrario en lugar de sobre el

origen de coordenadas. Para poder aplicar la rotación o la escala sobre un punto arbitrario se debe:

- Trasladar el o los puntos a transformar de manera que el punto arbitrario pase a ser el origen de coordenadas.
- Aplicar la rotación o la escala.
- Trasladar el o los puntos transformados de manera que el punto arbitrario vuelva a situarse en su posición original.

Para aplicar estas acciones, se podría aplicar cada transformación de manera individual a los puntos. Sin embargo, para combinar distintas transformaciones, lo que haremos es multiplicar las matrices de transformación de las distintas transformaciones que se quieran aplicar a un punto para obtener una única matriz de transformación, que se multiplicará por el punto al que se quieran aplicar las perturbaciones.

Así, conseguimos obtener el punto transformado P' mediante solamente la multiplicación de una matriz de transformación M_t por el punto P que se quiere transformar:

$P' = M_tP$

Esta aproximación nos permite obtener la posición de P' realizando muchas menos operaciones que si se aplicara cada transformación individualmente a todos los puntos.

Es importante tener en cuenta que el producto de matrices no es conmutativo, por lo que es relevante el orden en el que se multipliquen las matrices. Las matrices de transformación individuales de cada transformación que se quiere combinar se tienen que multiplicar en orden inverso al orden en el que se quieren aplicar las transformaciones.

A continuación, se muestra un caso que sirve de ejemplo para ilustrar esta situación.

En primer lugar, se quieren trasladar los puntos para que un punto arbitrario se convierta en el origen de coordenadas, para lo que tenemos la matriz T(-Dx,-Dy).

A continuación, se quieren rotar los puntos en un ángulo Θrespecto al origen de coordenadas, para lo que disponemos de la matriz R(Θ).

Por último, se quieren trasladar todos los puntos de manera que el punto arbitrario respecto al que se ha realizado la primera traslación, se volvería a situar en la posición original. Para ello, disponemos de la matriz T(Dx,Dy).

Así pues, dado que las matrices de transformación de cada una de las transformaciones que se quieren aplicar se tienen que multiplicar en orden inverso al orden en el que se desean aplicar, tenemos que

M_t = T(Dx,Dy) R(Θ) T(-Dx,-Dy)

y como hemos visto antes

$P' = M_tP$, por lo que

P' = [T(Dx,Dy) R(Θ) T(-Dx,-Dy)] P

4.2 Agregación

Este método de sanitización consiste en mostrar la información geográfica de las trazas agregada por zonas. La agregación se puede llevar a cabo de dos maneras.

En primer lugar, si se agrupa la información por zonas y cada polígono que representa una zona se identifica mediante un identificador, que sustituirá a las coordenadas de aquellas trazas que se encuentren en esa zona, estaríamos hablando de una agregación por área.

En este trabajo no se ha implementado una agregación por área. En su lugar, se ha implementado una segunda aproximación de la agregación, llamada agregación de puntos. A grandes trazas, la agregación de puntos consiste en representar un conjunto de localizaciones mediante el uso de un solo punto geográfico, es decir, agregar todas las localizaciones originales en una sola localización.

En concreto, en este trabajo se ha implementado la micro-agregación, y se ha aplicado este método a las coordenadas geográficas del dataset original D. Tal como se define en [5], la micro-agregación de puntos consiste en aplicar los siguientes dos pasos para cada traza del dataset: (i) identificar un set de registros o trazas que son geográficamente similares al registro o traza, y (ii) reemplazar las coordenadas del registro con unas coordenadas calculadas a partir del set de registros geográficamente similares.
Para la realización de este trabajo, se ha implementado una micro-agregación de puntos en que se aplican los siguientes pasos sobre cada traza del dataset: (i) seleccionar todos aquellos registros cuya latitud y longitud coincidan en precisión 0,001º con la traza, y (ii) reemplazar las coordenadas de la traza por la media (average) de la latitud, y de manera independiente de la longitud, de todos los registros similares identificados previamente en (i).
Aplicando la Micro-agregación a todas las trazas de D, obtenemos un dataset sanitizado D_s que, a diferencia de si se hubiera aplicado una agregación por áreas, podemos explotar de la misma manera que se pueden explotar los datos del dataset original D. En el caso de usar agregación por áreas, en lugar de coordenadas tendríamos identificadores de área o de polígonos (que definen las áreas), de manera que no se podrían explotar sus datos como en D.

4.3 Swapping

Este método de sanitización, descrito en [6], se basa en el intercambio de trayectorias parciales. La idea principal de este algoritmo consiste en que, cuando dos sujetos del dataset se encuentran cerca, se intercambian sus trayectorias parciales. Claramente, se hace necesario definir qué significa que dos sujetos estén *cerca*, lo cual se consigue con dos umbrales: un umbral de tiempo τ y un umbral de distancia χ Hay que tener en cuenta la frecuencia de muestreo, así como el hecho de que dos sujetos no pueden estar en el mismo punto exacto a la vez. Si una traza de la trayectoria T de un sujeto, siendo T el conjunto de trazas ordenadas para ese sujeto, se encuentra dentro del umbral de distancia χcon la traza de otro sujeto y esto sucede con una diferencia menor o igual al umbral de tiempo τ, se considera que ambos individuos se han *econtrado* en la fecha de sus trazas, denotado como i≈j, y son candidatos para aplicar el swapping.
En la siguiente figura se define el algoritmo usado para implementar este método de sanitización, que está basado en el algoritmo SwapMob definido en [6], aunque se ha escrito de nuevo el pseudocódigo en lugar de usar el pseudocódigo original. Se han introducido ligeras diferencias respecto al algoritmo SwapMob original para reflejar más fielmente el algoritmo usado en este trabajo.

Algoritmo de swapping

Input Dataset D, umbrales τ, χ
Output Dataset sanitizado D_s
$\cup \tau_j$ ←Dividir D en intervalos de tiempo τ en función de la fecha de las trazas
para cada par de registros i,j en el intervalo τ_j **hacer**
  si distancia(i(lat, long), j(lat, long)) $< \chi$**hacer**
    añadir i, j a la lista de posibles swaps (intercambios) $S_{\tau j}$ para el intervalo τ_j
    hacer match aleatorio entre los posibles swaps de $S_{\tau j}$
  fin si
fin para
ordenar todos los swaps de $\cup S_{\tau j}$ por fecha ascendente

para cada pareja i≈j en $\cup S_{\tau j}$ **hacer**

swap de las trayectorias parciales de *i* hasta la fecha del swap (T_i) con las trayectorias parciales de *j* hasta la fecha del swap (T_j)

fin para

retorna Dataset sanitizado D_s

Al ejecutar este algoritmo, se obtiene para cada conjunto de trazas de intervalo de tiempo τ, el conjunto de sujetos que se han encontrado durante el intervalo. Una vez disponemos del conjunto de sujetos candidatos al swapping del intervalo ($S_{\tau j}$), se realiza un matching aleatorio entre ellos. Este matching se realiza por parejas de la siguiente manera: se divide el conjunto de posibles swaps $S_{\tau j}$ en 2, se ordena cada uno de los dos subconjuntos de manera aleatoria e independiente, y, por último, se hace el match entre los dos subconjuntos para aquellos sujetos con el mismo orden en cada uno de los subconjuntos. Queda claro que, si el conjunto $S_{\tau j}$ es impar, uno de los sujetos se quedará sin tener match.

Por último, una vez disponemos del conjunto definitivo de swaps $S_{\tau j}$, se realiza el swap entre las trayectorias parciales de *i*,*j*.

5 Unicidad de las trazas de un conjunto

La unicidad, tal como se describe en [7] permite estimar el número de puntos p que son necesarios para identificar de forma única la traza de movilidad de un individuo. Cuantos menos puntos se necesiten, más únicas son las trazas, y por tanto, más fácil será re-identificarlas usando información externa.

Dado I_p, un conjunto de puntos espacio-temporales, y D, un dataset de trazas de movilidad, se evalúa la unicidad ε de las trazas extrayendo del dataset D un subconjunto $S(I_p)$ de trazas que coincidan con los p puntos de I_p. Se considera que una traza es única si $|S(I_p)|=1$, es decir, si el subconjunto $S(I_p)$ contiene solamente una traza.

Es importante tener en cuenta que ε depende de la resolución tanto temporal como espacial del dataset D.

Para este trabajo, se han agrupado las localizaciones de las trazas en celdas cuadradas de 111,32 metros -o lo que es lo mismo, $0{,}001^o$ decimales-. En cuanto a la resolución temporal, puesto que no es posible que en el mismo instante de tiempo dos objetos se encuentren en el mismo punto exacto, se ha tenido que usar un umbral de tiempo τ respecto a la fecha de la traza.

Con tal de evaluar la unicidad de las trazas del dataset D usado en este trabajo, se ha optado por una solución basada en la fuerza bruta. Esta solución consiste en, para las trazas de cada uno de los individuos de D, seleccionar al azar p puntos espacio temporales para formar el conjunto I_p, y buscar el subconjunto $S(I_p)$ de individuos del dataset D cuyas trazas son compatibles con p, es decir, cuyas trazas contienen todos los puntos p escogidos al azar. Los puntos p que forman I_p se seleccionan de forma aleatoria entre todas las trazas del individuo en D, sin aplicar ninguna restricción.

Aunque la unicidad de las trazas se puede estimar como el porcentaje de 2500 trazas aleatorias que son únicas dado p, en este trabajo se ha evaluado para cada individuo en D el conjunto I_p, buscando el subconjunto $S(I_p)$ en cada una de las trazas de todo el dataset D, que está compuesto de más de 11 millones de trazas.

6 Ataques de inferencia de puntos de interés

Existen múltiples métodos para deducir o inferir los puntos de interés de un individuo descritos en la literatura actual, algunos de los cuales se han implementado en este trabajo.
Los métodos de deducción de POIs implementados nos permitirán evaluar la efectividad de los distintos métodos de sanitización descritos previamente en este documento y que se han implementado para este trabajo.
A continuación, se enumeran los métodos de deducción implementados:
- Deducción de hogares por mayor número de repeticiones. [6]
- Deducción de "stays" usando el algoritmo DT-Cluster. [1][8]
- Heurística. Método Begin-End location finder. [1]

Ahora, veremos la descripción de los distintos métodos implementados.

6.1 Deducción de hogares

Este método de deducción de puntos de interés se describe [6] con tal de evaluar el algoritmo de swapping descrito en el propio paper.
El algortimo de deducción de hogares consiste en discretizar el espacio en celdas cuadradas de 111,32 metros -o lo que es lo mismo, 0,001º decimales- y contar cuáles son las celdas más pobladas para cada sujeto, considerando que la más poblada debería contener la localización de su casa u hogar.
Para este trabajo, se ha seguido exactamente el mismo algoritmo que el descrito previamente. En lugar de usar directamente la localización más repetida y establecerla como el hogar del individuo, se han obtenido las 5 localizaciones más repetidas para cada individuo con tal de poder estudiar con más detalle las distintas casuísticas.

6.2 Extracción de *stays*

Un *stay* o estancia hace referencia a pasar cierto tiempo en algún lugar determinado. Se podría considerar un stay como una visita de 5 minutos al baño, una visita turística de un día a una ciudad o un viaje de trabajo a una ciudad extranjera de 5 días. Como se puede ver, una estancia se puede dar en distintas escalas geográficas y en distinta escala de tiempo, de manera que un stay en una determinada escala geográfica y temporal puede ser útil para alguna aplicación, pero no para otras.
Queda claro que la definición de lo que se considerará un stay dentro de un conjunto de trazas de un sujeto quedará determinado por un umbral de tiempo τ y un umbral de distancia χ. En el caso de este trabajo, el umbral de distancia representa la distancia máxima a la que se puede alejar un individuo de un punto geográfico concreto para poder considerarlo un stay, y el umbral de tiempo representa el tiempo mínimo que un sujeto debe permanecer dentro del umbral de distancia respecto a un punto para considerarse una estancia o stay en ese punto.
Este algoritmo devuelve un conjunto de stays $S = \{s_i\}$, donde cada stay s viene definido por la latitud y longitud del lugar, el tiempo de inicio y el tiempo de fin, de manera que

$$s_i = (lat_i, long_i, t\_ini_i, t\_fin_i)$$

Este algoritmo se puede clasificar como un algoritmo iterativo de clústering. En [1], se ha etiquetado a este algoritmo como "Density-Time cluster (DT-Cluster)".

En la siguiente figura se define el algoritmo usado para implementar este método de deducción de hogares, que está basado en el algoritmo de clústering de [8], aunque se ha escrito de nuevo el pseudocódigo en lugar de usar el pseudocódigo original.

Algoritmo de extracción de stays

Input Dataset D
Output Conjunto stays S
para cada ID distinto en D **hacer**
ordenar trazas del subconjunto D_{id}
para cada traza ordenada del subconjunto D_{id} **hacer**
p ← min dt | dtj > dti + τ
si max dist(C) > τ **entonces**
pasar a la siguiente traza D_{id}
sino
se obtienen trazas posteriores en el tiempo
mientras traza <= χ **hacer**
j ← traza fin
fin mientras
S ← S ∪ s_i (centroide todas las trazas del subconjunto, fecha traza inicial, fecha traza final)
fin si
fin para
fin para

6.3 Heurística. Método Begin-End location finder.

El método de deducción de puntos de interés de inicio-fin (Begin-End) se basa en la heurística, asumiendo que el primer y último lugares registrados en una jornada de trabajo de un individuo corresponden a los puntos de salida y llegada desde un punto de interés (POI). Se deduce que, si no existe ninguna traza registrada durante un período de tiempo determinado, el sujeto ha tenido un parón en su trabajo, suponiendo que el lugar en el que ha realizado el parón es un punto de interés.
Cuando el umbral de tiempo τ, durante el cual se interpreta que si no se registra ninguna traza se trata de un POI, es lo suficientemente grande, los puntos de interés obtenidos pueden corresponder a la vivienda en la que el sujeto descansa después de la jornada laboral.
A continuación, se muestra el algoritmo que se ha implementado durante este trabajo.

Algoritmo Begin-End location finder

Input Dataset D
Output Conjunto POIs B
para cada ID distinto en D **hacer**
ordenar por fecha trazas del subconjunto Did
para cada traza ordenada del subconjunto Did **hacer**
si primera traza de Did **entonces**
begin = primera traza
última traza = primera traza
sino
si fecha traza actual – última fecha > τ **entonces**

```
            B = B + begin
            B = B + traza anterior a la actual
            begin = traza actual
        fin si
        última traza = traza actual
    fin si
  fin para
fin para
```

7 Evaluación empírica

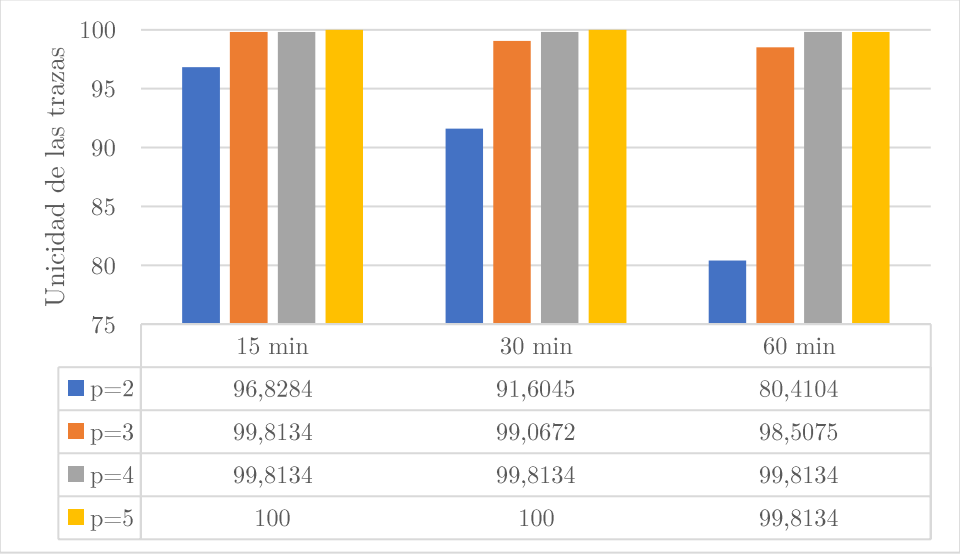
A continuación, se evalúan los distintos métodos de sanitización implementados durante el trabajo, utilizando distintos parámetros de entrada según el caso.
Para la realización de este trabajo se ha usado el dataset CabSpotting. Este dataset consiste en un conjunto de trazas de movilidad correspondientes a 536 taxis de la ciudad de San Francisco del estado de California de EEUU, tomadas entre mayo y junio del año 2008. En total, el dataset dispone de más de 11 millones de trazas.
El dataset CabSpotting se puede conseguir con fines de investigación y docentes registrándose en Crawdad, una comunidad del Dartmouth College de EEUU (https://crawdad.org//~crawdad/epfl/mobility/20090224/cab/). Cada uno de los métodos de sanitización se ha aplicado sobre el dataset CabSpotting del Dartmouth College.

7.1 CabSpotting – Dataset original

Para poder evaluar los métodos de sanitización implementados en este trabajo, en primer lugar, llevaremos a cabo los distintos ataques contra el dataset CabSpotting. Los resultados obtenidos al atacar este dataset serán la base para la evaluación de los resultados obtenidos de aplicar los mismos ataques a los distintos datasets sanitizados.

7.1.1 Unicidad

Para este trabajo, siempre se calculará la unicidad con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5. En la siguiente imagen, se pueden ver los resultados del cálculo de la unicidad de las trazas del dataset D.



	15 min	30 min	60 min
p=2	96,8284	91,6045	80,4104
p=3	99,8134	99,0672	98,5075
p=4	99,8134	99,8134	99,8134
p=5	100	100	99,8134

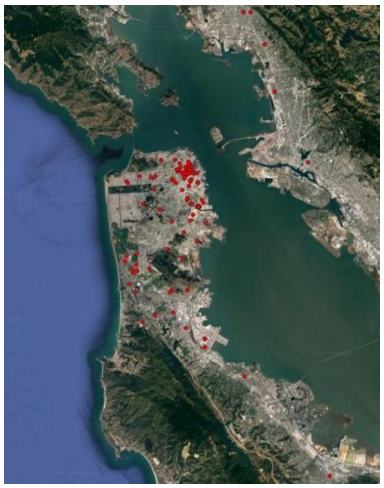
La unicidad εse expresará siempre en porcentaje (%). En el caso del dataset CabSpotting, vemos que para 2 puntos, la unicidad es muy alta cuando el umbral τ es de 15 minutos, para 30 minutos está por encima del 90% y, en el caso de τ a 60 minutos, vemos que εes más bajo, rondando el 80%. Para p>=3, ε toma valores muy elevados, por encima del 98% independientemente del umbral de tiempo.

7.1.2 Ataques

A continuación, se presentan los resultados de aplicar los distintos ataques implementados en este trabajo.

7.1.2.1 Deducción de hogares

A continuación, se estudian los resultados de aplicar el ataque de deducción de hogares sobre el dataset original.
En la siguiente imagen, se pueden observar los hogares deducidos para todos y cada uno de los sujetos del dataset original. Estos hogares corresponden con las localizaciones más repetidas para cada sujeto.



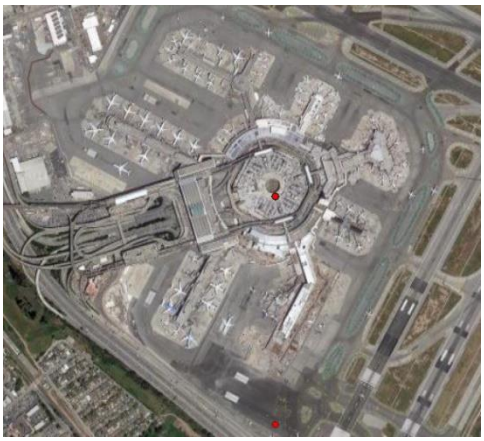
Se puede ver fácilmente en el mapa que hay algunas localizaciones que es muy probable que correspondan realmente con los hogares de los taxistas. Corresponden con aquellas que se encuentran marcadas por rectángulos rojos en la siguiente imagen. Estas localizaciones son las que se encuentran más alejadas de la ciudad de San Francisco (de hecho, no se encuentran en la ciudad), por lo que, teniendo en cuenta que los taxistas desarrollan su trabajo en esa ciudad, el hecho de que la localización más repetida para un individuo se encuentre tan alejada significa, muy probablemente, que se trata de su hogar.

Mediante la observación con detalle de cada uno de los puntos, se comprueba que, efectivamente, en todos los casos se trata de zonas residenciales que no se encuentran en la ciudad de San Francisco, por lo que se deduce que deberían ser sus viviendas.
Vemos en la siguiente imagen una captura de imagen de la localización (37,868, -122,296), que corresponde con el hogar deducido para el sujeto *icwiroic*. Realizando una búsqueda con Google

Maps, vemos que se trata de una Inn (posada u hostal) en la ciudad de Berkeley, en la que probablemente durme el trabajador.

En la siguiente imagen se muestra la localización (37,916, -122,309), que corresponde con el hogar deducido para el individuo *oncalku*. Si buscamos la localización, vemos que se trata de una casa particular en un barrio residencial de El Cerrito.

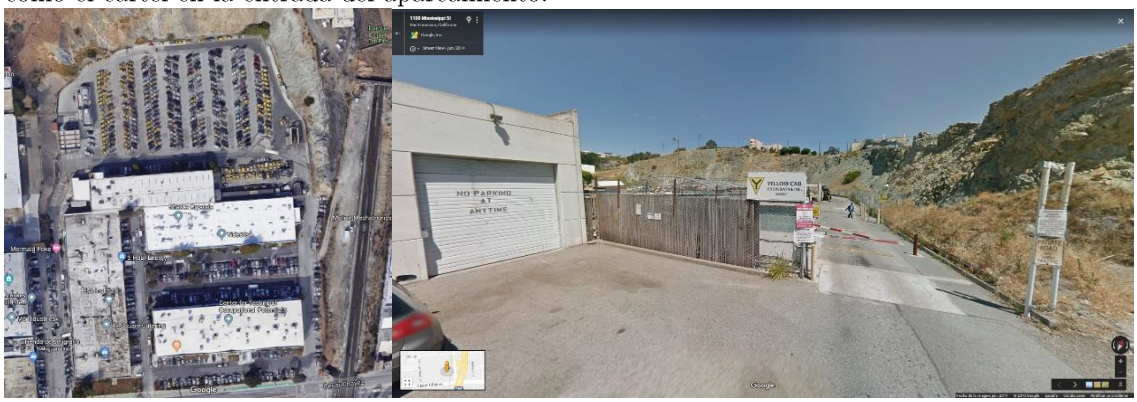
De los 536 taxistas, existen 104 de ellos, es decir, un 19,4%, cuyo hogar coincide exactamente en el punto (37,616, -122,386). Si buscamos este punto, vemos que se trata del parking del aeropuerto internacional de San Francisco, por lo que se deduce que cerca del 20% de los hogares deducidos no corresponden con el hogar real de los sujetos, sino con un lugar que frecuentan mucho debido a su trabajo. En la siguiente figura se puede ver el mapa en el que se muestra el hogar deducido en este caso.



Pero sin duda, la localización que más veces se ha deducido como el hogar de un individuo, corresponde con (37,751, -122,394). 258 de los 536 sujetos del dataset, es decir, el 48,1% de los sujetos, tienen este punto como hogar deducido por el ataque. Los resultados revelan que, además, los puntos (37,751, -122,395) y (37,751, -122,393), corresponden al 6,7% y el 2,8% de hogares deducidos respectivamente. Por tanto, más de la mitad de hogares deducidos corresponden a la

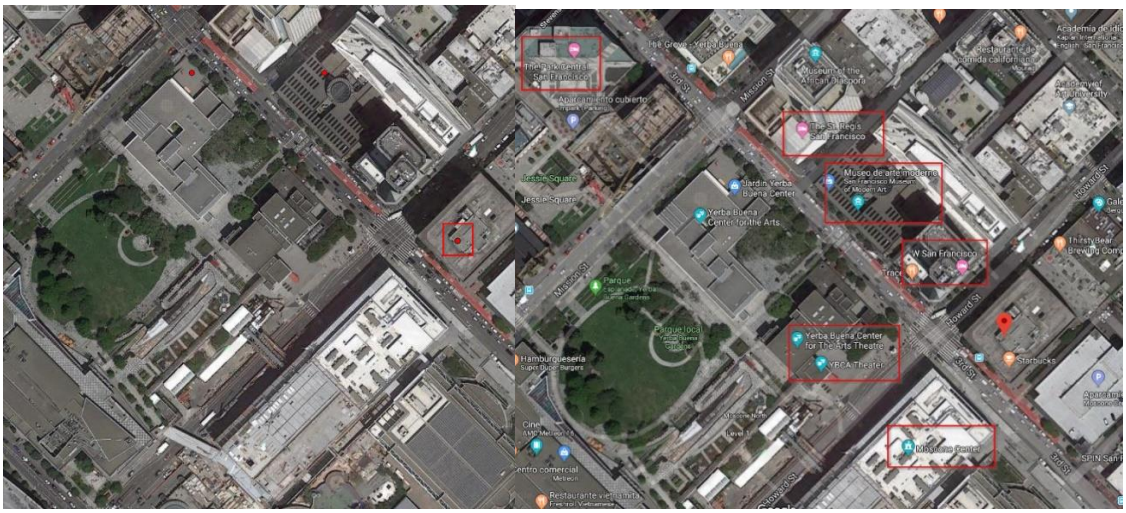
misma zona de unos pocos cientos de metros. En el siguiente mapa se muestran las tres localizaciones que corresponden con los hogares deducidos.

Si se investiga lo que existe en esta zona usando Google Maps, descubrimos que esta localización corresponde con el aparcamiento de la empresa Yellow Cab Cooperative Inc. donde los taxistas aparcan sus taxis. Todas las trazas del dataset CabSpotting corresponden a la empresa Yellow Cab. En la siguiente imagen se puede ver una vista aérea en la que se ven los taxis aparcados, así como el cartel en la entrada del aparcamiento.



Más del 70% de hogares deducidos mediante este ataque corresponden al aparcamiento de la empresa de taxis o al aeropuerto. Aunque no corresponde con los lugares en los que habitan los taxistas, en este caso el ataque aporta información relevante, ya que permite deducir el lugar de trabajo de los sujetos.

El hogar deducido que ocupa el cuarto lugar en cuanto a más repetido corresponde a la localización (37,785, -122,4). Alrededor de este punto, a muy poca distancia, se han deducido otros hogares, tal como se puede ver en el siguiente mapa, en el que la posición indicada viene marcada con un recuadro. Si en la misma imagen de Google superponemos la información de las localizaciones, vemos que en la zona en la que se han deducido los hogares hay varios hoteles importantes, así como centros culturales relevantes. Eso, sumado a que no es una zona de viviendas, indica claramente que se han deducido lugares en los que los taxistas pasan mucho tiempo, pero que no son sus viviendas.



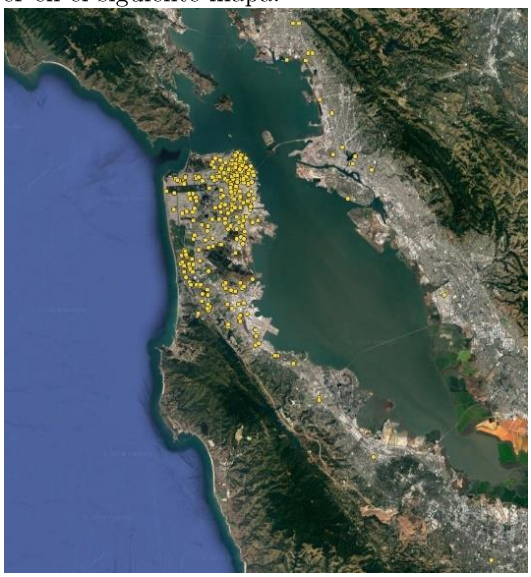
Revisando los distintos hogares deducidos, vemos que los hogares deducidos en la zona noreste de San Francisco (ver siguiente imagen), entre los que se encuentran los comentados justo arriba, corresponden con zonas en que los taxistas pasan mucho tiempo por razones de trabajo, ya que no se trata de zonas residenciales.

De los hogares deducidos, aproximadamente menos de un 10% encaja con las características del lugar en el que puede vivir un individuo. El resto, como en parte hemos visto más arriba, corresponde a lugares en los que los individuos pasan mucho tiempo por razones laborales.

Hasta ahora hemos comentado los resultados obtenidos al calcular los hogares como aquella localización que más se repite para un individuo. Si tenemos en cuenta las 5 localizaciones más repetidas para cada sujeto, tenemos que los puntos de interés deducidos más relevantes siguen siendo los mismos, aunque no en el mismo orden. El hogar deducido más repetido en este caso pasa a ser el aeropuerto de San Francisco. Las 3 localizaciones siguientes más repetidas corresponden al aparcamiento de la empresa Yellow Cab. En este caso, hasta el quinto lugar con más repeticiones sube la localización (37,776, -122,395), que si se tienen en cuenta solamente los lugares más repetidos, aunque ocupa el sexto lugar, tiene un número de repeticiones bajo muy similar a otros puntos sin mucho interés. Este lugar corresponde a la estación de tren de San Francisco.

Observando el set de registros obtenido al calcular los 5 lugares más repetidos, se puede deducir que aquellos lugares que solamente se repiten una vez, es decir, solamente aparecen como un punto

importante para un solo sujeto, es muy probable que corresponda al hogar de ese individuo. Esas localizaciones se pueden ver en el siguiente mapa.



Hay que tener en cuenta que aparecen localizaciones que corresponden al aparcamiento de la empresa de taxis, al aeropuerto, y a la zona noreste de San Francisco en la que se encuentra el distrito financiero y la zona comercial. Si no se tienen en cuenta las localizaciones de estas zonas, se puede considerar que el 10% de las localizaciones obtenidas teniendo en cuenta los 5 lugares más repetidos para cada individuo, corresponden efectivamente a sus hogares. Una revisión pormenorizada de los lugares corrobora esa aproximación.

Si aplicamos el ataque de deducción de hogares solamente sobre aquellas trazas que corresponden a los momentos en los que el taxi está libre, vemos que las 10 localizaciones más repetidas como hogar de un sujeto corresponden con las mismas localizaciones que en el ataque sobre el dataset completo. Esto significa que los hogares deducidos que se repiten entre varios taxistas corresponden a lugares de trabajo. En este caso, en primer lugar, sigue apareciendo el aparcamiento de la empresa de taxis, seguido del aeropuerto, con mucha diferencia respecto al resto. El resto de hogares que se repiten para más de un individuo siguen siendo los mismos que los deducidos en el ataque a todo el dataset. Si se comparan las localizaciones de hogares que solamente corresponden a un solo individuo, vemos que de 78 casos obtenidos con todo el dataset y 75 casos cuando el taxi está libre, la intersección de localizaciones coincide en 73 casos, en un 93,5% de los hogares deducidos sobre todo el dataset. Si comparamos todas las localizaciones deducidas para todo el dataset y cuando el taxi está libre, vemos que la intersección es de cerca del 96%, correspondiente a 90 de los 94 hogares deducidos para todo el dataset. Claramente vemos que los resultados obtenidos al deducir los hogares sobre las trazas cuando el taxi está libre y sobre todo el dataset son prácticamente idénticos, excepto por 4 hogares deducidos y la ligera diferencia en la frecuencia en que se repiten algunos hogares. Esto deja claro que, en todo el dataset, los lugares en los que más trazas aparecen para los taxistas corresponden con los momentos en los que el taxi está libre. El hogar más deducido corresponde con el aparcamiento de la empresa de taxis, hecho lógico puesto que la mayoría de los taxistas dejan el taxi aparcado en el aparcamiento de la empresa en lugar de llevarlo a su vivienda. El resto de hogares más repetidos entre los taxistas corresponden a momentos de espera a clientes -como en el caso del aeropuerto o la zona noreste de San Francisco en la que se encuentran gran cantidad de hoteles y de lugares culturales y de ocio-. Por último, aproximadamente un 10% de los hogares deducidos con este método corresponden muy probablemente con los lugares de residencia de los sujetos.

7.1.2.2 Stays

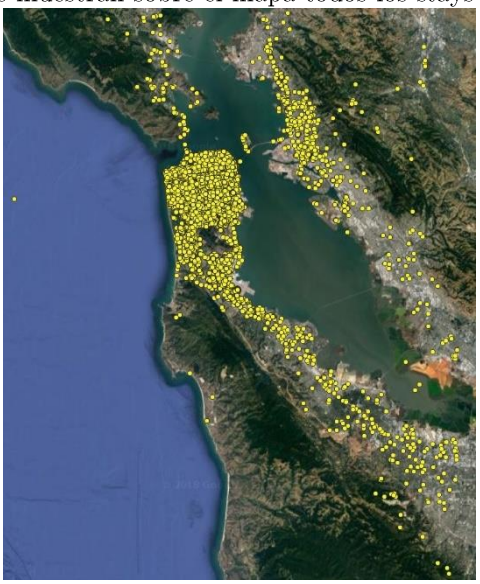
A continuación, se estudian los resultados de aplicar el ataque de extracción de stays mediante el algoritmo DT-Cluster sobre el dataset original.

En el caso de la extracción de stays, por cómo está diseñado el algoritmo DT-Cluster, para un mismo umbral de distancia χ, las localizaciones de los stays obtenidos para un umbral de tiempo $\tau < i$ están incluidos dentro del conjunto de stays que se extraen para un umbral de tiempo i. Es por eso que se ha decidido optar por ejecutar la extracción de stays usando un umbral de tiempo τ bajo, que correspondería a 5 minutos. Recordemos que la definición de un stay es la siguiente:

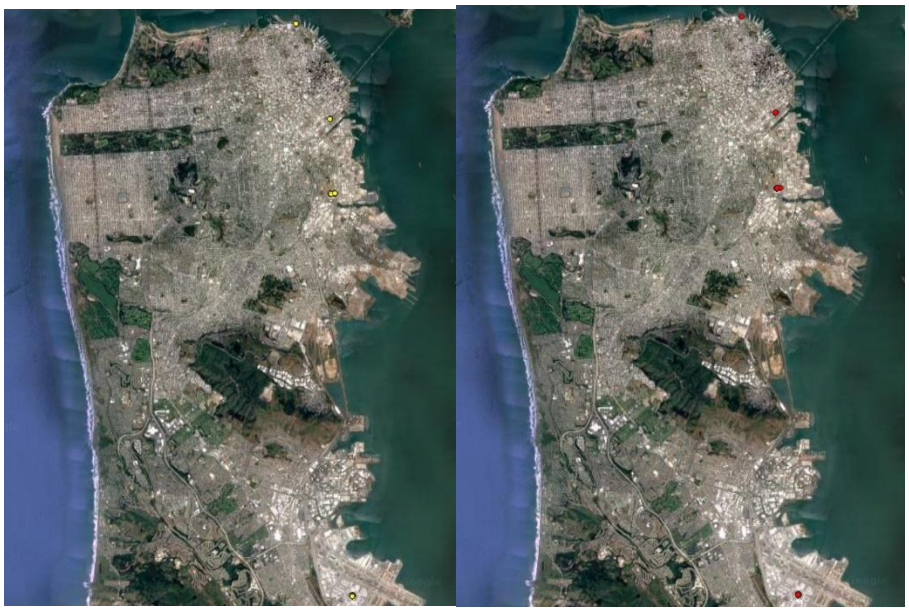
$$s_i = (lat_i, long_i, t\_ini_i, t\_fin_i)$$

Así pues, disponemos de la fecha de inicio y de fin de cada stay, lo que nos permite filtrar los stays extraídos en función de nuestras necesidades, y al usar un umbral de tiempo bajo, conseguiremos un conjunto de stays que incluirá desde stays de pocos minutos hasta stays de varias horas. Para el umbral de distancia χ, se han usado los valores de 50 y 100 metros. Así pues, a efectos prácticos, en este trabajo se considerará un stay como aquella localización en la que un sujeto pase al menos 5 minutos y no se aleje más de 50 o de 100 metros dependiendo del caso.

Al ejecutar el algoritmo DT-Cluster con el umbral de tiempo τ a 5 minutos y el umbral de distancia χa 50 metros con tal de extraer los stays para el dataset CabSpotting, se han obtenido un total de 105590 stays. Se ha extraído al menos un stay para cada uno de los 536 individuos de D. En la siguiente imagen se muestran sobre el mapa todos los stays extraídos para D.



Estudiando el conjunto de stays obtenidos, lo primero que buscaremos son aquellos POIs que son comunes para todos los sujetos del dataset D. Las localizaciones de stays que más se repiten entre todo el conjunto de stays extraído con umbral de tiempo τ a 5 minutos y umbral de distancia χ a 50 metros se pueden ver a la izquierda de la siguiente imagen en amarillo. A la derecha, en rojo, se muestran los puntos de stay más repetidos para τ a 5 minutos y χa 100 metros para poder comparar los resultados obtenidos con el umbral de distancia a 50 metros. Vemos que los POIs extraídos son exactamente los mismos.



Tal como se puede ver, son exactamente 4 las zonas a las que pertenecen los stays que más se repiten. El punto de interés con más stays corresponde al aparcamiento de la empresa de taxis Yellow Cab. El 17,2% de los stays extraídos con τ a 5 minutos y χa 50 metros corresponde a esta localización. El segundo POI con más stays corresponde al aeropuerto internacional de San Francisco, con el 11% de los stays obtenidos. Con mucha diferencia, estas dos localizaciones son aquellas en las que los sujetos hacen más paradas, ya sean muy pocos minutos hasta durante horas. El tercer lugar con más stays corresponde con la estación de tren. En este caso, en cambio, esta zona solamente contiene el 2% de los stays obtenidos. Por último, la zona del Pier 39, en la que se encuentra el acuario de San Francisco, obtiene el 1% de los stays. El detalle de estas dos últimas zonas se puede ver en la siguiente imagen.

El aparcamiento de taxis es la zona con más stays por la razón obvia de que los taxistas aparcan allí los taxis entre jornadas o hacen paradas más cortas en el aparcamiento, debido a que es la base de la flota de taxis. Los otros 3 casos que acabamos de comentar son los siguientes más repetidos porque al tratarse de zonas con parada de taxi, los taxistas realizan esperas en un espacio físico muy reducido y definido.

Para la obtención de POIs a partir de los stays calculados, por los datos de los que dispone cada stay, es muy útil subdividir los stays por la duración de éstos, ya que, dependiendo de la duración de los stays, podemos relacionar los puntos de interés con distintas actividades, tanto laborales

como personales. Como punto importante, hay que recalcar que el aparcamiento de la empresa de taxis se mantiene, en cualquier subconjunto de duración de los stays, como el punto de interés más repetido.

En el subconjunto de stays de duración de entre 15 y 30 minutos existen tres tipologías de POIs que prevalecen sobre cualquier otra: las esperas a clientes, los breaks o descansos de los taxistas y las paradas para repostar en gasolineras. Claramente, la duración de estas actividades está totalmente relacionada con que éstos sean los tipos de stays que imperan en este intervalo. En la siguiente imagen se muestran sobre el mapa gran parte de lo stays con duración entre 15 y 30 minutos.

En la zona noreste de San Francisco, en la que se encuentra el distrito financiero, instalaciones culturales y de ocio, y, sobre todo, en la que hay una gran cantidad de hoteles, se concentran gran parte de los stays que se encuentran en este intervalo de tiempo. En la siguiente imagen se puede observar la densidad de stays en toda la zona noreste a la izquierda, y a la derecha se ve con más detalle cómo los stays se agrupan en distintas zonas.

En la parte derecha de la imagen anterior, se extraen agrupaciones tan definidas de stays en esta zona debido a que cada agrupación de stays corresponde con un hotel relevante. En la imagen se ven grupos de stays alrededor de los hoteles W, Palace, Marriott, Hilton y Fairmont (el más repetido en el conjunto de todos los stays) entre otros. Para hacernos una idea de la magnitud de los stays calculados alrededor de hoteles, en la zona de hoteles que acabamos de comentar (derecha de la imagen anterior), los stays extraídos para τ a 5 minutos y χa 50 metros sin restricciones

de duración del stay, representan el 18,1% del total de stays obtenidos. Si ampliamos la zona evaluada a la zona que se encuentra dentro del recuadro rojo en la imagen anterior, tenemos que, los stays que se encuentran en la zona delimitada corresponden con el 36% de los stays obtenidos para τ a 5 minutos y χ a 50 metros sin restricciones de duración del stay. Queda claro que la zona noreste de San Francisco es una de las zonas con más actividad dentro del dataset CabSpotting y es una fuente muy importante para la obtención de POIs. En esta zona, las estancias suelen ser tirando a cortas. Más adelante veremos que (a excepción de la estación de tren), cuanto más dure un stay, menos probabilidades hay de que se encuentre en esta zona.
En la siguiente imagen se pueden ver marcadas sobre el mapa 3 gasolineras que se encuentran justamente al sur de la zona hotelera que acabamos de ver. Se puede apreciar la gran cantidad de stays entre 15 y 30 minutos que se agrupan sobre ellas.

En cuanto a los breaks o momentos de descanso de los conductores, en la siguiente imagen se puede observar una zona de varias calles en la que se agrupan stays sobre distintos restaurantes de comida asiática, pizzerías, mexicanos, etc. En la parte derecha, a modo ilustrativo, se muestran los stays obtenidos en la misma zona para el intervalo entre 1 y 3 horas. Queda claro que los descansos de los sujetos duran más de un cuarto de hora y menos de una hora.

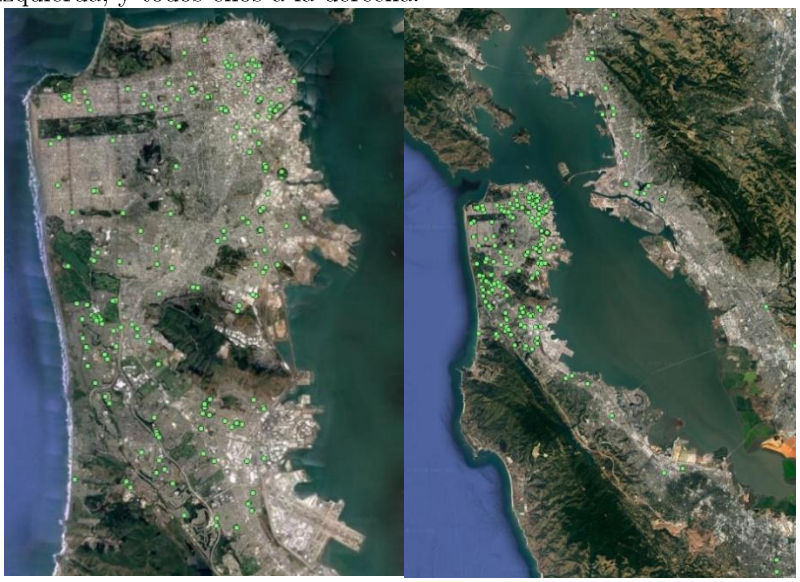
Existen gran cantidad de stays de entre 15 y 30 minutos que corresponden a cafeterías, a lavaderos de coches, a talleres de neumáticos, etc.

Observando los stays de entre 1 y 3 horas de duración, se obtienen stays interestantes porque mezclan los lugares de trabajo con lugares personales de los taxistas. Uno de los lugares en los que se detecta una densidad alta de stays, y que se encuentra lejos del centro de San Francisco y cerca del aeropuerto corresponde a un casino, aproximadamente sobre (37.627, -122.41051). No queda del todo claro si estos stays corresponden a esperas de clientes o a actividad personal de los sujetos.

En el siguiente ejemplo, podemos ver que los stays de entre 1 y 3 horas nos pueden dar información acerca de los domicilios de los individuos. La combinación con stays de entre 15 y 30 minutos ayuda a dejar claro que se trata de puntos de interés. En la siguiente imagen, a la izquierda podemos ver encuadrados por rectángulos rojos los stays que nos indican con muy alta probabilidad el hogar de dos individuos distintos. A la derecha, vemos un zoom sobre los puntos de interés que se encuentran más al sur. Se puede deducir incluso cuál es la casa del taxista por cómo están distribuidos los stays.

En la siguiente imagen se muestran parte de los stays de duración de entre 1 y 3 horas. Se puede observar que el aeropuerto y la zona noreste de SF siguen siendo zonas con alta densidad de stays, aunque obviamente ha bajado la densidad.

Cuando se tienen en cuenta los stays de más de 3 horas de duración, se detecta fácilmente sobre el mapa que la cantidad de stays deducidos para duraciones grandes se ve muy reducido. De hecho, solamente el 2,2% de los stays extraídos para τ a 5 minutos y χa 50 metros corresponden a stays de más de 3 horas, y sólo 481 de los 536 sujetos tienen algún stay calculado para esta duración. En la siguiente imagen se pueden observar parte de los stays extraídos para esta duración a la izquierda, y todos ellos a la derecha.

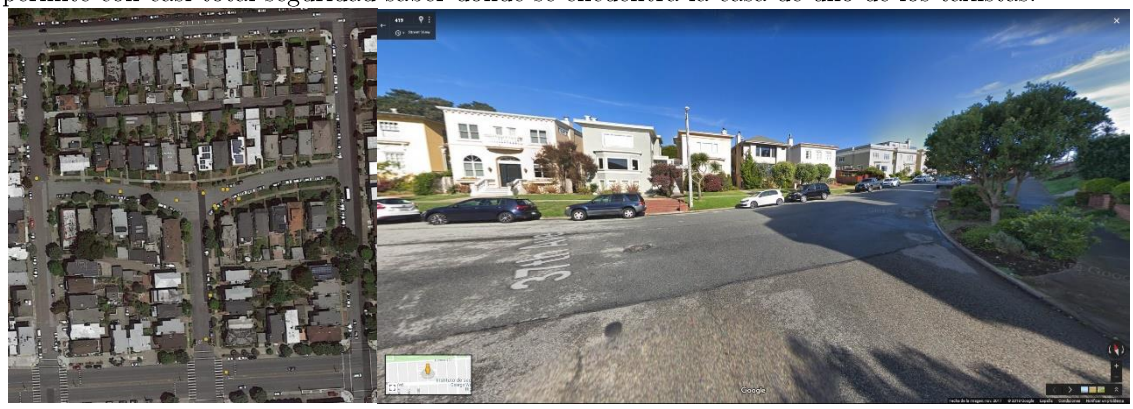


Uno de los hechos más llamativos cuando se observa este subconjunto de stays, y que se puede ver claramente en la anterior imagen, es que no existe ningún stay en el aeropuerto internacional de San Francisco. Todos los stays extraídos que corresponden a ese POI son de menos de 3 horas de duración. En este subconjunto de stays de más de 3 horas de duración, se puede considerar que, aquellos stays que no se encuentran en zonas que ya hemos visto previamente que son POIs comunes para todos los sujetos -como la estación de tren o el aparcamiento de taxis- es muy probable que correspondan a lugares de descanso o de ocio personal de los taxistas.

Si analizamos aquellos stays de más de 5 horas de duración, vemos que sólo el 1,1% de todos los stays extraídos para τ a 5 minutos y χa 50 metros corresponden a stays de más de 5 horas, y sólo 481 de los 536 sujetos tienen algún stay calculado para esta duración. En la siguiente imagen se pueden ver todos los stays de duración de más de 5 horas obtenidos.

Si no tenemos en cuenta los stays del aparcamiento de taxis, vemos que aproximadamente el 17 % del subconjunto de stays de más de 5 horas corresponde con una probabilidad muy alta a los

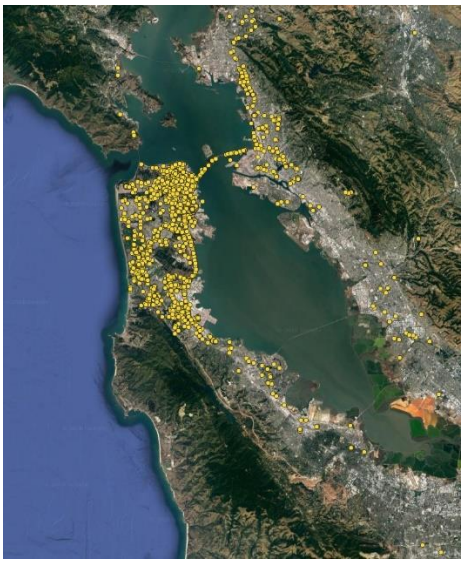
hogares o lugares en los que han descansado un largo tiempo los sujetos. Para el total de stays extraídos, esto representa aproximadamente el 0,2% de todos los stays. Todavía existen stays en la zona de la estación de tren y en la zona hotelera del noreste. Sin embargo, en esas zonas se concentran varios stays para los mismos individuos, lo que significaría que realmente aparcan en esa zona el taxi. En el caso de la estación de tren, podría indicar que los sujetos se van hacia otra zona en tren.
En la siguiente imagen se puede observar un caso muy claro para stays de más de 5 horas que permite con casi total seguridad saber dónde se encuentra la casa de uno de los taxistas.



Para el sujeto con pseudónimo *oasthul*, existen 13 stays correspondientes a 13 días distintos no consecutivos en torno a (37,780, -122,497). Por la distribución de los stays, se deduce que su vivienda puede ser alguna de las mostradas a la derecha en la imagen superior.
Queda claro que este ataque consigue una gran cantidad de información sobre los *stays* de los sujetos, lo que permite a un atacante, mediante el análisis de los resultados, obtener una gran cantidad de POIs, tanto para los hábitos laborales como para los personales de los individuos.

7.1.2.3 Begin-End locations

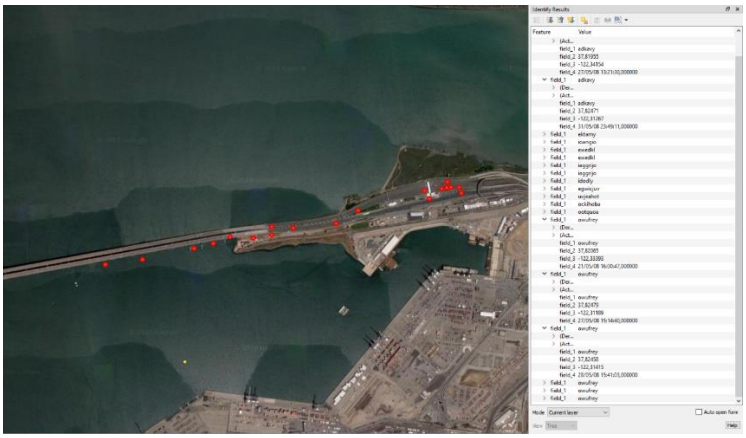
A continuación, se estudian los resultados de aplicar el ataque de extracción de Begin-End (inicio-fin) sobre el dataset original.
Por cómo está diseñado conceptualmente este ataque, en este caso obtenemos diferente cantidad de localizaciones para cada sujeto, dependiendo de su comportamiento. Se ha aplicado este ataque con distintos valores para el umbral de tiempo τ. Los ataques se han realizado con τ =28800 segundos (8 horas), τ =21600 s (6 horas), τ =18000 s (5 h), τ =16200 s (4,5 h) y τ =14400 (4 h).
Con un umbral de 8 horas (τ =28800 s) entre dos trazas para un individuo, obtenemos 4042 localizaciones, de las cuales solamente en 13 casos la misma localización se repite dos veces para el mismo individuo. Hay que tener en cuenta que la precisión de las trazas es de 0,00001º decimales, lo que implica una precisión de pocos metros, por lo que, si el taxi no permanece en prácticamente en el mismo punto exacto, la localización no coincidirá exactamente en distintas ocasiones.
Con este valor de τ, se han obtenido localizaciones solamente para 445 sujetos, un 83% de los taxistas que existen en el dataset. Para el resto de individuos, no existe en ningún caso en el dataset un período de tiempo de al menos 8 horas sin registrar su localización, seguido de una nueva traza. Podríamos decir que desde el punto de vista de este ataque, solamente existe una sola jornada para estos sujetos en todo el dataset.
En el siguiente mapa, se muestran todos los resultados obtenidos.



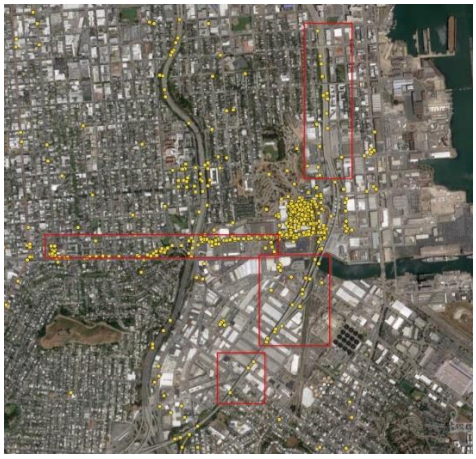
Aquellos lugares encontrados que se encuentran alejados del centro de San Francisco y que corresponden a zonas residenciales, corresponden con mucha probabilidad a lugares importantes en la vida cotidiana del taxista, ya que implica que en esa ubicación ha empezado o acabado un período de 8 horas sin reportar su posición.

Si vemos con detalle en el mapa algunas de las localizaciones extraídas, vemos que en varias zonas distintas hay agrupaciones de trazas muy cercanas. Tal como se puede ver en la siguiente imagen, esto se debe a que, muchas de las tratas corresponden al mismo individuo. Debido a la precisión de pocos metros de las trazas, si el taxista aparca en distintos lugares cada día, acaba creando zonas con una alta densidad de trazas encontradas.

Estudiando el mapa vemos algunas características curiosas. Una de ellas es que muchos de los POIs encontrados se encuentran sobre el puente que une San Francisco con Oakland o sobre autopistas que acceden a San Francisco desde el sur y sureste. Estudiando estas trazas, tanto por su localización como por la fecha y hora, deducimos que los sujetos pueden encender y apagar el dispositivo que envía las trazas a voluntad. Estos casos corresponden con el inicio o fin de la jornada laboral de los taxistas a los que pertenecen las trazas, lo que sugiere que empiezan o dejan de enviar su localización mientras se dirigen a trabajar a San Francisco o mientras se dirigen a sus hogares después de la jornada laboral. Además, las localizaciones de los mismos individuos se agrupan en zonas muy cercanas. En la siguiente imagen de QGIS se pueden ver varias trazas obtenidas con el ataque, que se encuentran sobre el puente SF-Oakland. Se puede ver que varias de las trazas extraídas corresponden a los mismos sujetos en distintos momentos, lo que sugiere que de manera repetida encienden o apagan el localizador en esta zona.



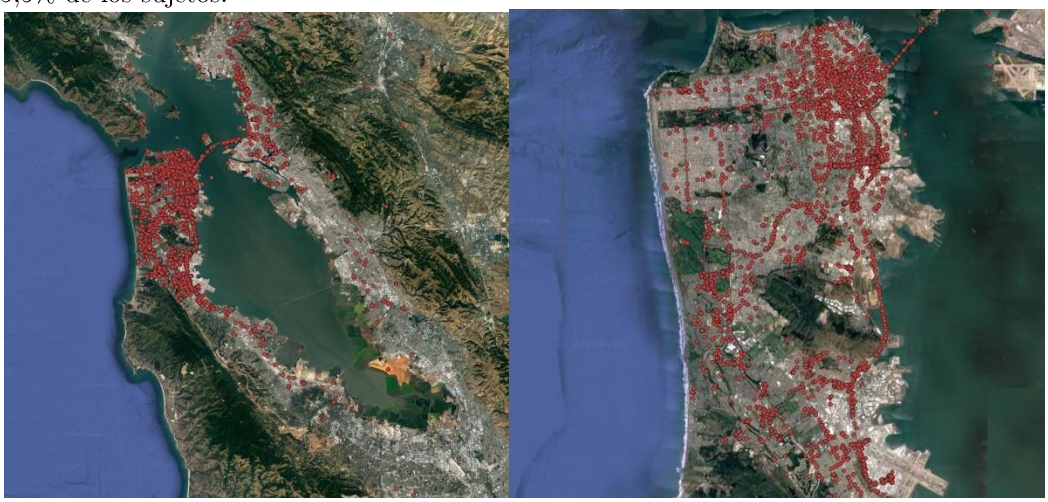
Relacionado con este comportamiento, vemos de nuevo el aparcamiento de Yellow Cab taxis, que en este ataque también recibe una gran densidad de las trazas extraídas, tal como se ve en la siguiente imagen.



Las trazas extraídas por el ataque se encuentran en las grandes vías de acceso al aparcamiento de la empresa. Se deduce que los taxistas para los que se han encontrado estas trazas empiezan a enviar su localización una vez han empezado su jornada de trabajo y ya han abandonado el aparcamiento, y dejan de enviarla cuando están volviendo al aparcamiento, pero aún no han llegado a éste.

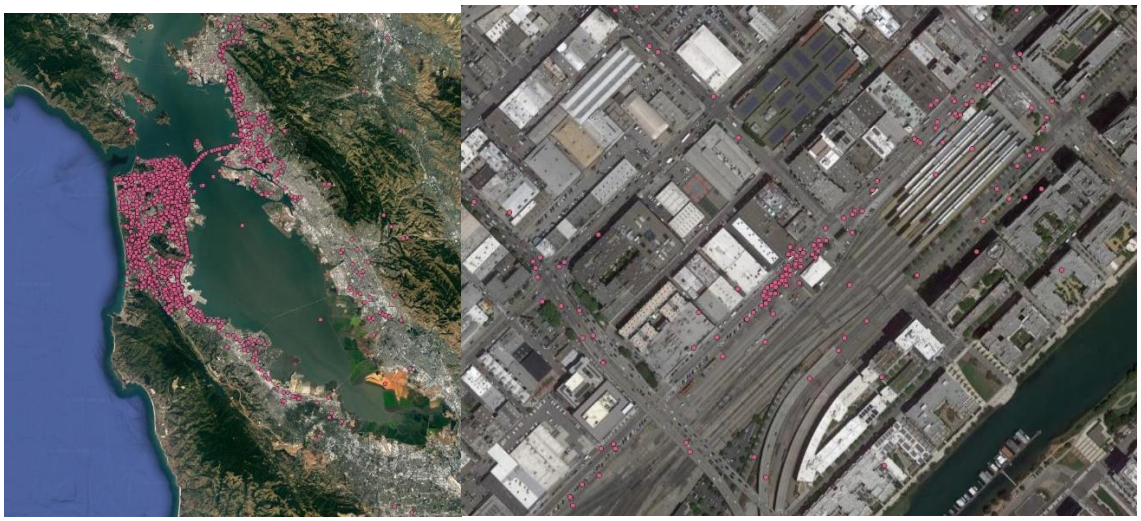
Si bajamos la precisión de los resultados obtenidos a 0,001º decimales y agrupamos los resultados para cada sujeto con esa precisión, vemos que los resultados siguen siendo válidos pese a la pérdida de precisión. Agrupar los resultados para cada taxista con una precisión más baja, favorece el estudio de los resultados sobre el mapa cuando hay mucha densidad de localizaciones en la misma zona. En la imagen de debajo a la derecha se puede ver la comparación entre las localizaciones encontradas por el ataque en amarillo con las localizaciones agrupadas por individuo en azul, en dos zonas con alta densidad de puntos.

Con un valor τ =21600 s (6 horas), el ataque obtiene un total de 6848 localizaciones, pero lo más importante es que el número de individuos para los que se extrae alguna localización sube a 512, el 95,5% de los sujetos.



Con este umbral de tiempo τ de 6 horas, se suman nuevos resultados a los obtenidos con τ a 8 horas que, en las zonas residenciales y zonas alejadas de San Francisco, claramente indican el descubrimiento de nuevos puntos de interés que probablemente corresponden a los hogares de nuevos individuos. Al bajar el intervalo de tiempo, se suman más repeticiones de localizaciones en zonas que ya tenían una elevada densidad de puntos, como son el aparcamiento de la empresa de taxis o el aeropuerto. En este caso, también las carreteras y el puente SF-Oakland quedan mucho más marcados debido a la aparición de nuevas localizaciones.

Realizando el ataque con un valor τ =14400 s (4 horas), el ataque obtiene un total de 12692 localizaciones. El número de individuos para los que se extrae alguna localización sube a 530, es decir, el 98,8% de los sujetos. Hay que tener en cuenta que al bajar tanto el umbral τ, el ataque obtiene casi el doble de localizaciones que cuando corresponde a 6 horas, y un poco más del triple que cuando el umbral se establece a 8 horas.



Mientras en el caso de τ a 6 horas la mayoría de las localizaciones de zonas residenciales dentro de la propia ciudad de San Francisco, aparte de lugares como centros de transportes y del aparcamiento de la empresa de taxis, correspondían muy probablemente a los lugares de residencia de los individuos, cuando se baja el valor de τ a 4 horas, empezamos a obtener localizaciones que pueden corresponder a puntos de interés de los individuos distintos de su residencia, como podrían ser lugares de descanso dentro de la jornada laboral o lugares de ocio, lo que significaría dejar de enviar trazas durante ese descanso. El ataque también obtiene más densidad de localizaciones en lugares como la estación de tren de San Francisco, tal como se ve en la imagen anterior, indicador de que algunos taxistas podrían dejar de enviar trazas mientras esperan clientes en el aparcamiento.
En general, si se bajara todavía más el umbral τ, se obtendrían aún más localizaciones, pero el concepto de la heurística por el que, al dejar un umbral de tiempo suficiente, se obtendrá un POI del sujeto al empezar y terminar su jornada laboral (begin-end), se vería difuminado puesto que en principio entre una jornada laboral de un taxista y la siguiente, debería haber al menos 4 horas, cuando no más. Hay que tener en cuenta que los identificadores del dataset CabSpotting están anonimizados, por lo que no sabemos cuántos taxis conduce más de un sujeto, por lo que la idea de principio y fin de jornada en estos casos no aplica.

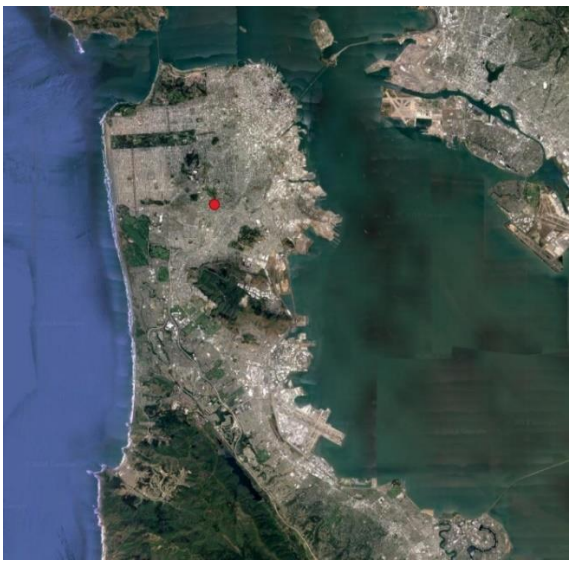
7.2 Métodos de sanitización

A continuación, se evalúan los distintos métodos de sanitización implementados durante el trabajo, utilizando distintos parámetros de entrada según el caso.
Cada uno de los métodos de sanitización se ha aplicado sobre el dataset CabSpotting del Dartmouth College, que ya ha sido descrito previamente en este capítulo.

7.2.1 Perturbación

En primer lugar, evaluaremos el conjunto de métodos de sanitización que se engloban dentro de la perturbación. Básicamente, la perturbación consiste en aplicar sobre un dataset la traslación de sus puntos, la escala, la rotación, o una combinación de varios de estos métodos.
Previamente a evaluar los resultados obtenidos al aplicar la sanitización del dataset original, hay que especificar que para este trabajo se ha escogido la localización (37,735085, -122,441601) como el punto de origen arbitrario sobre el que se aplicarán las transformaciones de escala y rotación. Tal como hemos visto anteriormente en este trabajo, para aplicar estas dos transformaciones sobre

un punto distinto al origen de coordenadas, se deben trasladar todos los puntos que se quieran transformar de forma que el punto arbitrario quede en el origen de coordenadas, aplicar la transformación, y volver a desplazar todos los puntos transformados respecto al punto arbitrario. En la siguiente imagen se puede observar el punto respecto al que se realizarán la escala y la rotación en este trabajo.



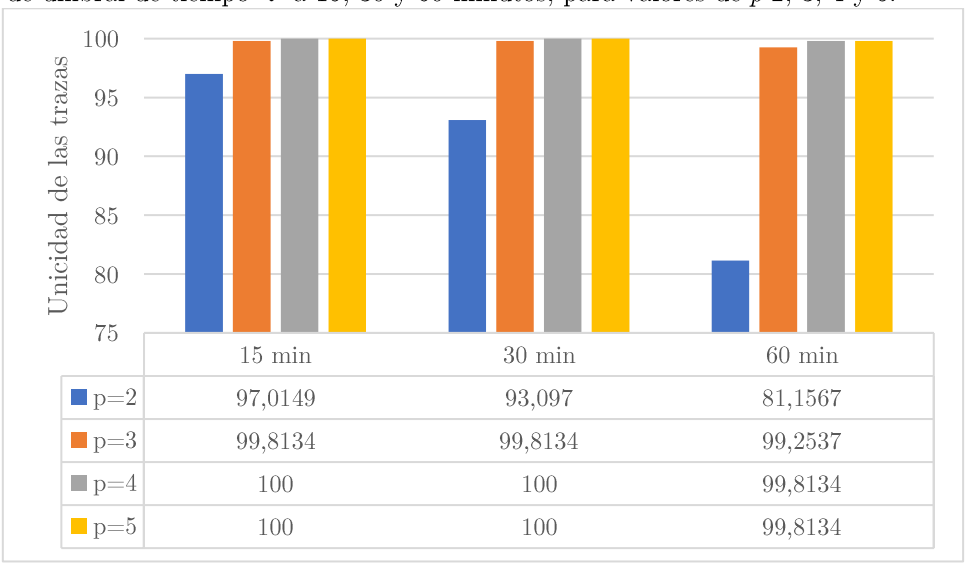
7.2.1.1 Rotación de 2º

Al aplicar una rotación de 2º respecto al punto (37,735085, -122,441601), se obtiene un dataset sanitizado D_s con todas las localizaciones del dataset original en una nueva posición (a excepción de aquellas trazas que se encuentren exactamente sobre el punto de rotación). En la siguiente imagen se pueden observar sobre el mapa todas las trazas de D_s para el sujeto con ID *abboip*.

Se puede ver de forma relativamente fácil el efecto que ha tenido la rotación respecto a las localizaciones. La forma más rápida de apreciar la rotación de los puntos consiste en ver localizaciones singulares, con una forma determinada, o conocidas. En las siguientes imágenes se pueden ver las trazas rotadas que se encontraban en origen en el aeropuerto de San Francisco, en el puente entre San Francisco y Oakland y en el aparcamiento de Yellow Cab taxi. En todas ellas se aprecia claramente hacia dónde se han desplazado las localizaciones de las trazas.

7.2.1.1.1 Unicidad

En la siguiente imagen, se muestra la unicidad ε de las trazas del dataset D_s calculada con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5.



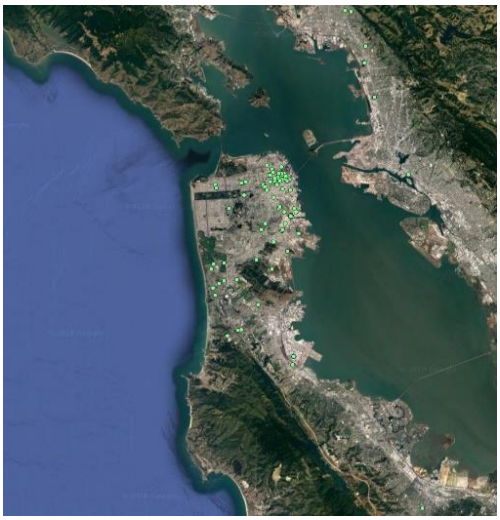
	15 min	30 min	60 min
p=2	97,0149	93,097	81,1567
p=3	99,8134	99,8134	99,2537
p=4	100	100	99,8134
p=5	100	100	99,8134

Comparándola con la unicidad ε calculada para el dataset D, ε para D_s es igual o ligeramente más elevada en todos los casos.

7.2.1.1.2 Ataques

7.2.1.1.2.1 Deducción de hogares

En la siguiente imagen se muestra la localización más repetida para cada uno de los individuos, es decir, el hogar deducido para cada uno, al aplicar el ataque de deducción de hogares sobre el dataset rotado en 2º.



De todos los hogares deducidos para el dataset rotado, solamente uno de ellos coincide con los hogares calculados para el dataset original D. Además, el hogar se ha deducido porque existe el mismo número de trazas en la zona con precisión 0,001 grados que en el dataset original. En este caso en concreto, las 59 trazas que han convertido esta localización en el hogar de un sujeto, se han mantenido en la misma posición con precisión 0,001 que en el dataset original. La principal razón para suceder esto es la proximidad de las trazas respecto al punto de origen sobre el que se han rotado los puntos. Aparte, está claro que ninguno de los puntos se encontraba en la zona limítrofe con la siguiente zona de aproximadamente 111 metros en la dirección de la rotación, porque sino alguna de las localizaciones hubiese pasado al siguiente sector. Este punto se puede ver superpuesto en el mapa en color amarillo. En color rosa se muestra el punto sobre el que se ha rotado el dataset original.

Si se tienen en cuenta las 5 localizaciones más repetidas para cada individuo, 28 de las localizaciones encontradas coinciden con las localizaciones más repetidas para los sujetos en D. En estos 28 casos se incluye el único hogar que coincide con los hogares deducidos sobre D, que hemos comentado previamente.

La mayoría de estos casos no coinciden con la situación del único hogar deducido que coincide con el deducido a partir de D. Excepto el caso comentado previamente y otro de los 27 casos restantes, en el resto de casos (26) no coincide el número de repeticiones de la localización con el número de repeticiones obtenidas con el dataset original. Hay dos razones principales por las que ha sucedido esto. La primera de ellas es que algunas trazas, al ser rotadas, han pasado a formar parte de un nuevo POI deducido que ha coincidido que se encuentra en el mismo punto con precisión 0,001 que un POI deducido para el dataset original. Esto sucede, sobre todo, en zonas en las que ya existía una densidad elevada de puntos de interés deducidos para un mismo individuo. En la siguiente imagen se superponen en el mapa los puntos deducidos con el dataset original D (en azul), los deducidos con el dataset rotado D_s (en amarillo) y los puntos que coinciden (la intersección) entre ambos ataques en color rojo.

La otra razón principal por la que ha sucedido esto son aquellas trazas que, como hemos visto previamente en el caso del único hogar que coincide entre D y D_s, no se han movido del punto con precisión 0,001 en el que se encontraban en D cuando han pasado a D_s.
La combinación de las dos razones principales que acabamos de describir explica las intersecciones de las localizaciones más repetidas para cada individuo entre D y D_s.

7.2.1.1.2.2 Begin-End

En la siguiente imagen se muestran parte de las localizaciones extraídas mediante heurística con el método de inicio-fin con un umbral de tiempo τ correspondiente a 8 (lila), 6 (naranja) y 4 horas (marrón).

Se han obtenido exactamente la misma cantidad de POIs y la misma cantidad de sujetos tienen al menos un punto de interés deducido para cada uno de los valores de τ que al aplicar el ataque sobre el dataset original D. Sin embargo, no se ha encontrado absolutamente ninguna intersección entre los puntos deducidos con el mismo valor de τ para D y para D_s, ninguna de las localizaciones extraídas para cada sujeto coincide. En este caso, al obtenerse localizaciones con una precisión de 0,00001, poco más de 1 metro, se han encontrado el mismo número de localizaciones, pero todas y cada una de ellas se han desplazado, incluso aunque se encuentren cerca del punto sobre el que se ha rotado.

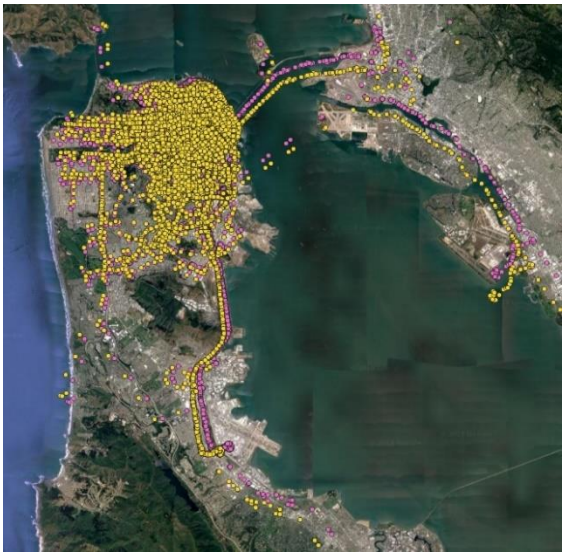
7.2.1.1.2.3 Stays

En la siguiente imagen se muestran todos los stays extraídos de D_s mediante la ejecución del algoritmo DT-Cluster usando los umbrales de tiempo τ a 5 minutos y de distancia χa 50 metros.

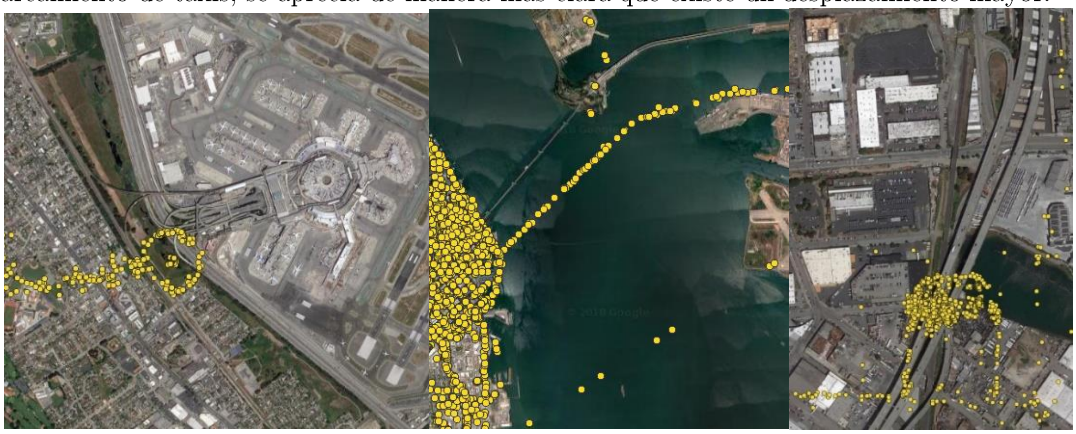
Se ha obtenido 1 stay menos que el obtenido sobre el dataset original D (105589 por 105590). Esto se puede explicar debido a que en D, la traza que convierte en stay este caso se encuentra justo en el límite de los 50 metros del umbral. Si, cuando se aplica la transformación a todos los puntos, la nueva traza en D_s se ha alejado aunque sea un metro y media, puede quedar fuera del umbral de distancia al calcular los stays para D_s. Se ha extraído al menos un stay para todos los sujetos de D_s, es decir 536, el mismo número que al extraer los stays de D. No se ha encontrado ninguna intersección entre los stays obtenidos al aplicar el ataque sobre D_s y sobre D. Al encontrarse todos los puntos de D_s desplazados respecto a los puntos de D, aunque la fecha de inicio y fin de los stays fuese la misma, las localizaciones no coinciden debido a que, aunque se trate de una media entre todas las trazas que componen el stay, la precisión obtenida es de pocos metros.

7.2.1.2 Rotación de 5º

Al aplicar una rotación de 5º respecto al punto (37,735085, -122,441601), se obtiene un dataset sanitizado D_s con todas las localizaciones del dataset original en una nueva posición (a excepción de aquellas trazas que se encuentren exactamente sobre el punto de rotación). En la siguiente imagen se pueden observar sobre el mapa todas las trazas de D_s para el sujeto con ID *abboip* en amarillo, y se puede comparar el resultado con las trazas para el mismo sujeto rotadas en 2º (en rosa).

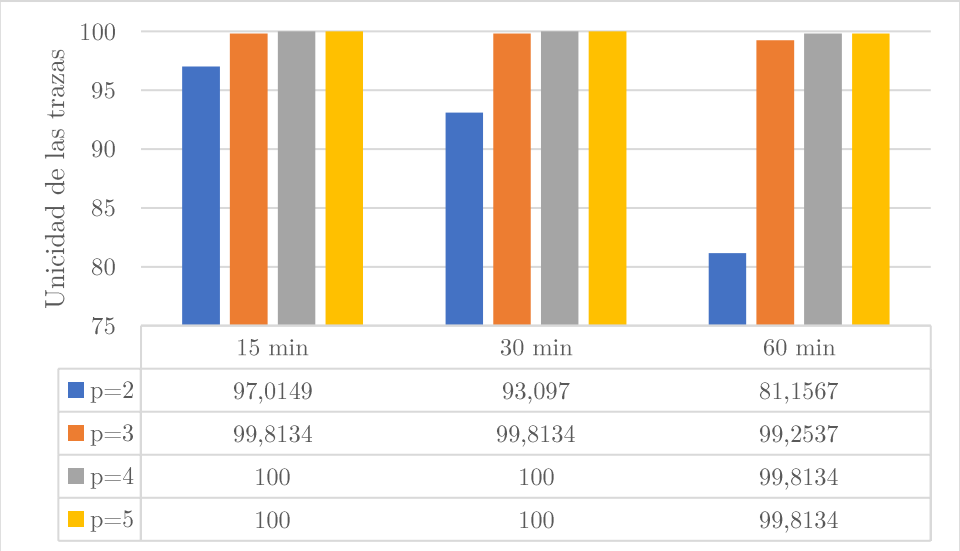


Al rotar 5º, el efecto de la rotación es todavía más obvio que al rotar 2º. Si revisamos de nuevo las trazas que en el dataset original corresponden al aeropuerto, al puente SF-Oakland y al aparcamiento de taxis, se aprecia de manera más clara que existe un desplazamiento mayor.



7.2.1.2.1 Unicidad

En la siguiente imagen, se muestra la unicidad ε de las trazas del dataset D_s calculada con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5.



	15 min	30 min	60 min
p=2	97,0149	93,097	81,1567
p=3	99,8134	99,8134	99,2537
p=4	100	100	99,8134
p=5	100	100	99,8134

Comparándola con la unicidad ε calculada para el dataset D, ε para D_s es igual o ligeramente más elevada en todos los casos.

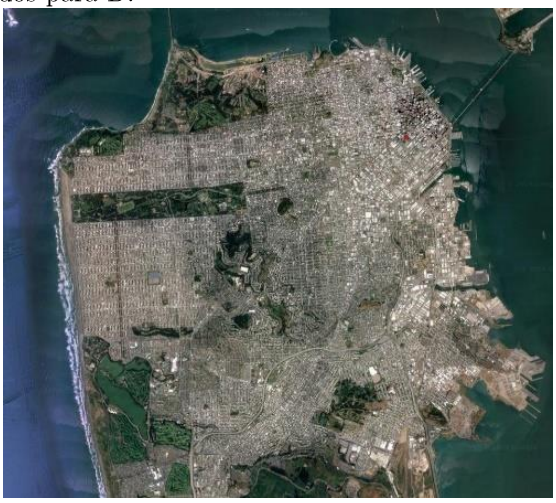
7.2.1.2.2 Ataques

7.2.1.2.2.1 Deducción de hogares

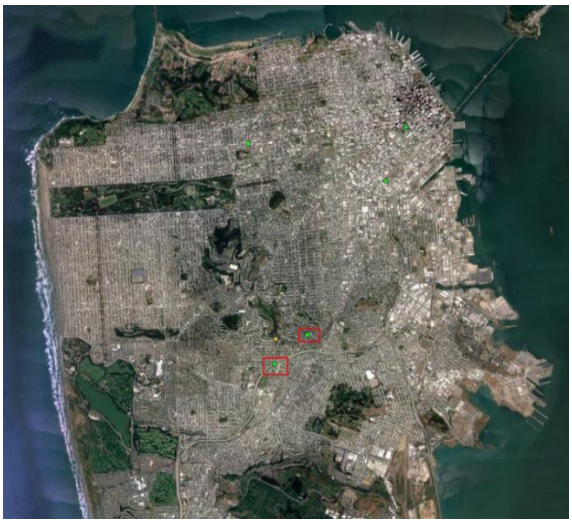
En la siguiente imagen se muestra la localización más repetida para cada uno de los individuos, es decir, el hogar deducido para cada uno, al aplicar el ataque de deducción de hogares sobre el dataset rotado en 5º.

De todos los hogares deducidos para el dataset rotado, solamente uno de ellos coincide con los hogares calculados para el dataset original D, aunque en este caso el hogar deducido sobre D_s se ha obtenido con menos trazas. En este caso, ninguna de las trazas que han provocado que el ataque considere esta localización un hogar del sujeto coincide con las trazas del dataset D. Por lo tanto, las trazas que ahora hacen de este punto un hogar corresponden a trazas que en D se localizaban en otro punto con precisión 0,001. En la zona en la que se encuentra este punto, la densidad de trazas era muy alta. Esto, sumado a que las trazas que formaban el hogar original pueden haber quedado repartidas en varios puntos de precisión 0,001 después de la rotación, ha

provocado esta singularidad. En la siguiente imagen se muestra el único hogar deducido de D_s que intersecta con los deducidos para D.



Si se tienen en cuenta las 5 localizaciones más repetidas para cada individuo, 6 de las localizaciones encontradas coinciden con las localizaciones más repetidas para los sujetos en D. En estos 6 casos se incluye el único hogar que coincide con los hogares deducidos sobre D, que hemos comentado previamente.



Existen dos tipos de situaciones distintas para la intersección entre los POIs correspondientes a las 5 localizaciones más repetidas en D y en D_s. Los dos puntos más cercanos al punto de origen de la rotación, que se indican mediante recuadros rojos en la imagen anterior, repiten como puntos de interés de sendos individuos debido a que, después de rotar 5º los puntos que formaban originalmente el POI en D, muchas de las trazas originales rotadas en D_s siguen manteniendose dentro del mismo recuadro de 0,001 grados. Concretamente 90 de 91 y 207 de 257 trazas originales de D se han mantenido en el mismo punto de precisión 0,001º en D_s. En el resto de puntos, ninguna de las trazas coincide con las originales debido a su mayor distancia con el punto de rotación. Simplemente, vuelve a suceder que son zonas con gran densidad de trazas, y al rotar los puntos, localizaciones que no encontraban en origen en ese lugar, lo han convertido de nuevo en uno de los POIs deducidos.

7.2.1.2.2.2 Begin-End

En la siguiente imagen se muestran parte de las localizaciones extraídas mediante heurística con el método de inicio-fin con un umbral de tiempo τ correspondiente a 8 (verde), 6 (rojo) y 4 horas (amarillo).

De nuevo, se han obtenido exactamente la misma cantidad de POIs y la misma cantidad de sujetos tienen al menos un punto de interés deducido para cada uno de los valores de τ que al aplicar el ataque sobre el dataset original D, y no se ha encontrado absolutamente ninguna intersección entre los puntos deducidos con el mismo valor de τ para D y para D_s. La razón es la misma que para la rotación de 2º, todos los puntos se han movido de su posición original.

7.2.1.2.2.3 Stays

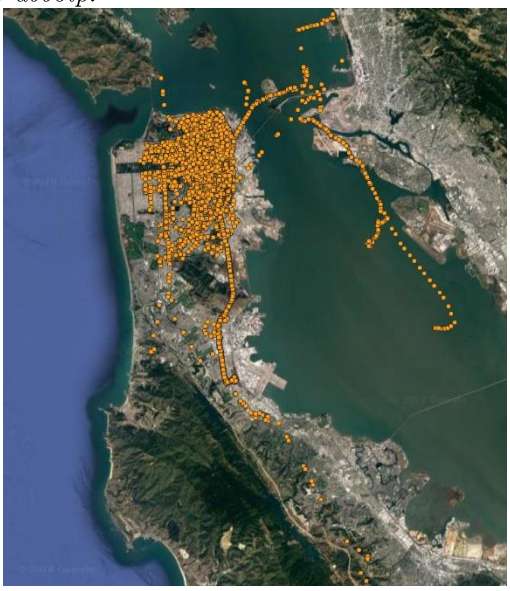
En la siguiente imagen se muestran todos los stays extraídos de D_s mediante la ejecución del algoritmo DT-Cluster usando los umbrales de tiempo τ a 5 minutos y de distancia χa 50 metros.

En este caso, se ha obtenido 105605 stays, 15 más que los obtenidos sobre el dataset original D. Teniendo en cuenta que las trazas solamente se han desplazado, esto se puede explicar de la misma forma que con D_s rotado en 2º. En este caso, varias trazas se encontraban en el límite de la distancia umbral respecto a otro punto, y al desplazarse los puntos, han quedado dentro del límite, llevando a obtener más stays. Se ha extraído al menos un stay para todos los sujetos de D_s, es decir 536, el mismo número que al extraer los stays de D. No se ha encontrado ninguna intersección entre los stays obtenidos al aplicar el ataque sobre D_s y sobre D. Al encontrarse todos los puntos de D_s desplazados respecto a los puntos de D, aunque la fecha de inicio y fin de los stays fuese la

misma, las localizaciones no coinciden debido a que, aunque se trate de una media entre todas las trazas que componen el stay, la precisión obtenida es de pocos metros.

7.2.1.3 Escala de latitud 1 y longitud 0,7

Al aplicar una escala de latitud 1 y longitud 0,7 respecto al punto (37,735085, -122,441601), se obtiene un dataset sanitizado D_s con todas las localizaciones del dataset original en una nueva posición (a excepción de aquellas trazas que se encuentren exactamente sobre la misma longitud que el punto de escala). En la siguiente imagen se pueden observar sobre el mapa todas las trazas de D_s para el sujeto con ID *abboip*.

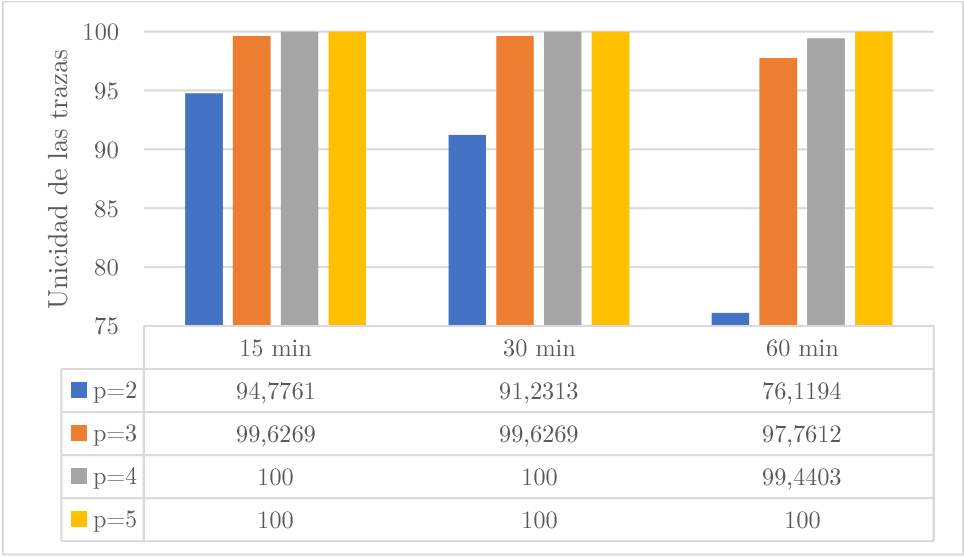


A simple vista ya se puede observar cómo ha afectado la escala a las trazas de este sujeto. La longitud de todas las trazas se ha modificado para acercarse al punto de origen -el valor es menor que 1-, mientras que la latitud se ha dejado inalterada (valor 1). Si revisamos de nuevo las trazas que en el dataset original corresponden al aeropuerto, al puente SF-Oakland y al aparcamiento de taxis, vemos el efecto de la aplicación de la escala sobre D.

Vemos que, al mantener la latitud constante y modificar las trazas para acercar su longitud al punto de origen, la proporción de las localizaciones que las trazas dibujan sobre el mapa se ve alterada. Esto es muy evidente en el aeropuerto o la parada de taxi, donde, aparte del desplazamiento hacia el punto de origen, el dibujo que las trazas crean sobre ellos se ve achatado.

7.2.1.3.1 Unicidad

En la siguiente imagen, se muestra la unicidad ε de las trazas del dataset D_s calculada con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5.



	15 min	30 min	60 min
p=2	94,7761	91,2313	76,1194
p=3	99,6269	99,6269	97,7612
p=4	100	100	99,4403
p=5	100	100	100

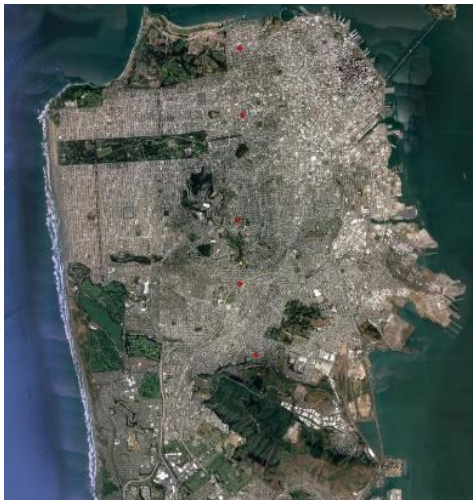
Para τ a 15 minutos, εes ligeramente más baja para p=2 en D, aunque para el resto de casos es muy parecida o mayor. Para τ a 30 minutos, εes muy parecida para D y para D_s. En el caso de τ a 60 minutos, εes 4 puntos más baja para p=2 en Ds respecto a D. En el resto de casos, ε es muy parecida, aunque ligeramente menor para p 3 y 4.

7.2.1.3.2 Ataques

7.2.1.3.2.1 Deducción de hogares

En la siguiente imagen se muestra la localización más repetida para cada uno de los individuos, es decir, el hogar deducido para cada uno, al aplicar el ataque de deducción de hogares sobre el dataset escalado D_s.

De todos los hogares deducidos para cada uno de los sujetos del dataset escalado D_s, ninguno coincide con ningún hogar calculado sobre D.
Si se tienen en cuenta las 5 localizaciones más repetidas para cada individuo, 6 de las localizaciones encontradas coinciden con las localizaciones más repetidas para los sujetos en D, que se pueden ver en la siguiente imagen (el punto amarillo representa el punto respecto al que se ha escalado). En el mapa aparecen solamente 5 puntos porque hay dos sujetos para los que el POI corresponde al mismo lugar.



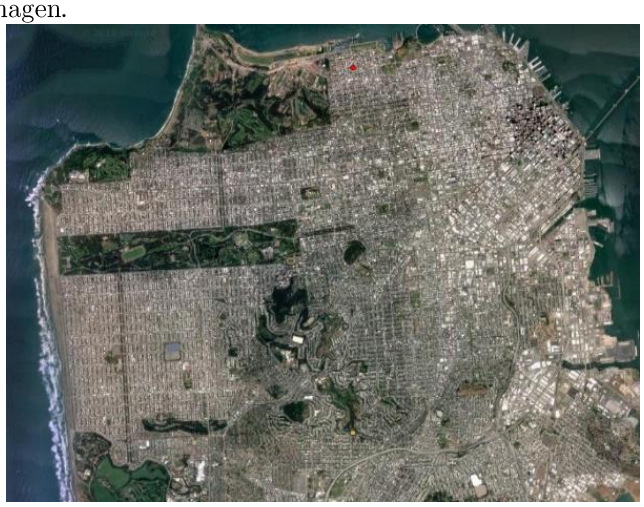
Si nos fijamos en la imagen previa, vemos rápidamente una peculiaridad de todos los POIs que intersectan con los deducidos a partir del dataset original: se encuentran prácticamente en la misma longitud que el punto respecto al que se ha escalado. Como que la longitud de las trazas que provocan que estas localizaciones sean escogidas como puntos de interés es tan parecida a la longitud del punto de origen, se mantienen prácticamente inalteradas cuando se tiene en cuenta con precisión 0,001 grados. El punto más cercano por el sur al origen del escalado tiene la peculiaridad de que las trazas que lo convierten en uno de los 5 puntos más visitados para ese sujeto son exactamente las mismas en D que en D_s, aunque todas se hayan escalado. Cuatro del resto de casos conservan entre el 95% y el 100% de las trazas de D, y el caso restante conserva el 60% de las trazas. Precisamente el punto que menos trazas conserva en D_s del dataset original, es el que obtiene más trazas nuevas que lo convierten en el hogar calculado (más repeticiones) para el individuo. Esto se debe a que, al estar prácticamente sobre la longitud del punto de origen de

la escala, si la densidad de puntos en D con la misma latitud y a poca distancia en longitud es alta, en D_s estas localizaciones se "comprimen" sobre la longitud del punto de origen, haciendo que más trazas entren dentro del punto con precisión 0,001º.

7.2.1.3.2.2 Begin-End

En la siguiente imagen se muestran parte de las localizaciones extraídas mediante heurística con el método de inicio-fin con un umbral de tiempo τ correspondiente a 8 (verde), 6 (morado) y 4 horas (naranja).

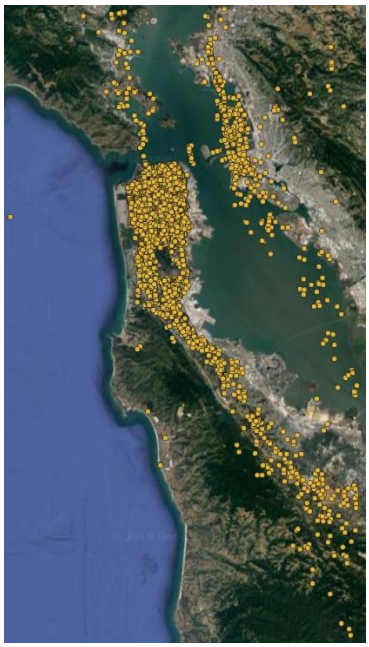
De nuevo, se han obtenido exactamente la misma cantidad de POIs y la misma cantidad de sujetos tienen al menos un punto de interés deducido para cada uno de los valores de τ que al aplicar el ataque sobre el dataset original D. A diferencia de los anteriores casos, se ha encontrado una intersección entre los puntos deducidos con el mismo valor de τ para D y para D_s, que se puede ver en la siguiente imagen.



Se trata de un caso excepcional, ya que, como se puede ver en la imagen, la longitud del POI extraído por el ataque parece la misma que la del punto sobre el que se ha escalado (en amarillo). Al analizarlo con más detalle vemos que, pese a no ser idéntica, la longitud del punto es de -122,44159, mientras que la del punto de origen es -122,441601. La diferencia es tan pequeña que el escalado no ha modificado su precisión 0,00001 grados.

7.2.1.3.2.3 Stays

En la siguiente imagen se muestran todos los stays extraídos de D_s mediante la ejecución del algoritmo DT-Cluster usando los umbrales de tiempo τ a 5 minutos y de distancia χa 50 metros.



Se han obtenido 108232, es decir, un 2,5% más de stays que para el dataset D. En este tipo de perturbación, al *comprimirse* todos los puntos respecto al punto de origen de la escala, eso provoca que muchas de las trazas pasen a estar más cerca unas de otras. Al estar más cerca entre ellas, una gran cantidad de trazas (se han obtenido 2642 stays más que en D) que se encontraban dentro del umbral de tiempo τ de 5 minutos, pero se encontraban más lejos del umbral χde 50 metros, han pasado a encontrarse también dentro del umbral debido a la escala. Se ha extraído al menos un stay para todos los sujetos de D_s, es decir 536, el mismo número que al extraer los stays de D. En este caso, se ha encontrado una intersección de 10 stays entre los obtenidos sobre D y sobre D_s, lo cual representa aproximadamente un 0,01% de intersección con los stays de D. En la siguiente imagen se muestran las 10 localizaciones de los stays que intersectan en amarillo, y el punto de origen del escalado en rosa.

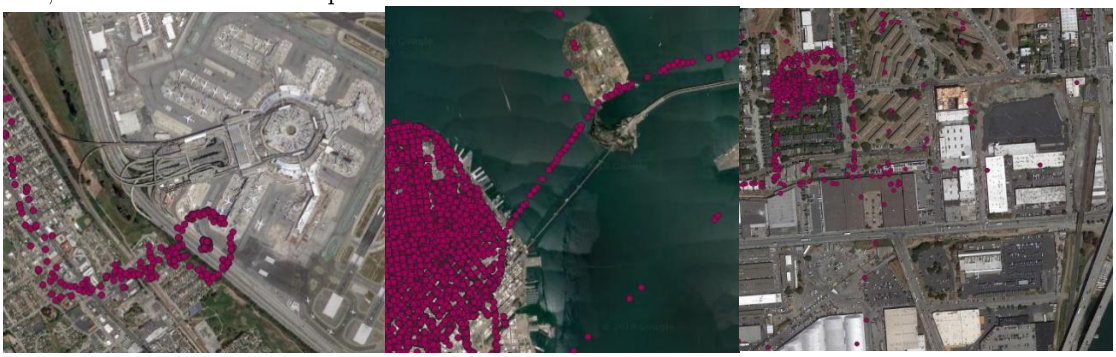
Salta a la vista que todos los puntos se encuentran prácticamente en la misma longitud que el punto de origen. Es por eso que, al estar tan cerca en la longitud, y al haberse transformado solamente la longitud y mantener la latitud constante respecto a D, las trazas de D_s que provocan que se conviertan en stays se encuentran exactamente en el mismo sitio, no se han desplazado.

7.2.1.4 Escala de latitud 1,05 y longitud 0,9

Al aplicar una escala de latitud 1,05 y longitud 0,9 respecto al punto (37,735085, -122,441601), se obtiene un dataset sanitizado D_s con todas las localizaciones del dataset original en una nueva

posición (a excepción de aquellas trazas que se encuentren exactamente sobre el punto de escala). En la siguiente imagen se pueden observar sobre el mapa todas las trazas de D_s para el sujeto con ID *abboip*.

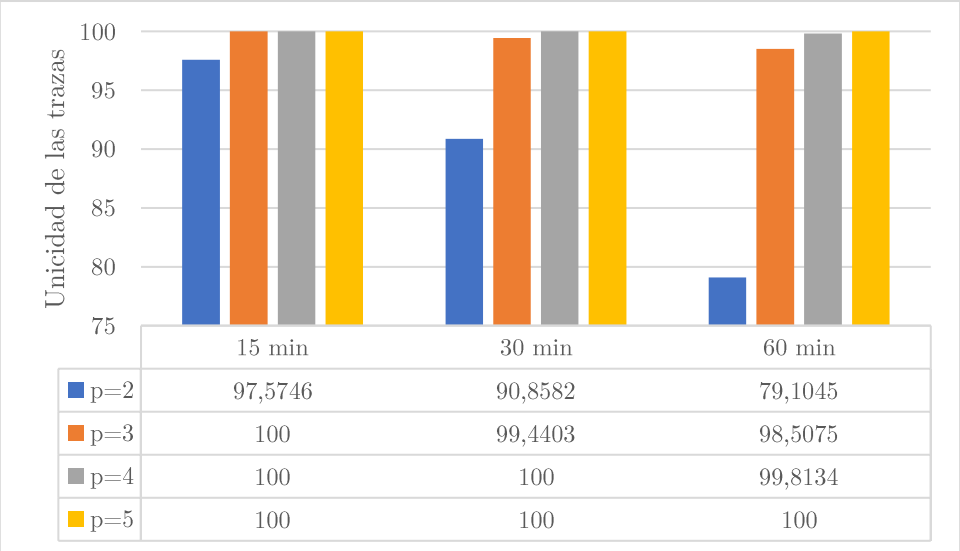
En este caso, el desplazamiento de las trazas es más sutil, debido a que los valores usados para la escala son más cercanos a 1. La longitud de todas las trazas se ha modificado para acercarse al punto de origen -el valor es menor que 1-, pero la latitud de las trazas se ha modificado para alejarse del punto de origen, ya que el valor es superior a 1. Si revisamos de nuevo las trazas que en el dataset original corresponden al aeropuerto, al puente SF-Oakland y al aparcamiento de taxis, vemos el efecto de la aplicación de la escala sobre D.



Vemos que dependiendo de dónde se encuentren los puntos respecto a la latitud del punto de origen, se sitúan más hacia el norte o más hacia el sur de su posición original en D. Se puede ver en el aeropuerto y el aparcamiento de taxis que la proporción no se mantiene y que el dibujo que crean las trazas sobre ellos está achatado.

7.2.1.4.1 Unicidad

En la siguiente imagen, se muestra la unicidad ε de las trazas del dataset D_s calculada con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5.



	15 min	30 min	60 min
p=2	97,5746	90,8582	79,1045
p=3	100	99,4403	98,5075
p=4	100	100	99,8134
p=5	100	100	100

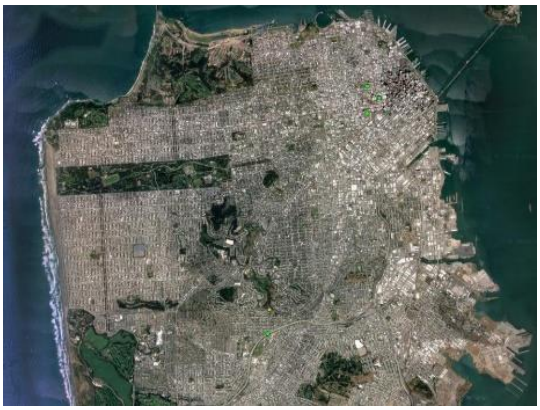
Para τ a 15 y 30 minutos, εes ligeramente más elevada para D que para D_s en todos los casos. Para τ a 60 minutos, εestá 1 punto por debajo en D_s respecto a D para p=2, aunque para el resto es prácticamente igual.

7.2.1.4.2 Ataques

7.2.1.4.2.1 Deducción de hogares

En la siguiente imagen se muestra la localización más repetida para cada uno de los individuos, es decir, el hogar deducido para cada uno, al aplicar el ataque de deducción de hogares sobre el dataset escalado D_s.

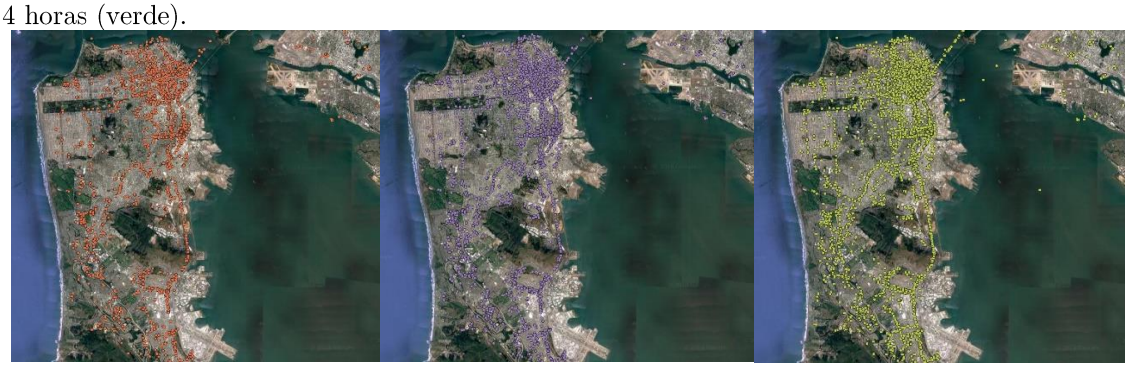
De todos los hogares deducidos para cada uno de los sujetos del dataset escalado D_s, ninguno coincide con ningún hogar calculado sobre D.
Si se tienen en cuenta las 5 localizaciones más repetidas para cada individuo, 7 de las localizaciones encontradas coinciden con las localizaciones más repetidas para los sujetos en D, que se pueden ver en la siguiente imagen (el punto amarillo representa el punto respecto al que se ha escalado). En el mapa aparecen solamente 4 puntos porque hay 3 de los POIs que comparten dos sujetos en cada caso.



El punto más al sur del punto de origen del escalado se encuentra muy cerca de éste, y sobre una longitud muy parecida. Esto provoca que ninguna de las trazas que hacen que sea considerado un POI en D se hayan desplazado del cuadrado de aproximadamente 111 metros (0,001 grados) en el que se encuentran al ser escaladas en D_s. El resto de puntos no conservan en D_s las trazas que los convertían en un POI en D. Ha sucedido lo mismo que ya hemos visto en casos anteriores, al tratarse de zonas con mucha densidad de trazas, las trazas escaladas han pasado a ser POIs de los mismos sujetos en la misma localización en la que en D otras trazas lo convertían en punto de interés.

7.2.1.4.2.2 Begin-End

En la siguiente imagen se muestran parte de las localizaciones extraídas mediante heurística con el método de inicio-fin con un umbral de tiempo τ correspondiente a 8 (naranja), 6 (morado) y 4 horas (verde).



De nuevo, se han obtenido exactamente la misma cantidad de POIs y la misma cantidad de sujetos tienen al menos un punto de interés deducido para cada uno de los valores de τ que al aplicar el ataque sobre el dataset original D. No se ha encontrado ninguna intersección entre los puntos deducidos con el mismo valor de τ para D y para Ds, debido a que todos los puntos se han movido de su posición original.

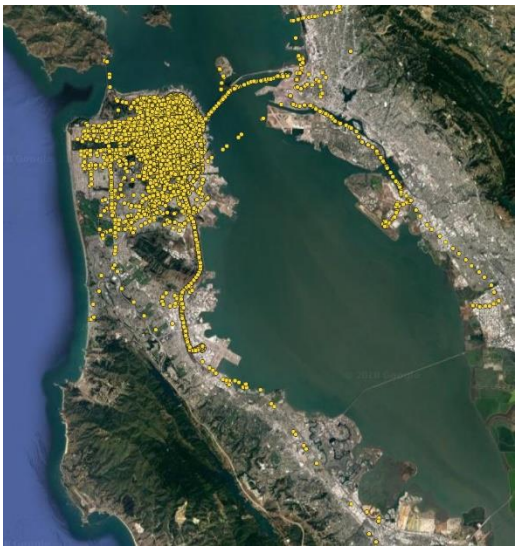
7.2.1.4.2.3 Stays

En la siguiente imagen se muestran todos los stays extraídos de D_s mediante la ejecución del algoritmo DT-Cluster usando los umbrales de tiempo τ a 5 minutos y de distancia χa 50 metros.

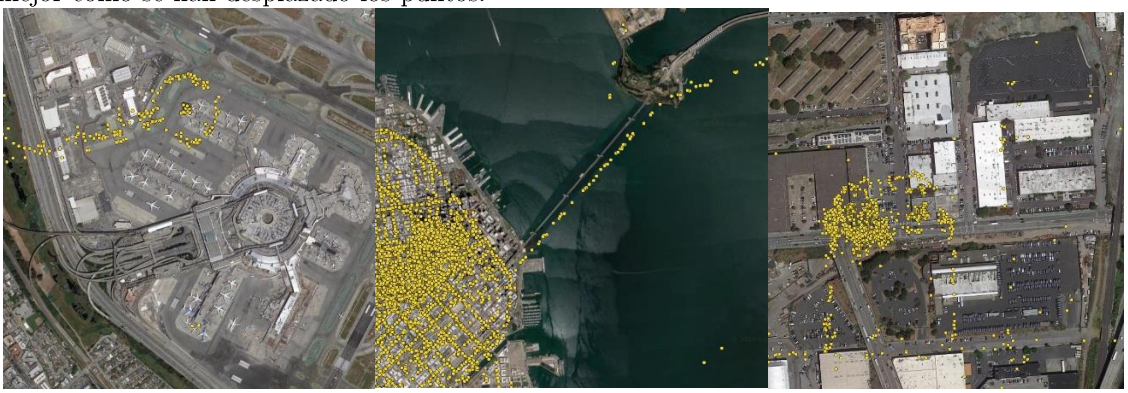
Se han obtenido 105979 stays en este caso, aproximadamente un 0,4% más que para el dataset D. En este caso, los puntos se han acercado al origen de la escala en longitud, pero se han alejado del origen en latitud, aunque la escala ha sido muy suave en ambos sentidos. La compresión en longitud de los puntos mezclada con el alejamiento en latitud ha provocado la extracción de 389 nuevos stays, muchos menos que en el experimento en el que solamente se escalaba la longitud. Se ha extraído al menos un stay para todos los sujetos de D_s, es decir 536, el mismo número que al extraer los stays de D. En este caso, no se ha encontrado ninguna intersección entre los stays obtenidos de D y de D_s. Al escalar tanto la latitud como la longitud, los puntos se han desplazado lo suficiente en ambas magnitudes para que no coincida ningún stay.

7.2.1.5 Composición de perturbaciones: Escala de 0,95 de latitud y 0,95 de longitud + Rotación de 1º

Se ha obtenido un dataset sanitizado D_s realizando una composición de perturbaciones. Se aplican las perturbaciones para escalar los puntos del dataset original D en 0,95 de latitud y 0,95 de longitud, y rotar los puntos en 1º. El orden en el que se aplican las transformaciones es relevante y es el que se acaba de indicar. El punto sobre el que se han aplicado las dos transformaciones es el definido para todas las perturbaciones, (37,735085, -122,441601). En la siguiente imagen se pueden observar sobre el mapa todas las trazas de D_s para el sujeto con ID *abboip*.

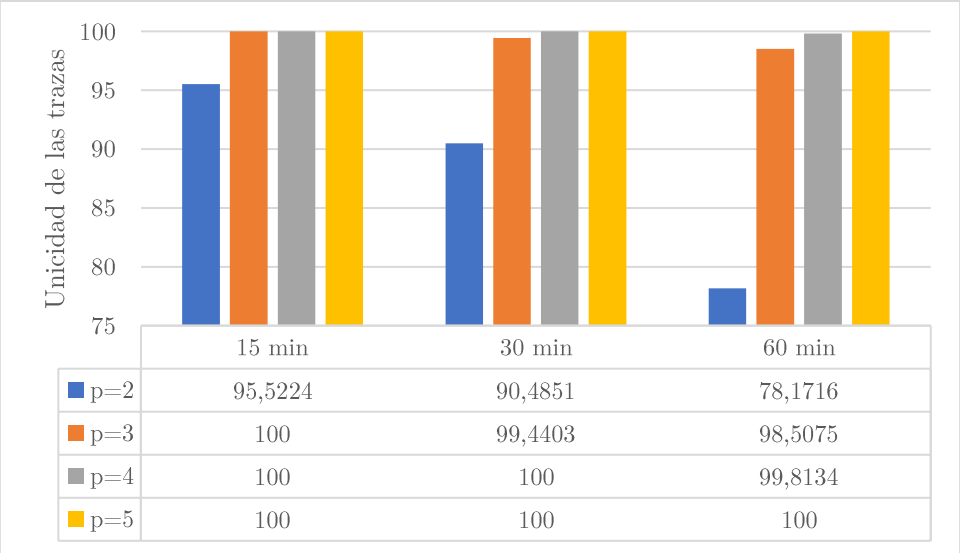


Al aplicar una escala tan sutil y una rotación de tan sólo 1 grado, el efecto de la composición de transformaciones es, probablemente, la menos evidente de las que hemos evaluado. Pese a todo, debido a las características geográficas de San Francisco, el desplazamiento de los puntos se hace muy evidente en zonas como los puentes. Si revisamos de nuevo las trazas que en el dataset original corresponden al aeropuerto, al puente SF-Oakland y al aparcamiento de taxis, se aprecia mejor cómo se han desplazado los puntos.



7.2.1.5.1 Unicidad

En la siguiente imagen, se muestra la unicidad ε de las trazas del dataset D_s calculada con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5.



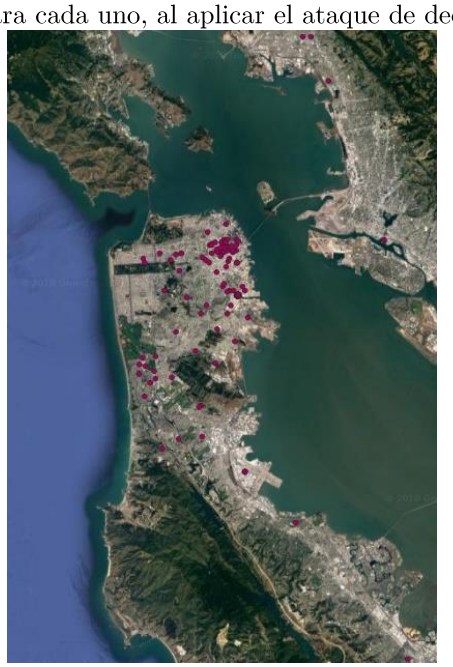
	15 min	30 min	60 min
p=2	95,5224	90,4851	78,1716
p=3	100	99,4403	98,5075
p=4	100	100	99,8134
p=5	100	100	100

Para τ a 15, εes ligeramente más baja para D que para D_s cuando p=2, pero en el resto de casos es muy similar o más elevada. Para τ a 30 minutos, la unicidad calculada es un punto más baja para D_s cuando p=2, pero es mayor cuando p>2. Para τ a 60 minutos, εestá 2 puntos por debajo en D_s respecto a D para p=2, aunque para el resto es prácticamente igual.

7.2.1.5.2 Ataques

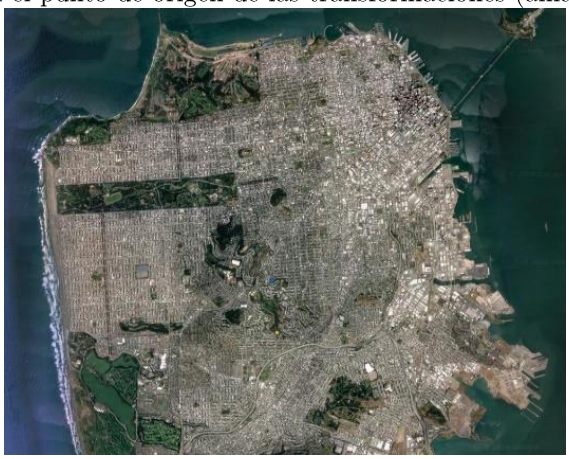
7.2.1.5.2.1 Deducción de hogares

En la siguiente imagen se muestra la localización más repetida para cada uno de los individuos, es decir, el hogar deducido para cada uno, al aplicar el ataque de deducción de hogares sobre D_s.



De todos los hogares deducidos para el dataset D_s, solamente uno de ellos coincide con los hogares calculados para el dataset original D. El 99% de las trazas que convierten en hogar de un sujeto esta localización en D se corresponden con las trazas modificadas de D_s que lo convierten en el más votado para este sujeto. Como en casos que hemos visto previamente, la localización cercana del POI respecto al punto sobre el que se han realizado las transformaciones es la causante de esta situación, en la que solamente dos de las trazas de D ha cambiado de recuadro de 0,001 grados.

En la siguiente imagen se muestra el único hogar deducido de Ds que intersecta con los deducidos para D (azul), junto con el punto de origen de las transformaciones (amarillo).



Si se tienen en cuenta las 5 localizaciones más repetidas para cada individuo, 7 de las localizaciones encontradas coinciden con las localizaciones más repetidas para los sujetos en D. En estos 6 casos se incluye el único hogar que coincide con los hogares deducidos sobre D, que hemos comentado previamente. Las 7 intersecciones se pueden ver en la siguiente imagen, junto al punto de origen de las transformaciones en amarillo.

El punto más cercano al origen de las transformaciones (por el este), mantiene en D_s exactamente las mismas trazas que en D le convierten en uno de los 5 puntos más visitados por un sujeto. De nuevo, la proximidad al punto de origen provoca que las trazas se desplacen tan poco que se mantengan en el mismo punto de precisión 0,001. Además, en este caso, el ángulo de rotación tan bajo sumado con la proximidad con el origen, provocan que ésta afecte muy poco a estas trazas. El tercer punto más cercano al punto de origen, siendo los dos primeros el que acabamos de comentar y el único hogar que intersecta con los obtenidos en D, tiene más del 90% de las trazas que lo convierten en POI tanto en D como en D_s. Cuanto más nos alejamos del punto de origen, menos trazas de las que convierten las localizaciones en puntos de interés coinciden entre los datasets original y D_s. De hecho, los 4 puntos restantes no conservan ninguna traza del dataset original, lo que indica que en D existían en esas zonas grandes densidades de trazas que al trasladarse han pasado a convertir la localización en punto de interés en lugar de otras que en D no se encuentran en el mismo recuadro de aproximadamente 111 metros.

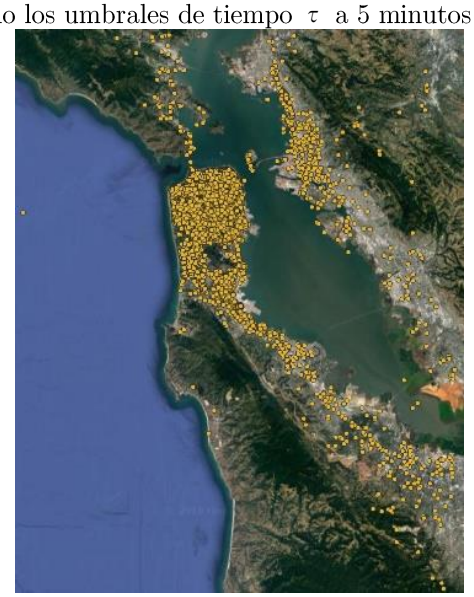
7.2.1.5.2.2 Begin-End

En la siguiente imagen se muestran parte de las localizaciones extraídas mediante heurística con el método de inicio-fin con un umbral de tiempo τ correspondiente a 8 (morado), 6 (rosa) y 4 horas (naranja).

De nuevo, se han obtenido exactamente la misma cantidad de POIs y la misma cantidad de sujetos tienen al menos un punto de interés deducido para cada uno de los valores de τ que al aplicar el ataque sobre el dataset original D, y no se ha encontrado ninguna intersección entre los puntos deducidos con el mismo valor de τ para D y para D_s.

7.2.1.5.2.3 Stays

En la siguiente imagen se muestran todos los stays extraídos de D_s mediante la ejecución del algoritmo DT-Cluster usando los umbrales de tiempo τ a 5 minutos y de distancia χa 50 metros.

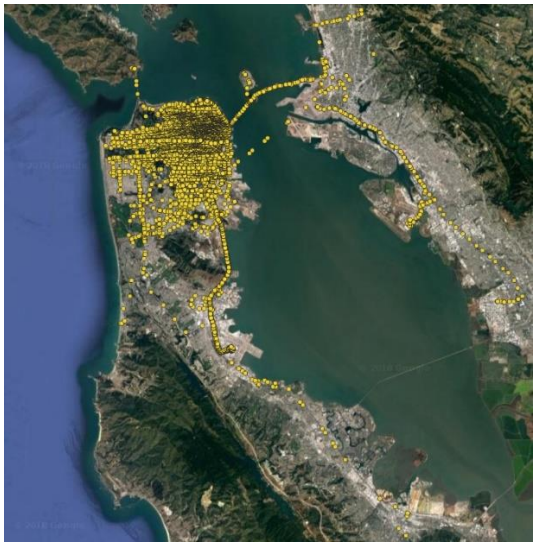


Se han obtenido 106593 stays en este caso, aproximadamente un 0,95% más que para el dataset D. En este caso, la escala se ha realizado de forma uniforme entre latitud y longitud usando un valor menor que 1 pero muy cercano a éste para que el desplazamiento no fuese muy evidente. Eso provoca que todos los puntos del dataset se acerquen y se compriman hacia el punto de origen de una forma uniforme, provocando que una gran cantidad de trazas (se han extraído 1003 stays más que para D) que se encontraban dentro del umbral τ de tiempo pero no dentro del umbral de distancia χen D, hayan pasado en D_s a encontrarse dentro de los dos umbrales. La rotación aplicada a todos los puntos después de la escala es de tan sólo 1º, por lo que, los puntos cercanos al origen de la rotación se han desplazado muy poco, permitiendo que las trazas que han provocado la extracción de los nuevos stays se mantengan dentro de los umbrales. Se ha extraído al menos un stay para todos los sujetos de D_s, es decir 536, el mismo número que al extraer los stays de D. En este caso, no se ha encontrado ninguna intersección entre los stays obtenidos de D y de D_s. Al

escalar y luego rotar todos los puntos del dataset D, los puntos del dataset resultante D_s se han desplazado lo suficiente para que no coincida ningún stay.

7.2.2 Agregación

A continuación, evaluaremos la micro-agregación de puntos mediante el estudio de la unicidad y del efecto de los ataques sobre el dataset sanitizado D_s obtenido de aplicar la micro-agregación al dataset original D.
En la siguiente imagen se pueden observar sobre el mapa todas las trazas de D_s para el sujeto con ID *abboip*, en amarillo.



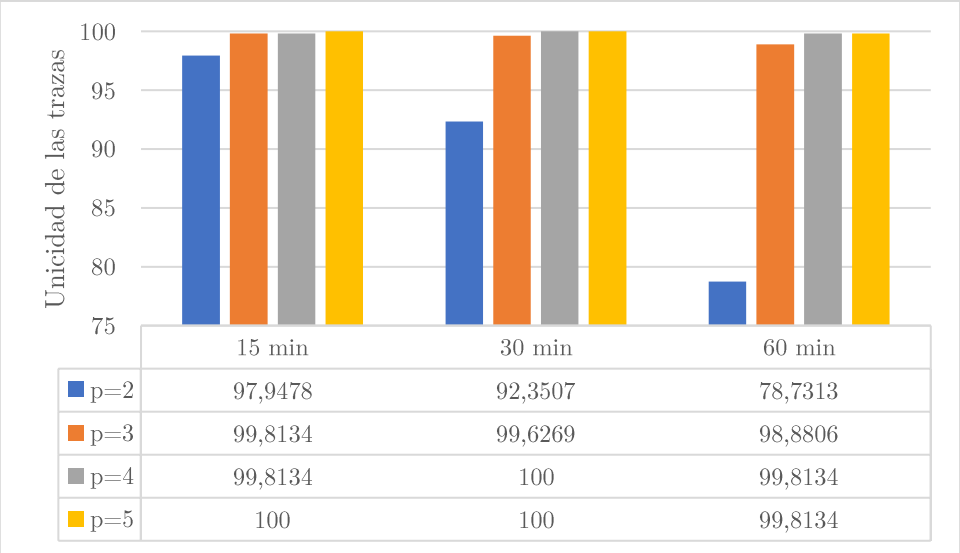
La aplicación de la micro-agregación de puntos crea un efecto cuadrícula que se puede ver en la imagen anterior. Esto se debe a que para cada recuadro de 0,001 grados (aproximadamente 111 metros), todos los puntos que se encuentran dentro del recuadro han visto modificadas sus coordenadas para situarse en el punto medio de latitud y longitud de todos los puntos del recuadro. Así pues, vemos sobre el mapa que para cada recuadro solamente tenemos una traza, creando el efecto de cuadrícula. Hay que resaltar que, aunque solamente aparezca un punto superpuesto sobre el mapa, cada punto puede contener una o varias trazas, pero al encontrarse todas en el mismo punto exacto, solamente vemos el punto que se encuentre arriba del todo. En la siguiente imagen se muestran los detalles del aeropuerto de San Francisco, de la zona del noreste de San Francisco y del aparcamiento de la empresa de taxis.

La zona noreste de San Francisco es una zona con una gran densidad de trazas debido a que se encuentra el distrito financiero y zonas culturales y de ocio. Por eso el efecto de la micro-agregación es muy evidente en esta zona. El parking de la empresa de taxis, el cual previamente hemos visto

que tiene una gran densidad de trazas repartidas por toda la zona, muestra claramente la cuadrícula compuesta por recuadros de 0,001 grados, en los que el punto en el que se agrega no es el centro sino la media de todas las trazas del recuadro.

7.2.2.1 Unicidad

En la siguiente imagen, se muestra la unicidad ε de las trazas del dataset D_s calculada con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5.



	15 min	30 min	60 min
p=2	97,9478	92,3507	78,7313
p=3	99,8134	99,6269	98,8806
p=4	99,8134	100	99,8134
p=5	100	100	99,8134

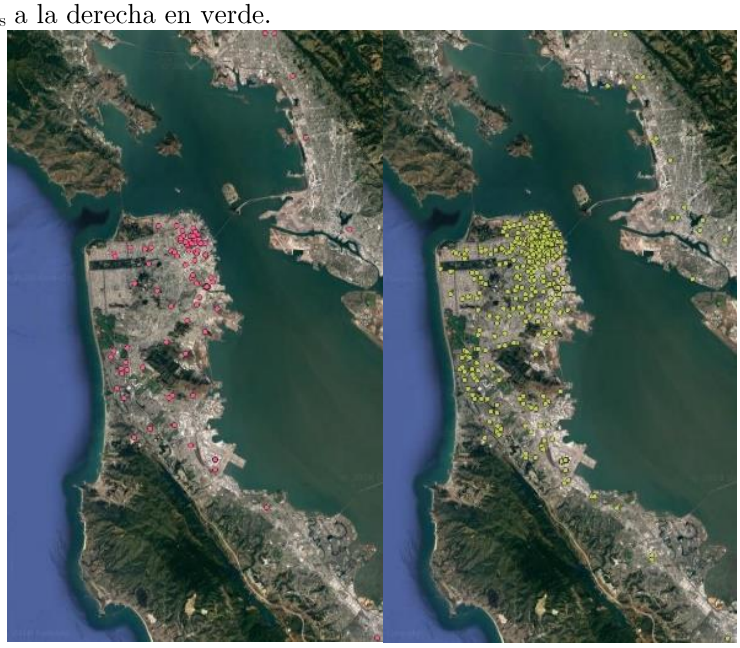
Para τ a 15, εestá un punto por encima en D_s cuando p=2, pero en el resto de casos es la misma. Para τ a 30 minutos, εes en general un poco más elevada que en D. Para τ a 60 minutos, la única diferencia entre D_s y D es que para p=2, εestá 2 puntos por debajo.

7.2.2.2 Ataques

7.2.2.2.1 Deducción de hogares

En la siguiente imagen se muestra la localización más repetida para cada uno de los individuos, es decir, el hogar deducido para cada uno, al aplicar el ataque de deducción de hogares sobre D_s.

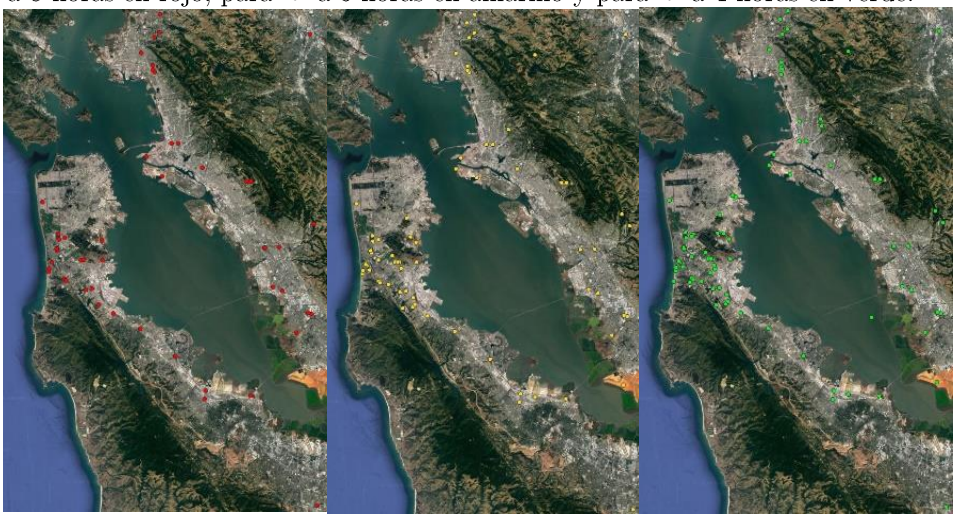
Todos y cada uno de los hogares deducidos para D_s coinciden con los hogares calculados para el dataset original D. Esto sucede porque tanto el algoritmo de sanitización aplicado en este caso -la micro-agregación- como el ataque de deducción de hogares se basan, en este trabajo, en agrupar trazas repartiendo el espacio de la misma manera: en cuadrados de 0,001 grados (aproximadamente 111 metros). De esta forma, todas las trazas de D, aunque se modifique su localización, siguen manteniéndose en el mismo recuadro. Desde el punto de vista del ataque, que cuenta las trazas que se encuentran en cada recuadro de 0,001 grados, las trazas se mantienen exactamente igual en D que en D_s, de ahí que la intersección sea total.
Si se tienen en cuenta las 5 localizaciones más repetidas para cada individuo, más del 99,9% de las localizaciones encontradas en D_s coinciden con las localizaciones más repetidas para los sujetos en D. Solamente dos POIs de dos sujetos distintos -de los 2680 puntos obtenidos- no coinciden. Esto se debe a que, en ambos casos, el quinto punto más repetido coincide con otra localización en número de repeticiones. El algoritmo implementado no ordena de una manera específica aquellos puntos con el mismo número de repeticiones, por lo que parece que en este caso la base de datos ha devuelto los puntos en un orden distinto al que se obtuvo para el dataset original. En la siguiente imagen se puede ver la intersección entre los hogares deducidos para D y para D_s a la izquierda en rosa, y la intersección entre los 5 puntos más repetidos para cada sujeto deducidos para D y para D_s a la derecha en verde.



7.2.2.2.2 Begin-End

En la siguiente imagen se muestran parte de las localizaciones extraídas mediante heurística con el método de inicio-fin con un umbral de tiempo τ correspondiente a 8 (verde), 6 (naranja) y 4 horas (amarillo).

De nuevo, se han obtenido exactamente la misma cantidad de POIs y la misma cantidad de sujetos tienen al menos un punto de interés deducido para cada uno de los valores de τ que al aplicar el ataque sobre el dataset original D. Se han encontrado intersecciones entre los POIs extraídos de D y de D_s para todos los valores de τ. De todos los puntos de interés extraídos por el ataque sobre D_s, un 1,34% para τ a 8 horas, un 1,08% para τ a 6 horas y un 0,85% para τ a 4 horas coinciden con los extraídos de D. En la siguiente imagen se pueden ver los puntos que intersectan para τ a 8 horas en rojo, para τ a 6 horas en amarillo y para τ a 4 horas en verde.

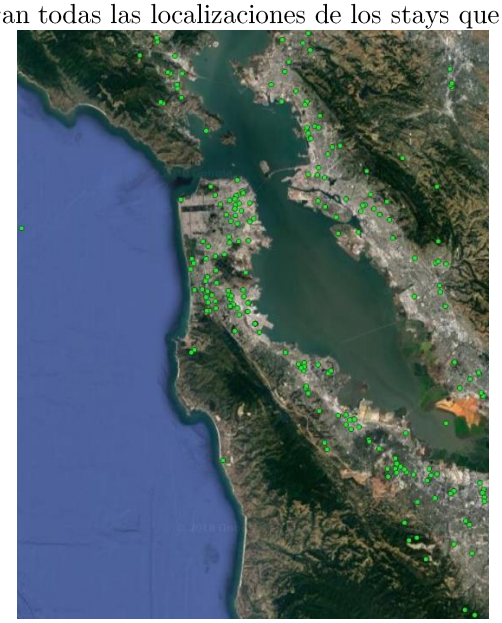


El ataque de begin-end aplicado a un dataset sanitizado mediante la micro-agregación de puntos extrae POIs que coinciden con los extraídos sobre D principalmente en zonas con poca densidad de trazas. Para los recuadros de 0,001 grados a los que solamente corresponda una traza, al realizar la micro-agregación se realizará la media de una sola traza y se quedará en el mismo punto exacto en el que se encuentra. En cuadrados a los que corresponda más de una traza, puede suceder que la localización de una de las trazas se encuentre exactamente en el punto medio de latitud y longitud entre todas las trazas del recuadro de 0,001 grados. En cualquier caso, observando la imagen anterior, se ve a simple vista que, en la zona de más concentración de trazas, en el noreste de San Francisco, no existe ninguna intersección en los POIs. Lo mismo sucede con el aeropuerto de San Francisco.

7.2.2.2.3 Stays

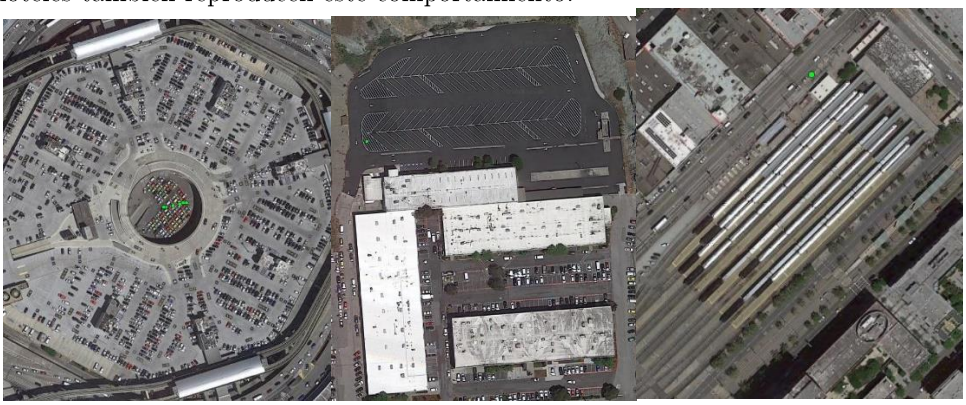
En la siguiente imagen se muestran todos los stays extraídos de D_s mediante la ejecución del algoritmo DT-Cluster usando los umbrales de tiempo τ a 5 minutos y de distancia χa 50 metros a la izquierda, y a la derecha se muestra un detalle de la zona noreste de San Francisco, en el que se aprecia el efecto malla de los stays obtenidos.

Se han obtenido 101801 stays en este caso, aproximadamente un 3,59% menos que para el dataset D. Este método de sanitización provoca que cada una de las trazas que comparten una precisión de 0,001 grados (unos 111 metros) vean trasladada su localización a la media de latitud y longitud de todas ellas. Este hecho provoca que en la nueva localización de D_s, aquellas trazas que se encontraban en zonas limítrofes del umbral de distancia χde 50 metros en D, pasen a encontrarse fuera del umbral. Es una cantidad considerable de información que el ataque ha perdido al aplicarse sobre D_s respecto a aplicarse sobre D (3789 stays menos, más de un 3,5% menos, como hemos visto previamente). En cualquier caso, en este caso también se ha extraído al menos un stay para todos los sujetos de D_s, es decir 536, el mismo número que al extraer los stays de D.
En este caso, se ha encontrado una intersección de 235 stays entre los obtenidos sobre D y sobre Ds, lo cual representa aproximadamente un 0,22% de intersección con los stays de D. En la siguiente imagen se muestran todas las localizaciones de los stays que intersectan entre D y D_s.



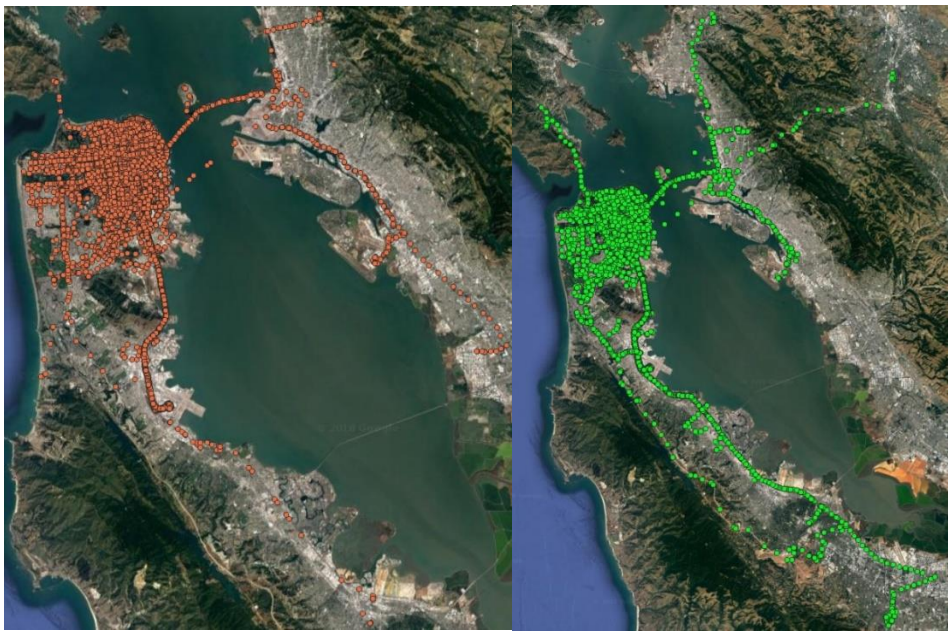
Tal como se puede ver sobre el mapa, una gran cantidad de los stays que coinciden se encuentran alejados del centro de San Francisco, algunos en zonas muy alejadas. Simplemente, por cómo se aplica la micro-agregación de puntos, aquellas zonas de 0,001º con muy pocas trazas tienen muchas más probabilidades de que las trazas se queden en el mismo punto exacto al hacer la media entre todos los puntos. Si, por ejemplo, en un cuadrado de 111 metros solamente se encuentra una traza, esta se quedará en el mismo lugar. Si dos zonas contiguas de precisión 0,001 grados (recordemos que el umbral χes de 50 metros), contienen cada una de ellas muy pocas trazas, es muy probable que las trazas que forman un stay se encuentren en el mismo lugar que cuando el stay se extrajo para D. De manera opuesta a las zonas con muy poca densidad de trazas, las zonas en las que se

concentra una gran cantidad de trazas en un espacio muy reducido, y por tanto se concentran una gran cantidad de stays, provocan que los stays calculados también puedan coincidir entre el dataset original y el sanitizado, debido a que es probable que, al haber tantos stays en poco espacio, alguno coincida. Este último comportamiento queda patente en zonas donde se concentran gran cantidad de stays como el aeropuerto, el aparcamiento de taxis o la estación de tren, como se puede ver en la siguiente imagen. Otras zonas como aquellas donde los taxis esperan clientes en los hoteles también reproducen este comportamiento.



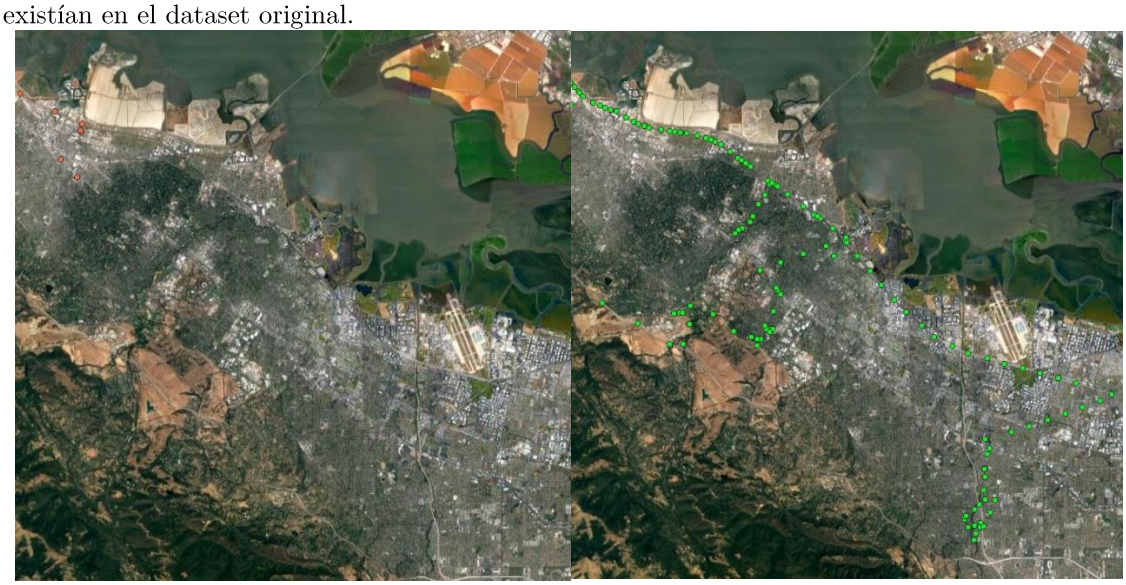
7.2.3 Swapping

A continuación, se evalúa el swapping, un método de sanitización que, como hemos visto previamente en este trabajo, a grandes rasgos consiste en intercambiar las trazas de dos individuos del dataset cuando éstas se encuentran lo suficientemente cerca en un intervalo de tiempo lo suficientemente corto. El umbral de distancia se denota como χ y el de tiempo como τ.
En este trabajo, se ha sanitizado el dataset D aplicando el método de swapping con un umbral de tiempo τ a 10 segundos y χ a 20 metros. Hay que tener en cuenta que el sampling del dataset CabSpotting es de una traza cada 1 segundo aproximadamente y 0,00001 grados decimales de precisión, lo que corresponde aproximadamente a 1,11 metros. Pese a todo, no todos los individuos envían continuamente su localización, por lo que las trazas de algunos sujetos se ven reducidas respecto al mismo tiempo para el resto de individuos. Al aplicar el swapping al dataset original D se obtiene un dataset sanitizado D_s. En la siguiente imagen, a la izquierda en naranja, se pueden observar sobre el mapa todas las trazas de D para el sujeto con ID *abboip*, mientras que, a la derecha en verde, se pueden observar sobre el mapa todas las trazas de D_s para el sujeto con ID *abboip*.

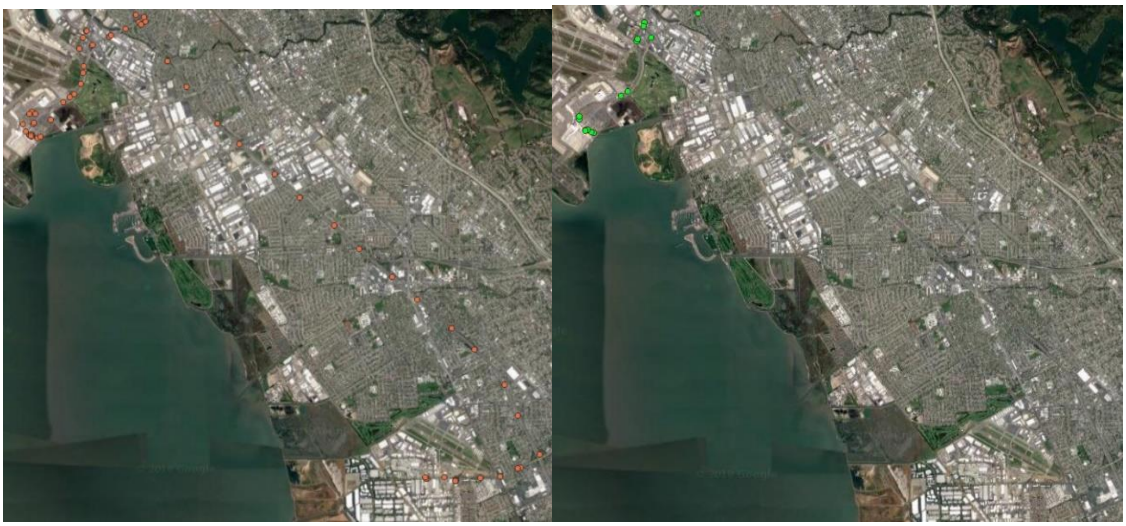


Con este individuo se puede ilustrar claramente cómo sus trazas se han intercambiado con las de otros individuos, creando trayectorias que no aparecían en D, y que por tanto el individuo *abboip* nunca realizó en realidad.

En las zonas con menos densidad de trazas se hace muy evidente el efecto del swapping de trazas. En la siguiente imagen se puede ver la diferencia entre las trazas de D, a la izquierda, y de D_s, a la derecha, al sur de San Francisco. En este caso, aparecen trazas en el dataset sanitizado que no existían en el dataset original.



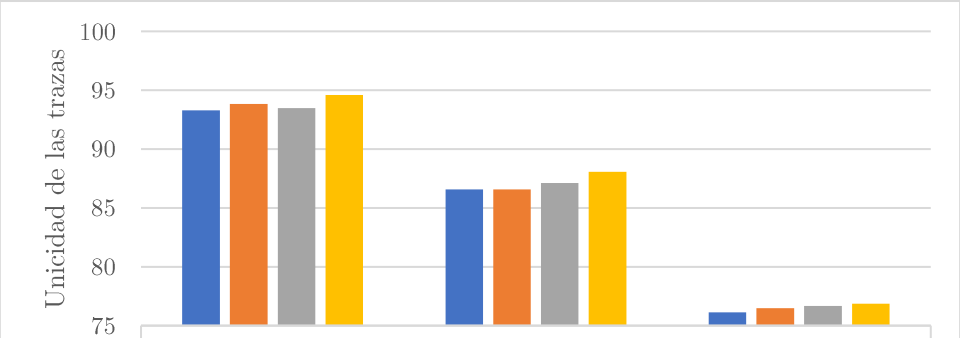
En la siguiente imagen vemos el caso inverso. A la izquierda vemos que en D existían trazas en la zona de las ciudades de Oakland, San Leandro, San Lorenzo y Hayward (al este de SF) que en D_s, a la derecha, ya no corresponden al mismo individuo.



En las zonas con alta densidad de trazas, es difícil ver qué trazas se han intercambiado durante el swapping, aunque si se superponen sobre el mapa las trazas de D y de D_s, es evidente que hay trazas que se han intercambiado. En la siguiente imagen se pueden ver la zona noreste de San Francisco y el aeropuerto para ilustrar este hecho.

7.2.3.1 Unicidad

En la siguiente imagen, se muestra la unicidad ε de las trazas del dataset D_s calculada con los valores de umbral de tiempo τ a 15, 30 y 60 minutos, para valores de p 2, 3, 4 y 5.

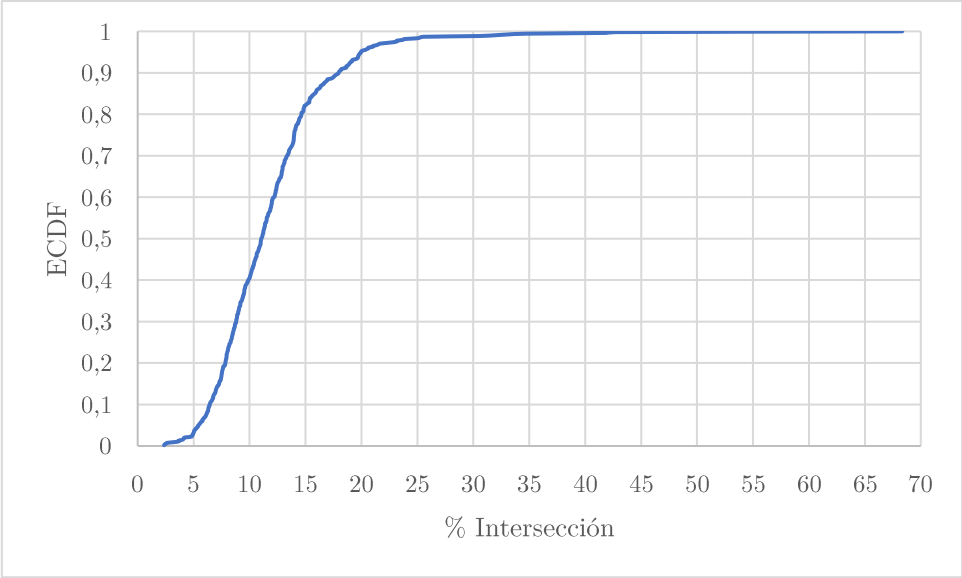


	15 min	30 min	60 min
p=2	93,2836	86,5672	76,1194
p=3	93,8433	86,5672	76,4925
p=4	93,4701	87,1269	76,6791
p=5	94,5896	88,0597	76,8657

La unicidad calculada ε para D_s, es más baja en todos los casos que para D. Se mantiene entre el 93% y 94% para cualquier valor de p con el umbral τ a 15 minutos, mientras que para D, ε toma un valor de cerca del 97% para p=2 y asciende hasta el 100% para p=5. Para τ a 30 minutos, la unicidad de las trazas siempre se mantiene por debajo del valor más bajo calculado para D en este umbral. ε se sitúa en este caso entre el 86,5% y el 88%, representando una disminución importante respecto a D, sobre todo cuanto mayor es p. Por último, para τ a 60 minutos, la diferencia entre la unicidad calculada para D y para D_s es la mayor que hayamos visto. Mientras que para p=2, la unicidad de D se sitúa en el 80%, para D_s se sitúa en el 76%. Lo más remarcable es que, para p>2, la unicidad de D_s se mantiene constante en el 76%, mientras que en D, ε toma valores superiores al 98%. Existe pues, una diferencia de aproximadamente 20 puntos entre la ε calculada para D y para D_s cuando p>2 y τ a 60 minutos.

7.2.3.2 Intersección de trayectorias

Si un atacante conociera exactamente una serie de trazas (tanto la localización como el tiempo exactos de las trazas), podría intentar reidentificar trayectorias en el dataset D. Después de aplicar el método de swapping a D, las trazas del dataset sanitizado D_s sólo intersectaran con las de D en algunos casos. Así pues, aunque el atacante consiguiera identificar una o varias trazas exactas en D_s, probablemente no conseguiría toda la trayectoria original de D. En la siguiente imagen se muestra la función de distribución empírica acumulativa (ECDF, Empirical Cumulative Distribution Function, en inglés) de la intersección entre las trazas de D y D_s.

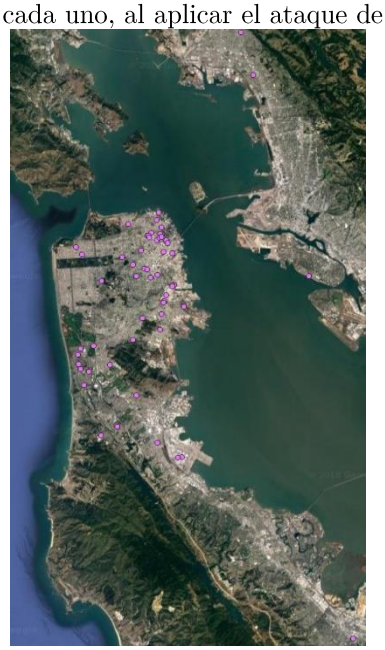


Se puede observar que la mitad de todas las trazas del dataset sanitizado intersectan con las trazas originales en menos del 11,05%. El 80% de las trazas de D_s intersectan con menos del 14,63% de las trazas de D. El 95,3% de las trazas de sanitizadas intersectan con menos del 20,01% de las trazas originales. Más del 98% de las trazas intersectan con menos del 25,5% de las trazas originales. Por último, el 100% de las trazas de D_s intersectan con menos del 68,35% de las trazas de D.

7.2.3.3 Ataques

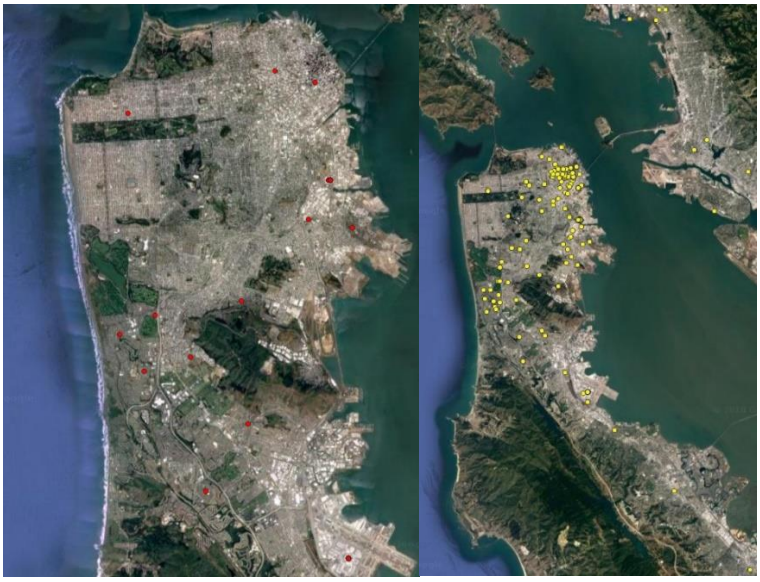
7.2.3.3.1 Deducción de hogares

En la siguiente imagen se muestra la localización más repetida para cada uno de los individuos, es decir, el hogar deducido para cada uno, al aplicar el ataque de deducción de hogares sobre D_s.



De todos los hogares deducidos para el dataset D_s, 251 de ellos, el 46,8% del total, coincide con los hogares calculados para el dataset original D. Si se tienen en cuenta las 5 localizaciones más repetidas para cada individuo, la intersección entre hogares de D y de D_s asciende hasta el 49%. Se trata de un porcentaje de intersección muy alto. Pese a que muchas trazas se han intercambiado durante la aplicación del swapping, queda claro que eso no ha evitado que este ataque obtenga casi la mitad de POIs que corresponden con los reales obtenidos de D. Analizando los hogares

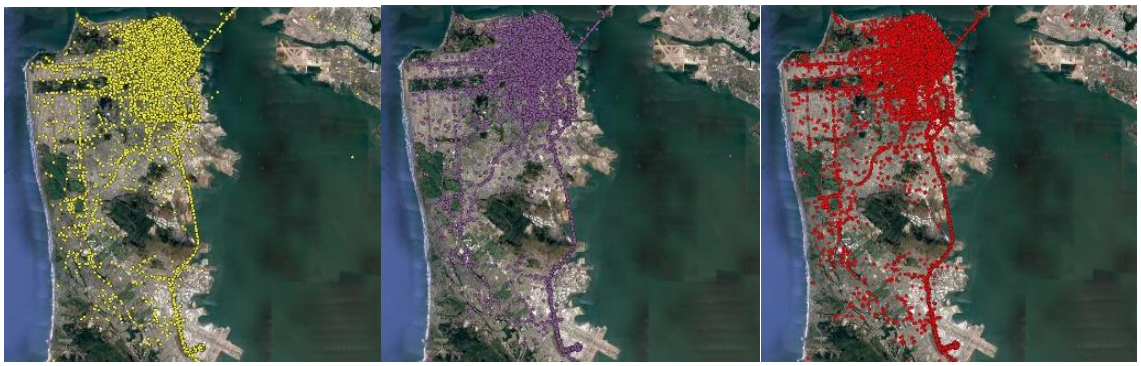
deducidos para D_s que intersectan con los deducidos para D, vemos que 217 de ellos, el 86,45% de todas las intersecciones de hogares, corresponden al aparcamiento de la empresa de taxis. 22 de las intersecciones corresponden al aeropuerto, lo que representa otro 8,76% del total de intersecciones. Si analizamos los 5 puntos más repetidos para cada individuo en lugar de los hogares deducidos (el más repetido), vemos que el 59,59% de las intersecciones corresponden al aparcamiento de taxis, y el 26,26% de las intersecciones corresponden al aeropuerto. Así pues, entre dos de los POIs que - como hemos ido viendo durante el trabajo-, son muy importantes en el dataset D, engloban más del 95% de las intersecciones. Aunque el swapping intercambia las trayectorias entre los sujetos, el hecho de que este ataque agrupe las localizaciones por zonas de 0,001 grados, sumado a la gran concentración de trazas en determinadas zonas como las comentadas, provocan que los hogares coincidan en más del 40% entre el dataset sanitizado y el original.
En la siguiente imagen se muestran las localizaciones de todas las intersecciones de hogares a la izquierda en rojo. A la derecha, se muestran todas las intersecciones de los 5 puntos más repetidos para cada individuo en amarillo.



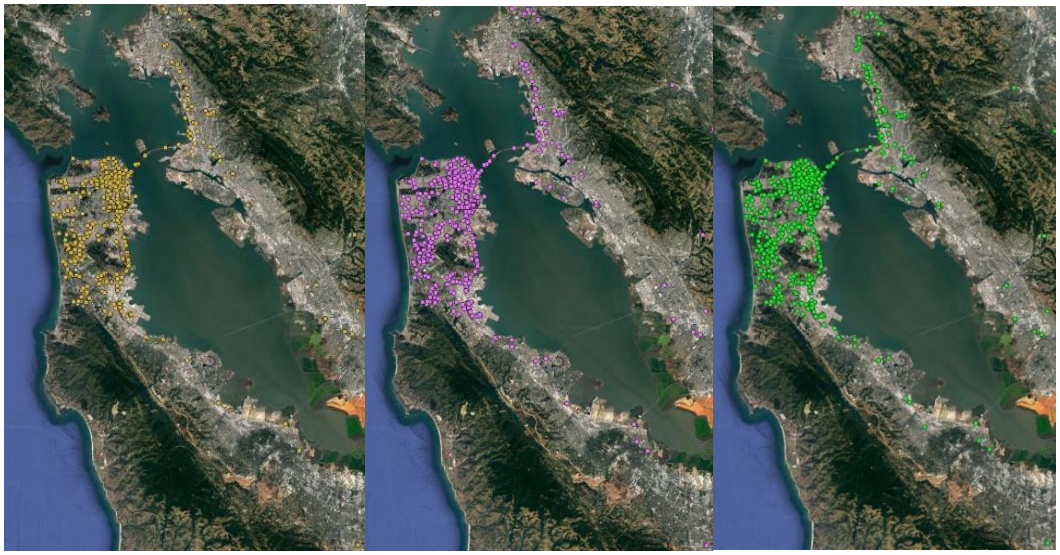
Queda claro que el algoritmo de swapping sufre cuando se aplica este ataque para aquellas zonas en las que hay una gran cantidad de trazas concentradas, y, como se ve en la imagen anterior a la derecha, en las zonas que se encuentran muy alejadas. Esto último se explicaría porque la traza no ha entrado dentro de ningún swap debido a que la trayectoria no ha estado lo suficientemente cerca en distancia y tiempo con las trazas de otro usuario.

7.2.3.3.2 Begin-End

En la siguiente imagen se muestran parte de las localizaciones extraídas mediante heurística con el método de inicio-fin con un umbral de tiempo τ correspondiente a 8 (amarillo), 6 (morado) y 4 horas (rojo).



Para este caso, se han obtenido con τ correspondiente a 8 horas, 19030 POIs, un 470,81% de puntos que con el mismo umbral para el dataset D. La intersección entre los puntos de interés obtenidos para D y para D_s con τ a 8 horas se sitúa en el 25,83% de los puntos obtenidos para D, correspondiente a 1044 puntos en común. Para τ correspondiente a 6 horas se han obtenido 19834 puntos de interés, un 289,63% de puntos que con el mismo umbral para el dataset D. La intersección entre los puntos de interés obtenidos para D y para D_s con τ a 6 horas se sitúa en el 18,27% de los puntos obtenidos para D, correspondiente a 1251 puntos en común. Para τ correspondiente a 4 horas se han obtenido 20902 puntos de interés, un 164,69% de puntos que con el mismo umbral para el dataset D. La intersección entre los puntos de interés obtenidos para D y para Ds con τ a 4 horas se sitúa en el 12,93% de los puntos obtenidos para D, correspondiente a 1641 puntos en común. Para cualquiera de los umbrales de tiempo usados, el número de sujetos para los que al menos se ha obtenido un POI es de 535, uno menos del total de sujetos. El ataque de begin-end contra D, en ningún momento había llegado a obtener un número tan alto de sujetos, lo que indica que las trazas efectivamente se han cambiado. En la siguiente imagen se pueden ver los puntos que intersectan para τ a 8 horas en amarillo, para τ a 6 horas en lila y para τ a 4 horas en verde.



7.3 Comparativa y análisis de los resultados obtenidos

A continuación, se revisarán los resultados de los ataques sobre cada uno de los datasets sanitizados D_s.

Primero se compara la unicidad de los distintos datasets sanitizados D_s en función del método de sanitización. En la siguiente imagen se muestra la unicidad ε en porcentaje para un umbral de tiempo τ de 15 minutos.

	p=2	**p=3**	**p=4**	**p=5**
CabSpotting (D)	96,8284	99,8134	99,8134	100
Rotación 2º	97,0149	99,8134	100	100
Rotación 5º	97,0149	99,8134	100	100
Escala 1 0,7	94,7761	99,6269	100	100
Escala 1,05 0,9	97,5746	100	100	100
Escala 0,95 0,95 + Rotación 1º	95,5224	100	100	100
Agregación	97,9478	99,8134	99,8134	100
Swapping	93,2836	93,8433	93,4701	94,5896

En la siguiente imagen se muestra ε para un τ de 30 minutos para los distintos D_s.

	p=2	**p=3**	**p=4**	**p=5**
CabSpotting (D)	91,6045	99,0672	99,8134	100
Rotación 2º	93,097	99,8134	100	100
Rotación 5º	93,097	99,8134	100	100
Escala 1 0,7	91,2313	99,6269	100	100
Escala 1,05 0,9	90,8582	99,4403	100	100
Escala 0,95 0,95 + Rotación 1º	90,4851	99,4403	100	100
Agregación	92,3507	99,6269	100	100
Swapping	86,5672	86,5672	87,1269	88,0597

En la siguiente imagen se muestra ε para un τ de 60 minutos para los distintos D_s.

	p=2	**p=3**	**p=4**	**p=5**
CabSpotting (D)	80,4104	98,5075	99,8134	99,8134
Rotación 2º	81,1567	99,2537	99,8134	99,8134
Rotación 5º	81,1567	99,2537	99,8134	99,8134
Escala 1 0,7	76,1194	97,7612	99,4403	100
Escala 1,05 0,9	79,1045	98,5075	99,8134	100
Escala 0,95 0,95 + Rotación 1º	78,1716	98,5075	99,8134	100
Agregación	78,7313	98,8806	99,8134	99,8134
Swapping	76,1194	76,4925	76,6791	76,8657

El estudio de la unicidad de las trazas de cada dataset nos da un indicador de la dificultad (o facilidad) para un atacante de reidentificar una traza y relacionarla con el sujeto al que pertenece. Queda claro, a la vista de los resultados, que las trazas, tanto del dataset CabSpotting como de todos los datasets sanitizados obtenidos, tienen un nivel muy alto de unicidad. Revisando todos los casos, el valor más bajo para ε que hemos obtenido es de un 76%. Ya de por sí es un valor alto, pero además hay que tener en cuenta que el umbral de tiempo τ usado es de 60 minutos, lo que podría ser un escenario muy optimista, ya que es probable que un atacante disponga de más precisión en el tiempo.

El dataset original D tiene una unicidad ε muy alta en todos los casos. Excepto para el caso de 2 puntos y umbral de tiempo de 60 minutos, que es el caso más optimista, ningún otro caso baja del 90%, lo que significa que las trazas de D son muy únicas, y por tanto, fácilmente reidentificables. Como era de esperar, cuanto mayor es p para un mismo umbral τ, mayor es el valor de ε, ya que el atacante dispone de más información y es más complicado que varios sujetos dispongan de todas las trazas p. Cuanto mayor es el umbral τ de tiempo, más baja es la unicidad para todos los casos, ya que es más probable que haya varias trazas con la misma localización en ese umbral de tiempo. En cualquier caso que se observe, el método de sanitización del swapping es claramente el mejor a la hora de bajar la unicidad de las trazas del dataset, que es lo que desearíamos para evitar que un atacante tenga fácil la reidentificación. Todos los valores de ε para el D_s obtenido mediante swapping son altos, pero si los comparamos con la unicidad de D, vemos que son los resultados bastante buenos y mejores que los obtenidos con cualquier otro método. Es remarcable que cuando p>2, mientras que para el resto de datasets la unicidad tiende al 100%, para el dataset sanitizado por swapping, el valor de ε se mantiene prácticamente constante, lo que es muy importante porque si un atacante dispone de más puntos, sigue teniendo las mismas probabilidades de éxito.

En la siguiente figura vemos en porcentajes la intersección entre los POIs obtenidos mediante los distintos ataques para el dataset original D y los obtenidos para los datasets sanitizados obtenidos aplicando cada método de sanitización a D.

	Hogares	**Top 5 (Hogares)**	**Begin-end τ=8H**	**Begin-end τ=6H**	**Begin-end τ=4H**	**Stays**
Rotación 2º	0,19	1,04	0	0	0	0
Rotación 5º	0,19	0,22	0	0	0	0
Escala 1 0,7	0	0,22	0	0	0,01	0,01
Escala 1,05 0,9	0	0,26	0	0	0	0
Escala 0,95 0,95 + Rotación 1º	0,19	0,26	0	0	0	0
Agregación	100	99,9	1,34	1,08	0,85	0,22
Swapping	46,8	49	25,83	18,27	12,93	-

Se puede observar fácilmente que los puntos de interés obtenidos para los datasets sanitizados mediante cualquiera de los métodos o combinación de métodos de perturbación prácticamente no intersectan con los POIs obtenidos para D. La mayor intersección en todos estos casos es de un 1%, lo que es una intersección muy baja. Como ya se ha visto previamente en la evaluación de los resultados, el hecho de que las trazas se desplacen debido a una perturbación arbitraria provoca que prácticamente todos los puntos de interés se obtengan en una localización que correspondería a aplicar la misma o mismas perturbaciones a los POIs obtenidos para D. En cualquier caso, aunque los valores de intersección son muy buenos, puesto que un atacante no obtendría prácticamente ningún punto de interés de los sujetos del dataset original D, es evidente que el atacante podría deducir las transformaciones que se han aplicado al dataset, revertirlas para obtener D y obtener todos los puntos de interés de D. Esto es más complicado que suceda para composiciones de perturbaciones, ya que al atacante le será más difícil deducir la posición original cuantas más combinaciones de perturbaciones se apliquen. El problema básico para el método de perturbación es que, si la transformación no es muy sutil -lo cual provocaría una mayor intersección de puntos de interés-, la información agregada no se conserva y la utilidad del dataset se ve muy menguada. En el caso de San Francisco, por su geografía, es un caso en el que la perturbación provoca que el dataset sanitizado tenga trazas que se sitúan en zonas en las que es

evidente que no se encontraban inicialmente, la mayoría sobre el mar o sobre la bahía. Esto provoca que estos datasets no sean demasiado útiles para la creación de mapas de movilidad o predicciones.

En el caso de la agregación de puntos, los resultados obtenidos al aplicar la deducción de hogares al dataset sanitizado son muy malos, puesto que prácticamente todos los POIs obtenidos coinciden con los de D. Como hemos visto previamente, al basarse tanto el método de sanitización como el ataque en la agrupación de puntos por recuadros de precisión 0,001 grados, prácticamente todos los puntos se sitúan en el mismo recuadro en el dataset sanitizado que en D. Para el ataque de heurística y la extracción de stays, este método consigue valores muy bajos. Digamos que los puntos de interés obtenidos con este método se encuentran muy cerca de los puntos de interés de D, pero cuando la precisión es muy alta, no existe intersección.

Por último, el dataset sanitizado mediante swapping obtiene intersecciones bastante altas en el caso de la deducción de hogares, cercanas al 50% de los puntos deducidos. Esto mejora al aplicar la heurística, ya que para cualquier umbral, la intersección de POIs se sitúa por debajo del 26%.

Por último, aunque se ha conseguido aplicar el algoritmo DT-Cluster para la extracción de stays para el resto de datasets sanitizados, lamentablemente no se ha conseguido extraer el conjunto de stays para el dataset sanitizado por swapping, debido a problemas con el rendimiento de la base de datos usada para este trabajo.

8 Conclusiones

Durante la realización de este trabajo se han implementado varios métodos de sanitización y varios ataques de deducción de puntos de interés, tal como se pretendía. Otro objetivo que se ha alcanzado ha sido la sanitización, mediante el uso de los métodos de sanitización implementados, de un dataset con más de 11 millones de trazas. La eficacia de estos métodos de sanitización se ha puesto a prueba realizando ataques contra los datasets sanitizados obtenidos. En el análisis de resultados se ha visto que, pese a que la intersección entre los puntos de interés obtenidos para el dataset original y los obtenidos para los datasets sanitizados mediante el método de perturbación es muy baja, la utilidad de los datos obtenidos al sanitizar el dataset original no tienen utilidad real para las grandes aplicaciones de estos datasets, como podrían ser los mapas de movilidad o la extracción de predicciones. Los datos agregados del dataset original tampoco se mantienen en los datasets sanitizados obtenidos mediante este método. El dataset sanitizado obtenido mediante el swapping obtiene muy buenos resultados respecto al resto de datasets, incluido el original, para la unicidad de sus trazas. Esto significa que la posibilidad de reidentificación de un sujeto para un dataset sanitizado por swapping es más bajo que para el resto de datasets. En el caso del swapping, se ha estudiado especialmente la intersección entre las trayectorias del dataset original y del sanitizado, arrojando buenos datos. El 80% de las trazas del dataset sanitizado intersectan con menos del 15% de las trazas del dataset original. Aunque los resultados obtenidos para la intersección entre los POIs deducidos para el dataset original y el sanitizado no son muy buenos en este caso, el método de swapping es el método estudiado más potente y con más posibilidades, aunque con los umbrales con los que se ha ejecutado en este trabajo, la calidad de la privacidad de los datos no es demasiado buena. Si el algoritmo de swapping se ejecutara con valores mayores para los umbrales de tiempo y distancia, los resultados serían mejores puesto que se intercambiarían más trayectorias de distintos sujetos. Por problemas de rendimiento con la base de datos usada, el swapping se ha aplicado de forma diaria, es decir, realizando los swaps por días, lo que con toda seguridad ha empeorado los resultados obtenidos.

9 Trabajo futuro

Parte del trabajo futuro con el que se podría continuar este trabajo consiste en:

- Agrupar todas las implementaciones de métodos de sanitización y ataques en una herramienta destinada a usuario.
- Mejorar el rendimiento de las implementaciones. Esto está directamente relacionado con el uso de una tecnología determinada. Sería importante valorar el uso de otros lenguajes de programación para la implementación, como podrían ser Matlab o R, que obtienen mejores rendimientos para algunas operaciones que en este trabajo han sido muy costosas en rendimiento.
- Por último, sería interesante implementar otros métodos de sanitización y ataques de deducción de puntos de interés.

10 Bibliografía

[1] Gambs, S.; Killijian, M.O.; Núñez del Prado Cortez, M.
(2011). Show Me How You Move and I Will Tell You Who You Are. Transactions On Data Privacy (Nº 4, Pág. 103–26).
[2] Foley, J.; Van Dam, A.; Feiner S.; Hughes, J. (1990). Geometrical Transformations. Computer Graphics. 2ª Edición. Massachusetts: Addison-Wesley, Pág 201-281. ISBN 9780201121100
[3] Geometric Transformations. (2001) [En línea]. Computer Graphics I. University of Illinois at Chicago. [Consulta: 16 de Noviembre de 2018]. Disponible en: https://www.evl.uic.edu/aej/488/index.html
[4] Homogeneous coordinates (2018) [En línea]. Wikipedia. [Consulta: 16 de Noviembre de 2018]. Disponible en: https://en.wikipedia.org/wiki/Homogeneous_coordinates.
[5] Armstrong, M.P.; Rushton, G.; Zimmerman, D.L. (1999). Geographically Masking Health Data To Preserve Confidentiality. Statistics In Medicine. (Nº 18, Pág. 497–525).
[6] Salas, J.; Megías, D.; Torra, V. (2018). SwapMob : Swapping trajectories for mobility anonymization. En: Domingo-Ferrer J., Montes F. (eds) Privacy in Statistical Databases. PSD 2018. Lecure Notes in Computer Science, vol 11126. Springer, Cham.
[7] de Montjoye, Y. A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. (2013). Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports (Nº 3, Pág. 1-5).
[8] Hariharan, R.; Toyama, K. (2004). Project Lachesis : Parsing and Modeling Location Histories. GIScience (Pág. 106–24).