

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 935 715**

51 Int. Cl.:

G06F 21/62

(2013.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **03.01.2020 PCT/EP2020/050097**

87 Fecha y número de publicación internacional: **16.07.2020 WO20144118**

96 Fecha de presentación y número de la solicitud europea: **03.01.2020 E 20700071 (2)**

97 Fecha y número de publicación de la concesión europea: **19.10.2022 EP 3908952**

54 Título: **Método para crear avatares para proteger datos sensibles**

30 Prioridad:

07.01.2019 FR 1900108

45 Fecha de publicación y mención en BOPI de la
traducción de la patente:

09.03.2023

73 Titular/es:

**BIG DATA SANTE (100.0%)
49 rue Condorcet
44100 Nantes, FR**

72 Inventor/es:

**NEDELEC, YOHANN y
BREILLACQ, OLIVIER**

74 Agente/Representante:

SÁNCHEZ SILVA, Jesús Eladio

ES 2 935 715 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Método para crear avatares para proteger datos sensibles

- 5 La presente invención se refiere a un método para crear avatares a partir de un grupo de datos sensibles iniciales, y que comprende atributos relativos a una pluralidad de personas.

Campo técnico

- 10 La protección de datos sensibles, en particular datos personales, publicados en línea o transmitidos a terceros para su uso, es un tema crucial. Las soluciones actuales que tienen como objetivo brindar a terceros, interesados en el potencial de la información, acceso a grupos de datos sensibles aseguran a un alto nivel la posibilidad de trabajar en archivos sin conocer su contenido. Aunque esto no excluye la necesidad de controlar el acceso a los datos, se requiere asegurar los datos en la fuente, a través de la anonimización real.

- 15 La anonimización se puede definir como la operación de borrar toda la información que permite la identificación directa o indirecta de una persona, contenida en un documento o una base de datos. Existen diferentes métodos para hacer esto.

- 20 Técnica anterior

- El llamado método de sustitución o seudonimización consiste en sustituir el identificador inicial de una persona por otro identificador arbitrario, un seudónimo. Para garantizar la trazabilidad y actualización de la información en la base de datos y evitar asociar a una persona datos relativos a otra persona, a falta de un identificador permanente, es necesario que, para cada persona, este seudónimo sea único.

- Para ello, se puede generar una tabla de correspondencia secreta, que asocia de una vez por todas todos los identificadores con los seudónimos que les han sido asignados. El nivel de seguridad de esta técnica de anonimización es bajo, donde la operación es reversible, ya que a partir del seudónimo se puede encontrar el identificador y quien posee la tabla puede leer perfectamente la base de datos.

- También se puede usar una función llamada "hash", que tiene la particularidad, en comparación con los algoritmos de cifrado estándar, de no ser reversible: no es posible encontrar el identificador inicial solo del seudónimo, incluso si se conoce la función hash utilizada. Sin embargo, a pesar de esta irreversibilidad en principio, esta técnica puede ser derrotada reconstituyendo, por reiteración, una tabla de correspondencia.

- Es posible reforzar la seguridad de la anonimización agregando previamente a los identificadores iniciales una clave secreta arbitraria, que luego se somete a la función hash. Quien desee reconstituir la tabla de correspondencias, por lo tanto, no sólo tendrá que probar todos los apellidos y nombres posibles, lo cual es relativamente fácil, sino también todas las modificaciones que estos identificadores pueden experimentar a partir de claves desconocidas. Sin embargo, la seguridad del dispositivo se basa una vez más en la confidencialidad de las herramientas utilizadas: la clave secreta por un lado, la función hash utilizada por el otro.

- Aún es posible reforzar la anonimización, realizando un doble hash con clave secreta, que consiste en realizar la operación una primera vez, y someter el seudónimo obtenido a una segunda función hash con clave secreta. Para garantizar la total confidencialidad, las claves se pueden renovar periódicamente.

- Los métodos de supresión, enmascaramiento o adición de ruido consisten en degradar la información inicial, suprimiendo determinados datos o, por el contrario, añadiendo información que distorsiona los datos iniciales. La desventaja de tales métodos es que distorsionan los datos desde el principio, lo que los hace menos relevantes para un uso distinto al original. La solicitud FR 3 047 586 divulga un método de introducción de ruido sobre datos sensibles divididos en subconjuntos, donde se aplican diferentes niveles de ruido, dependiendo del nivel de sensibilidad determinado para cada subconjunto.

- En la solicitud internacional WO 2017/093736, la anonimización de los datos se logra utilizando una o más técnicas que alteran los datos originales, como el uso de "tokens" en inglés, la generalización, el flujo de datos, inserción sintética de registros, eliminación o reorganización de registros. En estas dos solicitudes, la anonimización se realiza mediante sucesivas transformaciones sobre los datos sensibles individuales.

- La agregación es un método que consiste en agrupar múltiples valores en una sola clase. Por ejemplo, las personas, originalmente de 25, 27 y 30 años de edad, se clasificarán dentro del mismo grupo de edad de 25 a 30 años. Cuanto mayor sea el nivel de agregación, menor será el riesgo de reidentificación, pero, al mismo tiempo, menos precisa será la información. Por lo tanto, la autoridad a cargo de la base de datos debe encontrar un equilibrio satisfactorio entre la solidez de la anonimización garantizada por el nivel de agregación y la precisión de los datos que autoriza. La solicitud internacional WO 2018/028783 describe un método que permite anonimizar conjuntos de datos que contienen información sensible, utilizando dicho método de agregación.

Además, en un contexto de difusión exponencial de datos de todo tipo, públicos o privados, sobre todos los objetos posibles, lo que en ocasiones se denomina con el término "grandes datos", las técnicas de validación de la información son un medio muy eficaz para romper el anonimato de las bases de datos sensibles. Ciertos datos, más sensibles que otros, como los relativos al estado de salud, la vida familiar, los asuntos jurídicos o los bienes, también están sujetos a precauciones especiales, tal y como lo recomienda la Comisión Nacional de la Informática y las Libertades (CNIL) en Francia. Ninguna solución permite mantener toda la señal contenida en un grupo de datos eliminando cualquier riesgo de identificación. Sin embargo, se han producido muchas dificultades con las técnicas de anonimización.

Luisa Franconi ET AL: "Community Innovation Survey: comparable dissémination", Sesión de trabajo sobre confidencialidad de datos estadísticos, Manchester, 17-19 de diciembre de 2007, edición de 2009, 31 de diciembre de 2009, páginas 11-23, Luxemburgo, ISBN: 978-92-7-912055-8 (XP055633756) divulga la anonimización de una base de datos de tipo microdatos mediante microagregación.

El documento US 2012/030165 A1 divulga la anonimización de una base de datos.

VICENÀ TORRA ET AL: "Evaluating Fuzzy Clustering Algorithms for Microdata Protection", 29 de junio de 2004 (2004-06-29), PRIVACIDAD EN BASES DE DATOS ESTADÍSTICAS; NOTAS DE LECTURA EN CIENCIAS DE LA INFORMÁTICA, SPRINGER-VERLAG, BERLIN/HEIDELBERG, PÁGINA(S) 175 - 186, ISBN: 978-3-540-22118-0 (XP019006163) también divulga la anonimización de una base de datos de tipo microdatos mediante microagregación.

Divulgación de la invención

En consecuencia, existe la necesidad de mejorar aún más la protección de los datos sensibles, haciéndolos irreversiblemente no identificables, manteniendo los datos lo más útiles posible para fines de análisis.

El objeto de la presente invención es precisamente satisfacer esta necesidad.

Resumen de la invención

La invención consiste en un método, un dispositivo y un producto de programa como se define en las reivindicaciones.

La presente invención se refiere así a un método para crear avatares a partir de un grupo de datos sensibles iniciales almacenados en una base de datos de un sistema informático, dichos datos iniciales comprenden atributos relativos a una pluralidad de personas, el método que comprende:

- a) elegir, para los atributos relativos a una persona dado, un número k de vecinos más cercanos que se utilizarán entre el conjunto de personas del grupo de datos inicial,
- b) identificar, para esa persona, los k vecinos más cercanos entre las otras personas en el grupo de datos,
- c) generar, para al menos un atributo relativo a esta persona, un nuevo valor de atributo a partir de cantidades características de dicho atributo en dichos k vecinos más cercanos identificados, ponderados por un coeficiente,
- d) crear datos de avatares con los nuevos valores de atributo para hacer que los datos sensibles sobre la persona no sean identificables.

La invención proporciona una anonimización real en el sentido de la pérdida de la capacidad de volver a identificar a las personas de las que se originaron los datos. Esta creación de avatares, correspondientes a personas virtuales, de síntesis, denominada "avatarización", pretende oponerse a cualquier reidentificación, directa o indirecta.

La invención se centra en observaciones individuales ("basadas en datos" y "centradas en el paciente" en inglés), y no en un gran conjunto de personas que se parecen.

Esta se diferencia de algunos métodos conocidos que se basan en la observación de subgrupos. Esta siempre utiliza el entorno de la persona para crear su avatar. Esta permite gestionar grandes volúmenes de datos manteniendo los vínculos existentes entre las variables.

A diferencia de los métodos conocidos que se describen a continuación, la invención genera por completo un nuevo registro sintético: el avatar. Esta atraviesa una etapa no alcanzada por las técnicas actuales en el sentido de que propone generar íntegramente un dato resumen a partir del entorno local de los datos originales. El avatar es el resultado de una generalización local de su entorno. Por ejemplo, en el contexto de los datos médicos, el avatar de un paciente se crea a partir de personas similares, cuyo número es configurable.

La invención permite garantizar el mantenimiento del interés del grupo de datos, ofreciendo la posibilidad de realizar los mismos análisis, por ejemplo estadísticos o por métodos de inteligencia artificial, y obtener resultados comparables entre el grupo de datos inicial y los avatares.

Es posible la conservación de la estructura de los datos iniciales, es decir su granularidad, mientras que las soluciones conocidas con agregación o generalización ofrecen buena seguridad pero pierden retención de señal. Por ejemplo, el método según la invención permite mantener valores continuos para la edad, 30, 32 o 35 años, en lugar de una sola categoría, de 30 a 35 años.

La invención facilita el acceso a los datos y ofrece una solución completa, haciendo accesible el método a personas no expertas. Esto ahorra tiempo al usuario al ofrecer una solución llave en mano y no una herramienta que requiere la intervención de un experto ("data-scientist" en inglés), un abogado y un proveedor de soluciones para compartir datos.

Datos iniciales y número k de vecinos más próximos

Por "información sensible", según la definición de la CNIL, es necesario incluir información relativa al origen racial o étnico, opiniones políticas, filosóficas o religiosas, afiliación sindical, salud o vida sexual. Entre los datos sensibles se encuentran los datos personales, en particular los datos de salud.

El grupo de datos iniciales incluye atributos relativos a una pluralidad de personas, y también se denominan "microdatos". Estos datos son, por ejemplo, registros que contienen información sobre pacientes que acuden a urgencias o que acuden a un organismo sanitario, o que forman parte de una cohorte de estudio de una determinada enfermedad, de un ensayo clínico, o información diversa relativa a los habitantes de un municipio, o perfiles de usuarios de un determinado servicio.

El grupo de datos iniciales puede descargarse interactuando con una base de datos del sistema informático en el que se implementa el método según la invención, o con una base de datos de un sistema informático remoto.

Los datos iniciales pueden comprender al menos un atributo de valor numérico, denominado variable continua, y/o al menos un atributo perteneciente a una clase representativa de una característica de la persona, denominado variable categórica o atributo de clase.

El usuario puede especificar las restricciones lógicas vinculadas a la naturaleza de los atributos.

Ventajosamente se realiza un análisis multivariante sobre los datos iniciales, previo a la identificación de los k vecinos más cercanos de la persona, incluido el análisis de componentes principales (PCA). Este método matemático permite la reducción de la dimensionalidad de un grupo de datos al liberar los ejes más importantes de varianzas, discriminación, informatividad y la representación de la esencia de un grupo de datos de N dimensiones en un número menor de dimensiones respetando los parámetros definidos por usuario, en particular asignando pesos a cada variable.

Los valores de los atributos así modificados pueden luego ser proyectados en el mismo espacio geométrico.

El análisis multivariante, en particular en los componentes principales, se puede realizar en atributos categóricos y continuos.

Cada atributo también se puede definir como más o menos sensible. Por lo tanto, el usuario puede elegir perturbar el grupo de datos con más fuerza para ciertos atributos en particular. Durante el análisis multivariante, se pueden asignar diferentes pesos a cada atributo, lo que modifica la proyección de datos utilizada para la identificación del vecino más cercano.

Por ejemplo, habiéndosele asignado un peso significativo al atributo "tamaño", este último tiene un efecto exacerbado y distorsiona la proyección en el espacio euclidiano. Así, las distancias entre los puntos se ven particularmente afectadas por el atributo "tamaño" frente a los demás atributos: pequeñas diferencias de tamaño tienen mayores efectos sobre la proyección y el entorno de los pacientes.

Vecinos más cercanos

El número k de vecinos más próximos pueden ser función de los datos iniciales de la persona considerada. El número k de vecinos más próximos pueden ser variables, siendo diferentes para determinadas personas o de una persona a otro, para el mismo grupo de datos iniciales.

Este parámetro se puede elegir según un factor de sensibilidad de los datos iniciales y/o un factor de confianza en el destinatario destinado a recibir los datos del avatar: cuanto más elevado sea el factor de sensibilidad y/o más bajo sea el factor de confianza, más elevado será el número k, y cuanto más bajo sea el factor de sensibilidad y/o más elevado sea el factor de confianza, el número k se reducirá.

El factor de sensibilidad y el factor de confianza pueden estar entre 0 y 1.

La elección del número k vecinos más cercanos condiciona el equilibrio deseado entre la protección de los datos iniciales y la retención de la señal. Los grupos de datos iniciales pueden, de hecho, tener diversos grados de sensibilidad, es decir, que las consecuencias vinculadas a su conocimiento sean diferentes. Por ejemplo, en el caso de los datos médicos, conocer el índice de masa corporal de una persona es menos dañino que saber si esa persona tiene cáncer.

También es importante estimar la confianza que el usuario deposita en el destinatario de los datos de los avatares. Por ejemplo, en el contexto de un intercambio de datos dentro del mismo hospital, este nivel de confianza es más importante que si los datos están destinados a publicarse en línea.

La invención se basa en la realización de un modelado local a partir de los vecinos más próximos para cada registro correspondiente a una persona. En datos médicos, por ejemplo, esto consiste en identificar, para cada paciente, los k pacientes que se le parecen. Los tratamientos aplicados se realizan así según la cercanía de cada persona. Por lo tanto, la invención se centra en personas y no en un conjunto de personas que se parecen entre sí, y por lo tanto difiere de los métodos conocidos que se basan en observaciones grupales.

Datos de avatares

En el caso de que los datos iniciales comprendan al menos un atributo con valor numérico, el nuevo valor del atributo corresponde preferentemente a un baricentro ponderado estocástico calculado a partir de los valores de este mismo atributo de los k vecinos más cercanos, cada uno ponderado por un coeficiente elegido al azar de una distribución uniforme o exponencial.

Se puede elegir aleatoriamente un coeficiente diferente para cada valor del atributo de cada uno de los k vecinos más cercanos.

En el caso de que los atributos se proyecten en un mismo espacio euclidiano, para cualquier punto M , existe ventajosamente G , el baricentro con ponderación estocástica ponderado del polígono en k vértices, correspondientes a las posiciones de k vecinos más cercanos:

[Ecuación 1]

$$\sum_{i=1}^k \alpha_i \overrightarrow{MA_i} = \left(\sum_{i=1}^k \alpha_i \right) \overrightarrow{MG}$$

α_i : coeficiente de ponderación aleatoria,
 A_i : vértice del polígono (vecino),
 tal que :

[Ecuación 2]

$$\sum_{i=1}^k \alpha_i \overrightarrow{GA_i} = \vec{0}.$$

El baricentro ponderado estocástico G se calcula ventajosamente en un número N de dimensiones correspondiente al número de atributos que tienen un valor numérico.

El método según la invención permite volver a centrar los registros marginales de forma más significativa para eliminar el riesgo de reidentificación, en parte gracias a que el cálculo del nuevo valor del atributo no utiliza directamente el valor original, sino únicamente su cercanía para lograr una generalización local.

En el caso de que los datos iniciales incluyan al menos un atributo perteneciente a una clase representativa de una característica de la persona, el nuevo valor del atributo se elige ventajosamente de forma aleatoria según la representación ponderada de dicha clase entre los k vecinos más próximos, calculado según el número de valores pertenecientes a dicha clase en los k vecinos más cercanos ponderado por la cantidad de valores que pertenecen a esta clase en todo el grupo de datos iniciales.

Por ejemplo, en el caso de un atributo que indique si una paciente está embarazada o no, las proporciones de las clases "embarazada" y "no embarazada" se calculan dentro de los k vecinos más cercanos y luego ponderados por la representación de esas mismas clases en todo el grupo de datos. Luego, el nuevo valor del atributo se extrae aleatoriamente en función de esta representación ponderada.

Este enfoque permite combinar la estocasticidad del sorteo aleatorio garantizando la dificultad de reidentificación, con la seguridad de distorsionar este sorteo por la representatividad de cada clase entre los vecinos más cercanos.

Se puede utilizar una distribución uniforme, logarítmica normal o exponencial para los cálculos de coeficientes aleatorios de ponderación para los atributos numéricos y la elección aleatoria de nuevos valores de atributo de clase. Los nuevos valores de atributos numéricos y de clase se generan ventajosamente de forma conjunta para crear los datos de avatares. La invención, que conduce a la creación de datos de avatares, corresponde ventajosamente a una transformación vectorial de los datos iniciales de la persona.

El grupo de datos se reconstituye ventajosamente a partir de variables continuas y categóricas. La invención asegura la conservación del tipo de los atributos originales. Por ejemplo, si la edad es un atributo que toma valores enteros en el grupo de datos iniciales, tendrá la misma forma después de que se creen los avatares. Este resultado difiere de los métodos de generalización conocidos en los que varios pacientes de edades similares se encuentran en el mismo grupo de edad después de la anonimización de los datos que confirman una pérdida de información.

Se puede aplicar al menos una función de restricción a los datos de avatares, para limitar los valores aberrantes o la alteración de la relación entre atributos relacionados con la misma persona, en particular para restringir las distribuciones de los atributos numéricos, o para respetar una relación lógica entre los atributos de clase. El mínimo y el máximo de un atributo numérico inicial pueden utilizarse para modificar, si es necesario, el atributo de avatar resultante, en particular para no mostrar una edad inferior a la edad mínima del grupo de datos iniciales. Se puede aplicar una función de restricción, basada en reglas de relaciones lógicas entre los diversos atributos, previamente establecidas, por ejemplo, el hecho de que un hombre no puede estar embarazado. Esto aumenta aún más el parecido entre el grupo de datos iniciales y los datos del avatar y ayuda a mantener la coherencia de los datos.

Ventajosamente, el grupo de datos constituido por avatares tiene las mismas dimensiones que el grupo de datos iniciales.

Las operaciones de procesamiento aplicables al grupo de datos inicial siguen siendo válidas en los datos de los avatares.

Registro y acceso a los datos de avatares

Los datos de avatares se registran ventajosamente en una base de datos del mismo sistema informático y/o en la de un servidor remoto.

El grupo de datos de avatares se puede registrar en forma de archivo en una estación local, por ejemplo, en el contexto de uso interno. En una variante o en combinación, el grupo de datos de avatares se puede registrar en un servidor remoto, en particular mediante carga a través de un canal de transmisión, tal como Internet o intranet.

Se puede crear un acceso restringido para los destinatarios de los datos de avatares. Estos pueden ser notificados de la disponibilidad de dichos datos por un sistema de autenticación híbrido que comprende al menos dos etapas, basado en particular en la generación y transmisión de un enlace de descarga único, por ejemplo por correo electrónico, luego la comunicación de una clave con una vida útil limitada, por ejemplo por SMS. El usuario también puede decidir restringir el acceso a determinados atributos, personas o establecer un número máximo de solicitudes que puede realizar el destinatario.

Los destinatarios pueden acceder a los datos de avatares a través de varios medios, como descargar el archivo del avatar, por ejemplo en formato CSV, XML o XLS, directamente desde un enlace único personalizado, navegar y descargar el archivo a través de un protocolo seguro, por ejemplo, el protocolo SFTP, una solicitud segura a una base de datos en un servidor remoto, o una interfaz de programación (API) que permite al destinatario interactuar directamente con una de sus aplicaciones con los datos.

En este último caso, el usuario puede ventajosamente proporcionar una versión actualizada en tiempo real del grupo de datos del avatar. Esto es particularmente valioso en un contexto en el que es probable que se revoque el acceso a datos sensibles, por ejemplo, tras la implementación del Reglamento General de Protección de Datos, o cuando una cohorte continúa aumentando su número de registros, por ejemplo, en el caso de pacientes reclutados en un estudio médico en curso.

Se puede generar automáticamente un informe que contiene los detalles de la creación de los datos del avatar, incluidos los parámetros del análisis multivariado utilizado, el número de personas y atributos, numéricos y de clase, y el número k de vecinos más cercanos. Este informe permite la trazabilidad de las operaciones realizadas, y es útil para necesidades de archivo o legales. Este informe obviamente no contiene ninguna información que permita encontrar los datos iniciales. Tenga en cuenta que este informe no permite de ninguna manera revertir las operaciones de avatarización realizadas porque no guarda la traza de los valores generados aleatoriamente, inherentes a la operación del método.

Todas las etapas de la invención son ventajosamente implementadas automáticamente por un ordenador.

Producto de programa informático

5 Otro objeto de la invención, según otro de sus aspectos, es un producto programa informático para implementar el método según la invención para crear avatares a partir de un grupo de datos sensibles iniciales almacenados en una base de datos de un sistema informático, dichos datos iniciales comprenden atributos relativo a una pluralidad de personas, el producto de programa informático que comprende un medio y registra en este medio instrucciones legibles por un procesador para, cuando son ejecutadas:

- 10 a) un número k de vecinos más cercanos que se utilizarán entre el conjunto de personas del grupo de datos iniciales se elige para los atributos relacionados con una persona dada,
- b) para esa persona, los k vecinos más cercanos se identifican entre las otras personas en el grupo de datos,
- c) para al menos un atributo relacionado con esta persona, se genera un nuevo valor de atributo a partir de cantidades características de dicho atributo en dichos k vecinos más cercanos identificados, ponderados por un
- 15 coeficiente, y
- d) se crean datos de avatares que comprenden los nuevos valores de atributo para hacer que los datos sensibles sobre la persona no sean identificables.

20 Las características indicadas anteriormente para el método se aplican al producto del programa informático, y viceversa.

Dispositivo de creación de avatares

25 La invención también se refiere, según otro de sus aspectos, a un dispositivo para crear avatares a partir de un grupo de datos sensibles iniciales almacenados en una base de datos de un sistema informático, dichos datos iniciales comprenden atributos relativos a una pluralidad de personas, el dispositivo que comprende:

- a) un módulo de elección configurado para elegir un número k de vecinos más cercanos que se utilizarán entre el conjunto de personas en el grupo de datos iniciales, para los atributos relacionados con una persona dada,
- 30 b) un módulo de identificación configurado para identificar, para esta persona, los k vecinos más cercanos entre las otras personas en el grupo de datos,
- c) un módulo de generación configurado para, al menos para un atributo relativo a esta persona, generar un nuevo valor de atributo a partir de cantidades características de dicho atributo en dichos k vecinos más cercanos identificados, ponderados por un coeficiente, y
- 35 d) un módulo de creación de avatares configurado para crear datos de avatares que comprenden el o los nuevos valores de atributo, para hacer que los datos sensibles relacionados con la persona no sean identificables.

40 El dispositivo según la invención también puede comprender un módulo de registro configurado para registrar los datos de avatares en una base de datos del mismo sistema informático y/o de un servidor remoto.

El dispositivo para crear avatares puede comprender un módulo para especificar las restricciones lógicas específicas del grupo de datos, vinculadas a la naturaleza de los atributos.

45 El dispositivo también puede comprender un módulo para elegir el número k de vecinos más cercanos.

El dispositivo de creación de avatares puede comprender un módulo de evaluación de propiedades de los avatares según el riesgo de reidentificación y el riesgo de pérdida de información.

50 El dispositivo de creación de avatares puede comprender un módulo de alerta configurado para advertir a los destinatarios de los datos de avatares de la provisión de dichos datos por un sistema de autenticación híbrido que comprende al menos dos etapas, basado en particular en la generación y la transmisión de un enlace de descarga único, por ejemplo por correo electrónico, luego la comunicación de una clave con una vida útil limitada, por ejemplo por SMS.

55 Las características indicadas anteriormente para el método y el producto del programa informático se aplican al dispositivo, y viceversa.

Breve descripción de los dibujos

60 La invención podrá comprenderse mejor con la lectura de la descripción detallada que sigue, de los ejemplos de implementación no limitativos de la misma, y con el examen del dibujo adjunto, en el que:

65 [Figura 1] ilustra los pasos para crear datos de avatares a partir de datos sensibles iniciales de acuerdo con la invención,

[Figura 2] representa un ejemplo de creación de datos de avatares para un atributo de origen numérico,

[Figura 3] ilustra la distribución de la densidad de probabilidad de los valores de los atributos de avatares con respecto a los vecinos más cercanos en el ejemplo de la Figura 2,

[Figura 4] representa la posición en el espacio de los datos de avatares y los datos iniciales,

[Figura 5] representa la diferencia en la correlación entre diferentes atributos de un grupo de datos inicial y un grupo de datos de avatares,

[Figura 6] representa la evolución de la distancia entre un avatar creado según la invención y la persona de origen más cercano, según el número k de vecinos más cercanos y

[Figura 7] representa la evolución de las distancias entre el vecino más cercano y una persona, en un grupo de datos inicial, en función del número k de vecinos más cercanos.

Descripción detallada

En la figura 1 se ilustra un ejemplo de pasos para crear datos de avatares a partir de datos sensibles iniciales según la invención.

En este ejemplo, durante un paso 11, se descarga un grupo de datos iniciales al interactuar con una base de datos del sistema informático en el que se implementa el método según la invención, o con una base de datos de un sistema informático remoto. Este grupo de datos iniciales comprende atributos relativos a una pluralidad de personas, que pueden ser atributos numéricos y/o atributos pertenecientes a una clase representativa de una característica de la persona. El usuario puede especificar las restricciones lógicas vinculadas a la naturaleza de los atributos.

En un paso 12, como se ha descrito anteriormente, se fijan factores de sensibilidad y confianza, para elegir, en un paso 13, el número k de vecinos más cercanos. El número k de los vecinos más cercanos se pueden elegir de forma independiente para cada dato individual sensible, por lo que también se puede determinar localmente y variar de una persona a otra para el mismo grupo de datos.

Durante un paso 14, se realiza un análisis multivariante de los datos iniciales, en particular un análisis de componentes principales, asignando un peso a cada atributo en un paso 14bis. La identificación de los k vecinos más cercanos de la persona se realiza así sobre los datos modificados, en un paso 15. Por ejemplo, si se asigna un peso de 0,1 a la edad, un peso de 0,2 al IMC y un peso de 0,75 al número de embarazos, se especifica así que este último atributo es especialmente sensible, lo que distorsionará la proyección de los datos durante el análisis multivariante.

En el caso de un atributo numérico, durante un paso 16, se elige aleatoriamente un coeficiente para cada valor del atributo de cada uno de los k vecinos más cercanos, para calcular, en un paso 17, un baricentro con ponderación estocástica G a partir de los valores de este mismo atributo de los k vecinos más cercanos, cada uno ponderado por este coeficiente aleatorio para obtener los nuevos valores del dato de avatar. En el caso de que los datos contengan N atributos numéricos, se calcula el baricentro con ponderación estocástica G en las N dimensiones, dando así todos los valores nuevos para los atributos considerados.

En el caso de un atributo perteneciente a una clase representativa de una característica de la persona, durante un paso 18, se calcula el número de atributos pertenecientes a dicha clase en los k vecinos más cercanos, luego este número, en un paso 19, se pondera por el número de personas que pertenecen a esta clase en el grupo de datos inicial, para elegir aleatoriamente el nuevo valor del atributo en función de esta representación ponderada, en un paso 20.

En un paso 21, los nuevos valores de atributos numéricos y de clase se generan conjuntamente para crear los datos de avatares para hacer que los datos sensibles relacionados con la persona no sean identificables.

Durante un paso 22, las restricciones definidas por el usuario se aplican a los datos del avatar, para limitar los valores aberrantes o la alteración de la relación entre atributos relacionados con la misma persona, como se describió anteriormente.

El método según la invención se repite para cada persona. El método se centra en la observación individual sensible. Este permite generar a voluntad avatares de personas cuyos datos son sensibles, como se haría por los métodos conocidos de simulación por modelado global o por remuestreo.

Los datos de avatares así creados pueden registrarse en una base de datos del mismo sistema informático, en un paso 23, y/o en la de un servidor remoto, en un paso 24.

La gestión de los destinatarios de los datos de avatares se realiza durante un paso 26. Los datos pueden estar disponibles por varios medios, como descargar el archivo de avatares, por ejemplo en formato CSV, XML o XLS, directamente desde un enlace único personalizado, navegar y descargar el archivo a través de un protocolo seguro, por ejemplo, el protocolo SFTP, una solicitud segura a una base de datos en un servidor remoto, o una interfaz de programación (API) que permite al destinatario interactuar directamente con una de sus aplicaciones con los datos.

Los destinatarios pueden ser notificados de la disponibilidad de los datos mediante un sistema de autenticación híbrido que comprende al menos dos etapas, basadas en particular en la generación y transmisión de un enlace de descarga único, por ejemplo por correo electrónico, luego la comunicación de una clave con una vida útil limitada, por ejemplo por SMS.

Se puede generar automáticamente un informe que contiene los detalles de la creación de los datos de avatares, en un paso 25, comprendiendo este informe, por ejemplo, los parámetros del análisis multivariado utilizado, el número de personas y atributos, numéricos y de clase, y el número k de vecinos más cercanos.

La Figura 2 presenta un ejemplo de creación de un avatar a partir de datos iniciales que comprenden atributos cuantitativos (edad y recuento de glóbulos rojos, visibles en la Tabla 1 a continuación). Este ejemplo se realiza sobre datos brutos en aras de la simplicidad, pero la creación de los avatares se realiza ventajosamente después de un análisis multivariante, en particular en componentes principales.

[Tabla 1]

Tabla 1

Persona	Edad	Glóbulos rojos (M/ul)
O	25 3,5	
A	19	4,5
B	28	1,8
C	42	5,2

Como se ve en la Figura 2, los valores de los atributos de cada persona se proyectan en el mismo espacio euclidiano. En este ejemplo, se crea un avatar Y para la persona O usando 3 de sus vecinos, es decir $k = 3$. Se identifican los vecinos más cercanos de la persona O. Para cada vecino, se extrae aleatoriamente un coeficiente. Las coordenadas de cada vecino con los coeficientes asociados se utilizan para calcular el baricentro con ponderación estocástica ponderado y formar los datos del avatar. En el ejemplo considerado, el avatar Y de la persona O tiene un valor de atributo para su edad de 34,47 y 4,13 M/ul para el atributo de recuento de glóbulos rojos.

La Figura 3 ilustra la distribución de densidad de probabilidad de los valores de los atributos de avatares en relación con los vecinos más cercanos en el ejemplo anterior. Siendo el método según la invención parcialmente aleatorio, si se crea un gran número de avatares para una misma persona, estos se distribuyen en el espacio entre los k vecinos más cercanos según una densidad de probabilidad, como se ve en la Figura 3.

Ejemplo

A continuación se describirá un ejemplo de implementación y eficacia de la invención. En este ejemplo, el grupo de datos se refiere a pacientes reales y proviene de la base de datos de libre acceso "Base de datos de diabetes de los indios Pima" que comprende 768 personas, 8 variables continuas (número de embarazos, nivel de glucosa en sangre, presión arterial, grosor de la piel, nivel de insulina en sangre, edad, índice de masa corporal y función pedigrí de la diabetes), y una variable categórica (padece o no de diabetes).

Se realiza un análisis de componentes principales sobre los datos para proyectar el grupo de datos de 9 dimensiones, correspondiente a los 9 atributos, por ejemplo, en un espacio de 3 dimensiones cuando se desea visualizar los datos. El método según la invención, como se ha descrito anteriormente, se aplica a todos los datos. El número k de vecinos más cercanos se establece en 50.

La representación de la figura 4 muestra que, para $k = 50$ vecinos más cercanos considerados, el método según la invención permite terminar con datos de avatares que evolucionan en el mismo espacio que los datos iniciales y supone así una buena conservación de las propiedades del grupo de datos original. Por definición, un registro marginal, es decir, una persona con valores poco comunes, es más sensible a la reidentificación. Gracias a la invención, como puede verse en la Figura 4, los registros marginales son parcialmente desmarginados y sistemáticamente relacionados con otros registros. Por el contrario, un registro ordinario genera un avatar cercano a los muchos otros registros que son similares a él. El avatar generado se distingue del registro original por sus múltiples proximidades con otros registros individuales.

Un grupo de datos de interés estadístico está compuesto por variables que tienen relaciones entre ellas. El efecto de la creación de avatares sobre estas relaciones se evalúa calculando la diferencia de correlación entre los atributos del grupo de datos inicial y los de los datos de avatares. Se puede observar en la Figura 5 que la diferencia de correlación, para $k = 50$, es moderado y bastante equilibrado: el método según la invención no ha alterado demasiado significativamente la relación entre los atributos. Estos dos análisis permiten evaluar los efectos globales

de la invención sobre los datos. El análisis puede desglosarse localmente para evaluar los efectos locales de la invención en subgrupos de datos.

La conservación de la señal también se puede validar en un primer nivel realizando el mismo análisis sobre el grupo de datos inicial y sobre los datos de avatares: estableciendo un modelo de predicción de diabetes utilizando los otros ocho atributos del grupo de datos. Para esto, se genera en el 80 % del grupo de datos de avatares y en el 80 % del grupo de datos iniciales un algoritmo predictivo, por ejemplo, bosques aleatorios. Luego se evalúa el desempeño de los dos modelos usando el 20 % restante del grupo de datos original, midiendo su sensibilidad y especificidad. En nuestro ejemplo sobre el grupo de datos de diabetes, los valores obtenidos para estas dos mediciones son muy parecidos: sensibilidad de 0,575 para los datos iniciales frente a 0,553 para los datos de avatares, y especificidad de 0,868 para los datos iniciales frente a 0,89 para los datos de avatares.

Un modelo generado con avatares permite así predecir la aparición de diabetes con un rendimiento comparable al de un modelo generado con datos originales. El rendimiento de clasificación del algoritmo predictivo es comparable y valida la utilidad de los avatares en el contexto de los análisis, por ejemplo, estadísticos o mediante aprendizaje automático. Por lo tanto, la conservación de las propiedades originales permite aplicar a los datos de avatares el mismo procesamiento y análisis que los realizados sobre el grupo de datos iniciales.

En un ejemplo de escenario de ataque, a un atacante le gustaría saber si una persona, de la que conoce cierta información, ha elegido participar en un estudio científico relacionado con la diabetes. Para ello, en un caso de estudio desfavorable para la invención, este dispone de datos iniciales completos que comprenden todos los datos sensibles relativos a esta persona, y un grupo de datos de avatares.

Este caso es un ejemplo muy extremo donde la única información que no detiene al atacante es la presencia o no del paciente en la cohorte. El requisito previo del conocimiento del registro original completo del paciente es particularmente atrevido, pero permite el análisis de la situación más desfavorable. En la gran mayoría de los casos, sin embargo, el atacante tiene información parcial sobre el paciente, es decir, solo algunos atributos, y busca recopilar otros nuevos.

La figura 6 muestra la evolución de la distancia entre un avatar y el registro original más cercano, en función del número k de vecinos más cercanos. Esta distancia permite evaluar el riesgo de reidentificación en el contexto de ataques de proximidad, en los que se busca establecer el vínculo entre los datos del avatar y los datos iniciales, disponibles para el atacante en este escenario.

La Figura 7 representa la evolución de las distancias entre el vecino más cercano y una persona, en el grupo de datos iniciales, en función del número k de vecinos más cercanos. La distancia entre los vecinos más cercanos es importante para apreciar la densidad de un grupo de datos. Al comparar con la figura anterior, se observa que para k igual a 50, la distribución de la densidad del grupo de datos del avatar es comparable a la obtenida para el grupo de datos original.

Por supuesto, la invención no se limita a los ejemplos que se acaban de describir. En particular, se pueden utilizar otros métodos de análisis multivariante o sorteos aleatorios, así como otros medios de descarga y carga de datos. Estos dos ejemplos permiten subrayar la primacía del riesgo de reidentificación sobre la conservación de la señal en la invención, lo que la diferencia de otros métodos de simulación de datos conocidos.

Aplicaciones de la invención

La invención se puede utilizar en aplicaciones de intercambio de datos con fines analíticos. La avatarización abre la posibilidad de potenciar los datos sensibles en su granularidad individual, no limitada a los datos sensibles, con actores, internos o externos. El propósito de compartir estos datos es utilizar el potencial de la información contenida en los grupos de datos para servir intereses tanto en un aspecto retrospectivo como en una necesidad predictiva. Las técnicas de inteligencia artificial no pueden condenarnos a elegir entre la liberalización de los datos sensibles en detrimento de la protección de las personas y los beneficios colectivos que de ello se derivan.

Un establecimiento de salud pública, como un hospital, tiene una base de datos con información valiosa. Estos últimos representan un desafío para la investigación médica, pero compartirllos se vuelve laborioso debido a restricciones regulatorias legítimas. La invención permite, mediante la creación de avatares, tanto preservar la calidad de los datos como respetar la confidencialidad de los datos del paciente. De este modo, se puede dar una segunda vida a los datos de salud, cuyo intercambio mejorará la toma de decisiones, las prácticas y la investigación.

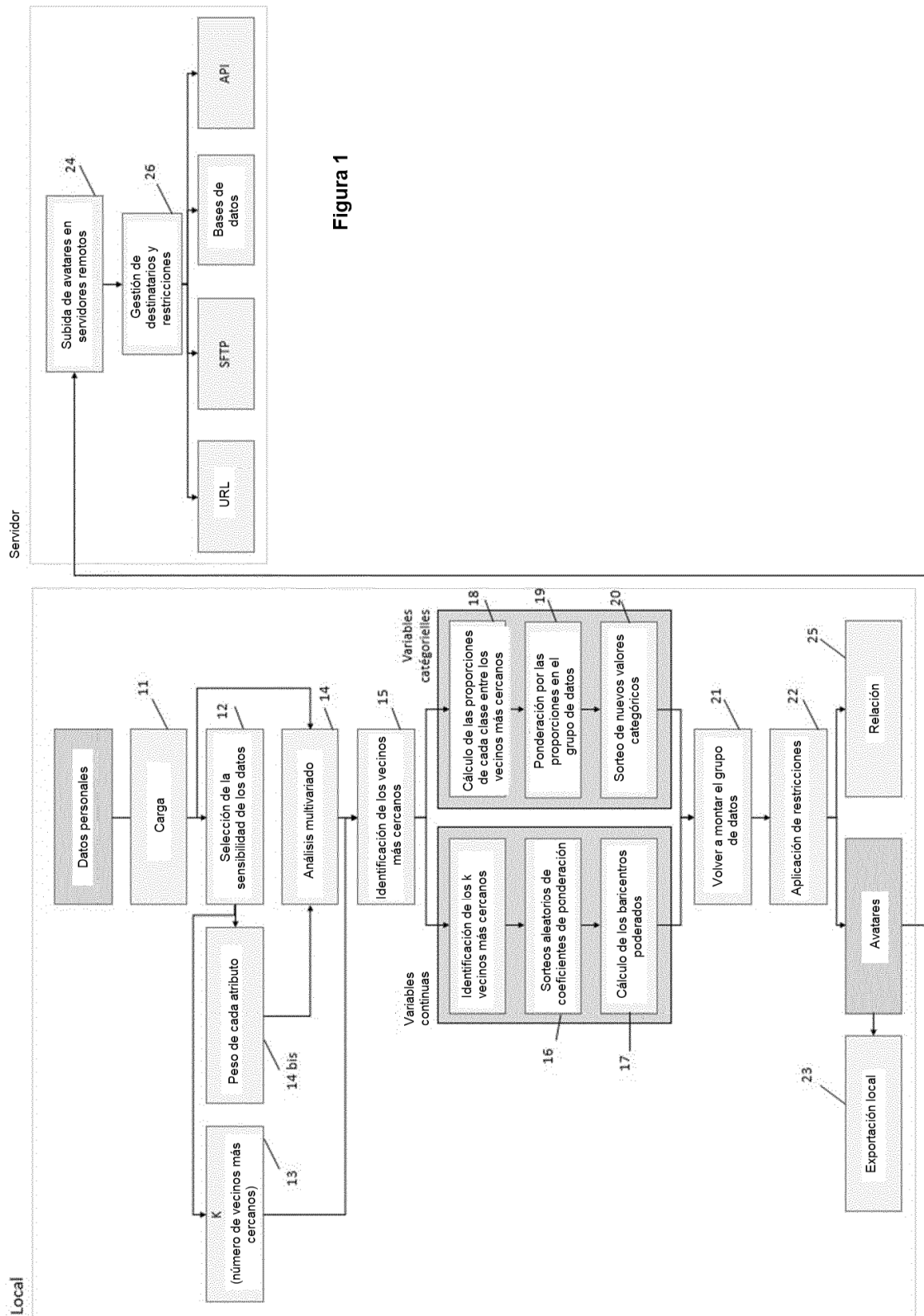
Un médico puede utilizar una recopilación de datos sobre pacientes que padecen una determinada patología como guía para aclarar su decisión, por ejemplo, para elegir entre diferentes tratamientos, según la edad y el estilo de vida del paciente. Una estructura de salud pública también podría crear avatares a partir de sus datos para predecir las necesidades necesarios para mejorar sus misiones y alertar de manera temprana a las autoridades públicas para llevar a cabo una campaña de sensibilización y planificar los recursos necesarios, por ejemplo, en el caso de una epidemia.

- Además, un municipio podría crear avatares a partir de los datos de su territorio para encomendárselos a un proveedor de servicios para detectar zonas o poblaciones que requieran una atención especial. Los avatares permiten valorizar datos con potencial, que en este caso sería una forma efectiva de prevenir la fractura social y el acceso a la atención. Esto también podría permitir orientar eficazmente las acciones preventivas e incluso establecer, a través de la predicción, en particular mediante el uso de técnicas de inteligencia artificial, una política que promueva la igualdad en los cuidados.
- Se podría imaginar, por ejemplo, la planificación de campañas de vacunación. Un municipio aún podría utilizar la información de salud de su población para planificar y desarrollar su política de desarrollo de espacios verdes, favoreciendo así los lugares que lo necesitarían de manera factual, o bien, adaptar su política en términos de alternancia de tráfico en el marco de las medidas contra la contaminación.
- Una empresa de servicios de suscripción que permita la transmisión de películas y series de televisión podría publicar avatares de sus clientes en Datos Abiertos para permitir que los productores de películas y series destaquen los perfiles de los clientes y así refinar su oferta.
- La perspectiva del intercambio de datos se puede considerar internamente para mejorar el acceso a los datos o externamente.
- La invención puede utilizarse para ayudar a calcular puntuaciones de riesgo en perfiles de clientes con un nivel de detalle nunca antes alcanzado, en particular por compañías de seguros o mutuas de seguros. Esta perspectiva se puede mejorar aún más con la posibilidad de cruzar bases de datos a través de avatares, lo que permite enriquecer las enseñanzas.
- La democratización de los grupos de datos es una oportunidad para la enseñanza y la calidad pedagógica, que permite mejorar las perspectivas de los enfoques pedagógicos. El acceso a grupos de datos en forma de avatares permite popularizar el acceso a los datos, sin tener que pasar por pasos regulatorios complejos y, en última instancia, limitantes. Los estudiantes pueden entrenarse con grupos de datos de avatares de cohortes reales, conservando las propiedades estadísticas.
- La invención se puede utilizar en aplicaciones de intercambio de datos con ánimo de lucro, por ejemplo, en el marco de una prestación de servicios remunerada. Una empresa farmacéutica interesada en una patología podría utilizar los datos del avatar del paciente para mejorar el desarrollo de sus fármacos.
- Los grandes minoristas podrían aprovechar los datos de los recibos para establecer valiosos recorridos del consumidor con los proveedores, incluidos, por ejemplo, sus hábitos, frecuencia y tiempo.
- El intercambio de datos se puede hacer como parte de la mejora de la estrategia y la gestión interna de una empresa. Proteger la privacidad de los clientes es una obligación para las empresas u organizaciones. Tener en cuenta la privacidad es garantía de seriedad para una empresa: de esta depende su reputación e imagen, y por tanto la fidelización de sus clientes a largo plazo. Para limitar el impacto de una posible falla informática, una empresa puede utilizar en la producción una base de datos de avatares, generada de acuerdo con la invención, y mantener los datos sensibles originales con total seguridad.
- La invención se puede utilizar en aplicaciones de intercambio de datos con el fin de desarrollar aplicaciones. Para desarrollar aplicaciones, hacerlas evolucionar o mantenerlas, es necesario disponer de suficientes datos. También es esencial que los datos de prueba tengan estructuras idénticas a los datos reales, por ejemplo, en un sitio web de un banco, para permitir el desarrollo de una nueva aplicación de pago bancario, el proveedor de servicios necesita grupos de datos que reflejen datos reales. Sin embargo, ya sea para proveedores de servicios o servicios internos dedicados, es esencial que los datos utilizados para este fin no sean datos sensibles. Los avatares según la invención cumplen este papel.
- La invención, con la creación de avatares, permite mantener el potencial de la información contenida en los datos sensibles y no tener que borrar los datos (para cumplir con el RGPD). Las empresas tienen interés en conservar la información estadística contenida en los grupos de datos, que representan un interés estratégico. Una empresa de gestión de carreteras podría mantener los avatares de datos de los conductores durante un período de tiempo ilimitado.
- También es posible que los avatares encuentren un papel en los procedimientos de gobierno, cuyas limitaciones aligeran. Se podría imaginar una gobernanza pragmática del uso de datos donde la prueba de concepto parcial en grupos de datos de avatares sería una parte integral de la autorización de acceso a los datos originales y el despliegue de algoritmos.

REIVINDICACIONES

- 5 1. Método para crear avatares a partir de un grupo de datos sensibles iniciales almacenados en una base de datos de un sistema informático, dichos datos iniciales comprenden atributos relativos a una pluralidad de personas, el método es del tipo identificativo de los vecinos más cercanos de una persona entre las otras personas del grupo de datos iniciales, caracterizado porque comprende
 - 10 a) la elección centrada en la persona consideradas, para atributos relativos a esta persona, de un número (k) de vecinos más cercanos que se utilizarán entre el conjunto de personas del grupo de datos iniciales,
 - b) identificar, para esa persona, los k vecinos más cercanos entre las otras personas en el grupo de datos,
 - c) generar, para al menos un atributo relativo a esta persona, un nuevo valor de atributo a partir únicamente de cantidades características de dicho atributo en dichos k vecinos más cercanos identificados, ponderados por un coeficiente,
 - 15 d) crear datos de avatares con los nuevos valores de atributo para hacer que los datos sensibles sobre la persona no sean identificables.
- 20 2. Método según la reivindicación 1, en donde el número (k) de vecinos más cercanos se elige en función de un factor de sensibilidad de los datos iniciales y/o de un factor de confianza en el destinatario destinado a recibir los datos de avatares, cuanto mayor sea el factor de sensibilidad y/o cuanto menor sea el factor de confianza, mayor será el número de vecinos más cercanos, y cuanto menor sea el factor de sensibilidad y/o cuanto mayor sea el factor de confianza, menor será el número de vecinos más cercanos.
- 25 3. Método según la reivindicación 1 o 2, en donde se realiza un análisis multivariante sobre los datos iniciales, previo a la identificación de los k vecinos más cercanos de la persona, en particular un análisis de componentes principales, los valores de atributo así modificados luego se proyectan en particular en el mismo espacio euclídiano.
- 30 4. Método según cualquiera de las reivindicaciones anteriores, en donde, los datos iniciales comprenden al menos un atributo que tiene un valor numérico, el nuevo valor del atributo corresponde a un baricentro con ponderación estocástica (G) calculado a partir de los valores de este mismo atributo de k vecinos más cercanos, cada uno ponderado por un coeficiente elegido aleatoriamente.
- 35 5. Método según la reivindicación anterior, en donde se elige aleatoriamente un coeficiente diferente para cada valor del atributo de cada uno de los k vecinos más cercanos.
6. Método según la reivindicación 4 o 5, en donde el baricentro con ponderación estocástica (G) se calcula en un número (N) de dimensiones correspondiente al número de atributos que tienen un valor numérico.
- 40 7. Método según cualquiera de las reivindicaciones anteriores, en donde, los datos iniciales comprenden al menos un atributo perteneciente a una clase representativa de una característica de la persona, el nuevo valor del atributo se elige aleatoriamente según la representación ponderada de dicha clase entre los k vecinos más próximos, calculados según el número de valores pertenecientes a dicha clase en los vecinos más cercanos ponderado por el número de valores que pertenecen a esta clase en todo el conjunto del grupo de datos iniciales.
- 45 8. Método según cualquiera de las reivindicaciones anteriores, en donde el número () de vecinos más próximos es variable, siendo diferente para algunas personas o de una persona a otra, para el mismo grupo de datos iniciales.
- 50 9. Método según cualquiera de las reivindicaciones anteriores, en donde se aplica al menos una función de restricción a los datos de avatares, con el fin de limitar los valores aberrantes o la alteración de la relación entre atributos relativos a una misma persona, en particular para restringir las distribuciones de los atributos numéricos, o para respetar una relación lógica entre los atributos de clase.
- 55 10. Método según cualquiera de las reivindicaciones anteriores, en donde los datos de avatares se registran en una base de datos del mismo sistema informático y/o de un servidor remoto.
- 60 11. Método según cualquiera de las reivindicaciones anteriores, en donde los destinatarios de los datos de avatares son notificados de la provisión de dichos datos por un sistema híbrido de autenticación que comprende al menos dos etapas, basado en particular en la generación y la transmisión de un enlace de descarga único, por ejemplo por correo electrónico, luego la comunicación de una clave con una vida útil limitada, por ejemplo por SMS.
- 65 12. Dispositivo para la creación de avatares a partir de un grupo de datos sensibles iniciales almacenados en una base de datos de un sistema informático, dichos datos iniciales que comprenden atributos relativos a una pluralidad de personas, el dispositivo es del tipo identificativo de los vecinos más cercanos de una persona entre las otras personas del grupo de datos iniciales, **caracterizado porque** comprende :

- a) un módulo de elección configurado para elegir, centrado en la persona considerada, un número () de vecinos más cercanos que se utilizarán entre el conjunto de personas del grupo de datos iniciales, para los atributos relacionados con esta persona,
- 5 b) un módulo de identificación configurado para identificar, para esta persona, los k vecinos más cercanos entre las otras personas en el grupo de datos,
- c) un módulo de generación configurado para, al menos para un atributo relativo a esta persona, generar un nuevo valor de atributo a partir únicamente de cantidades características de dicho atributo en dichos vecinos más cercanos identificados, ponderados por un coeficiente, y
- 10 d) un módulo de creación de avatares configurado para crear datos de avatares que comprenden el o los nuevos valores de atributo, para hacer que los datos sensibles relacionados con la persona no sean identificables.
13. Dispositivo según la reivindicación anterior, que comprende además un módulo de registro configurado para registrar los datos de avatares en una base de datos del mismo sistema informático y/o de un servidor remoto.
- 15 14. Dispositivo según la reivindicación 12 o 13, que comprende un módulo de alerta configurado para alertar a los destinatarios de los datos de avatares sobre la disponibilidad de dichos datos por un sistema híbrido de autenticación que comprende al menos dos pasos, basado en particular en la generación y la transmisión de un enlace de descarga único, por ejemplo por correo electrónico, luego la comunicación de una clave con una vida útil limitada, por ejemplo por SMS.
- 20 15. Producto de programa informático para implementar el método de creación de avatares a partir de un grupo de datos sensibles iniciales almacenados en una base de datos de un sistema informático, según cualquiera de las reivindicaciones 1 a 12, dichos datos iniciales que comprenden atributos relativos a una pluralidad de personas, el
- 25 producto de programa informático que comprende un medio y registros en este medio de las instrucciones legibles por un procesador para, cuando se ejecutan:
- a) un número () de vecinos más cercanos que se utilizarán entre el conjunto de personas en el grupo de datos iniciales se elige centrándose en la persona considerada, para los atributos relacionados con esta persona,
- 30 b) (b) para esa persona, los vecinos más cercanos se identifican entre las otras personas en el grupo de datos,
- c) para al menos un atributo relacionado con esta persona, se genera un nuevo valor de atributo únicamente a partir de cantidades características de dicho atributo en dichos vecinos más cercanos identificados, ponderados por un coeficiente, y
- 35 d) se crean datos de avatares que contienen los nuevos valores de atributo, para hacer que los datos sensibles relacionados con el individuo no sean identificables.



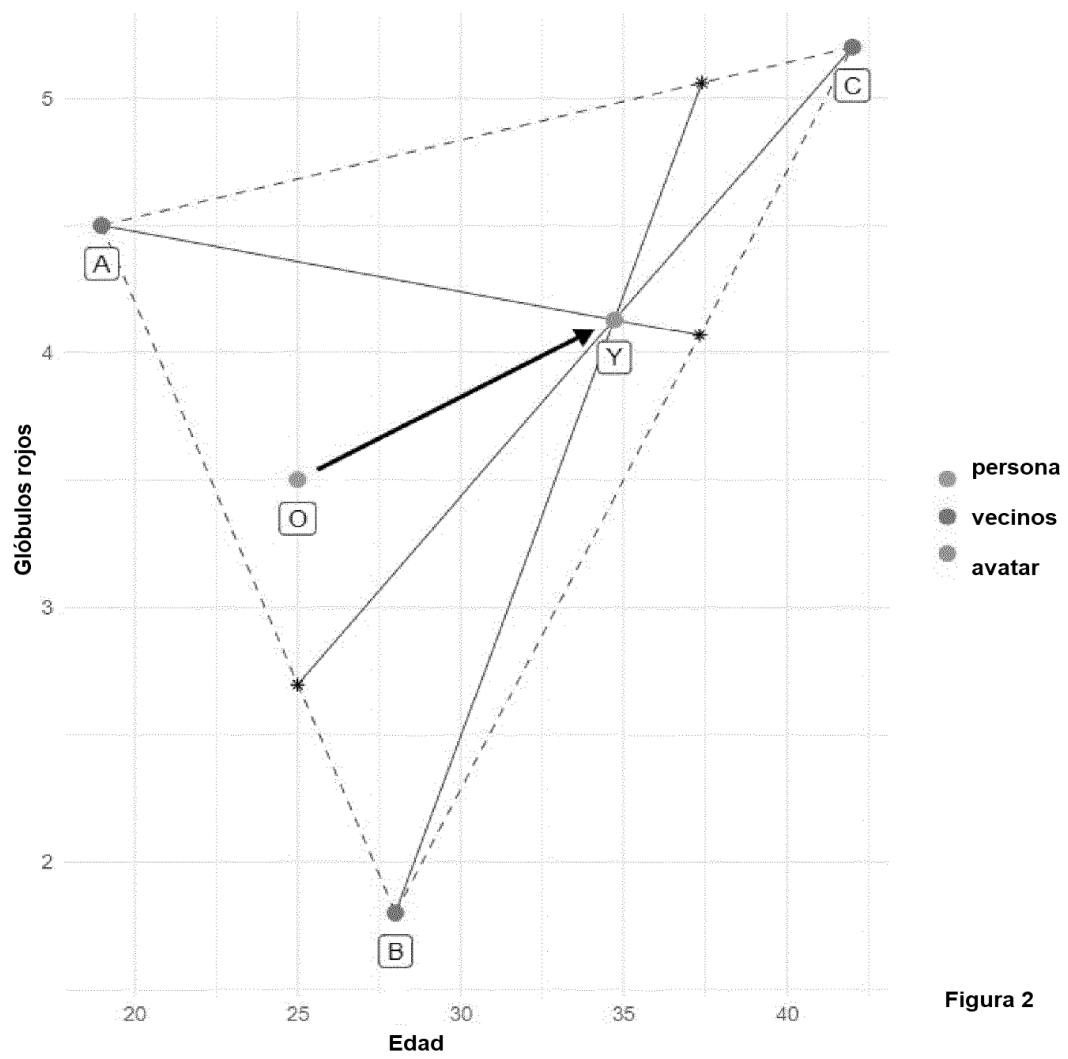


Figura 2

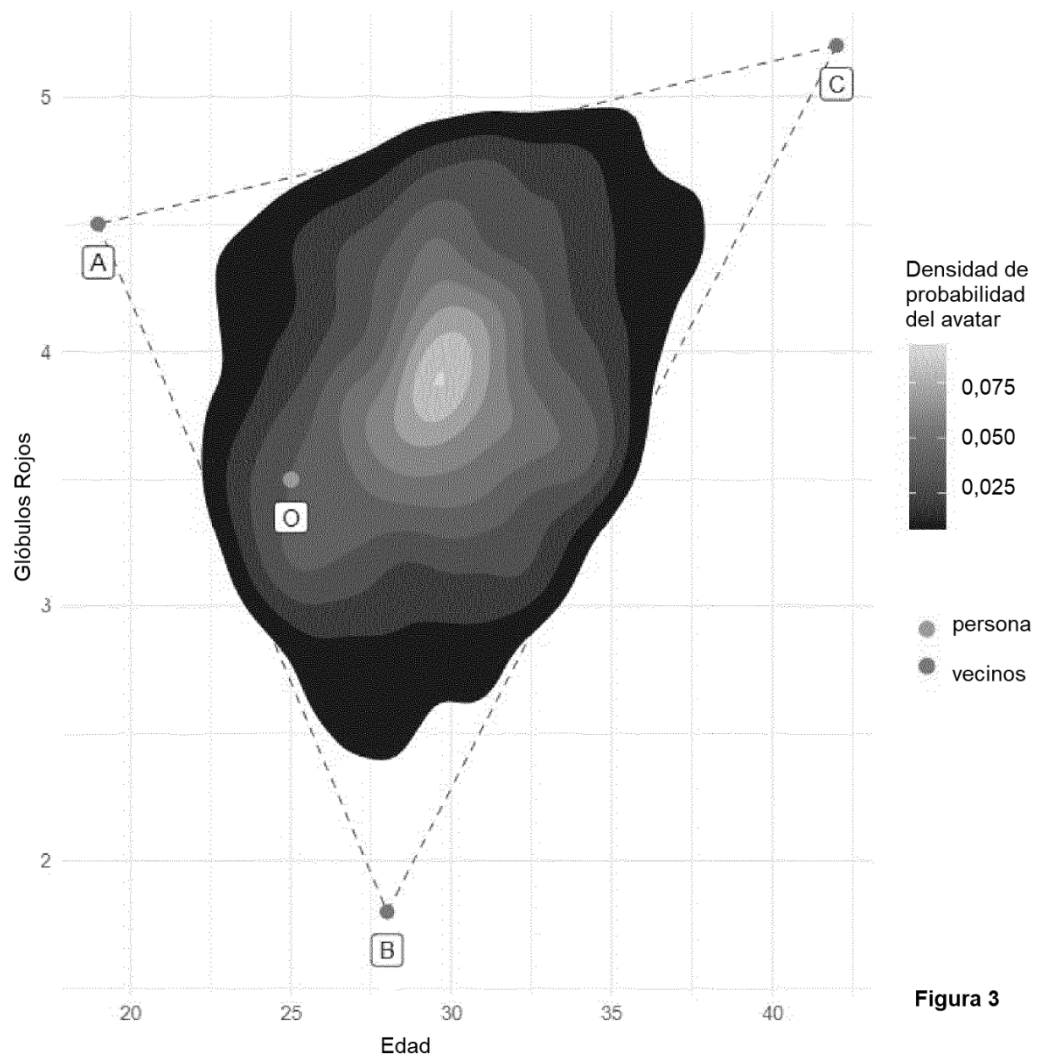
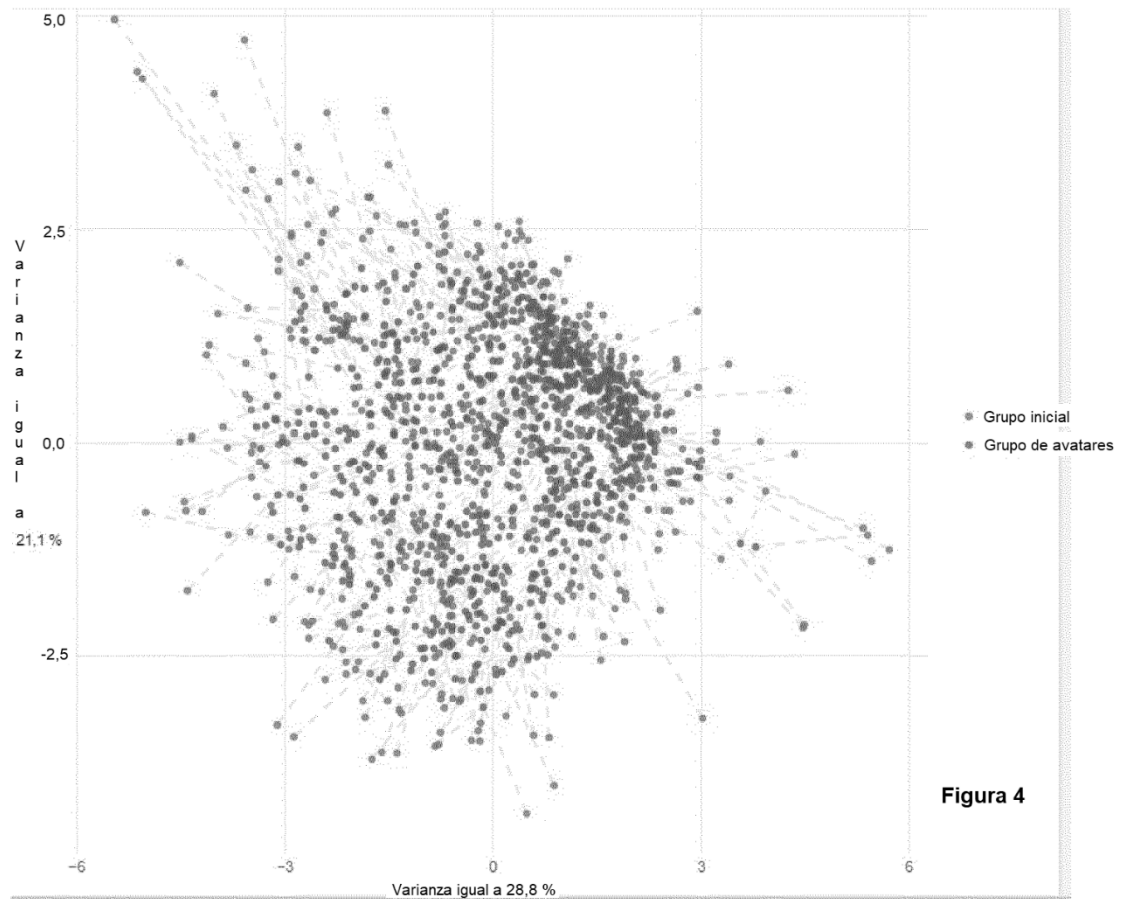
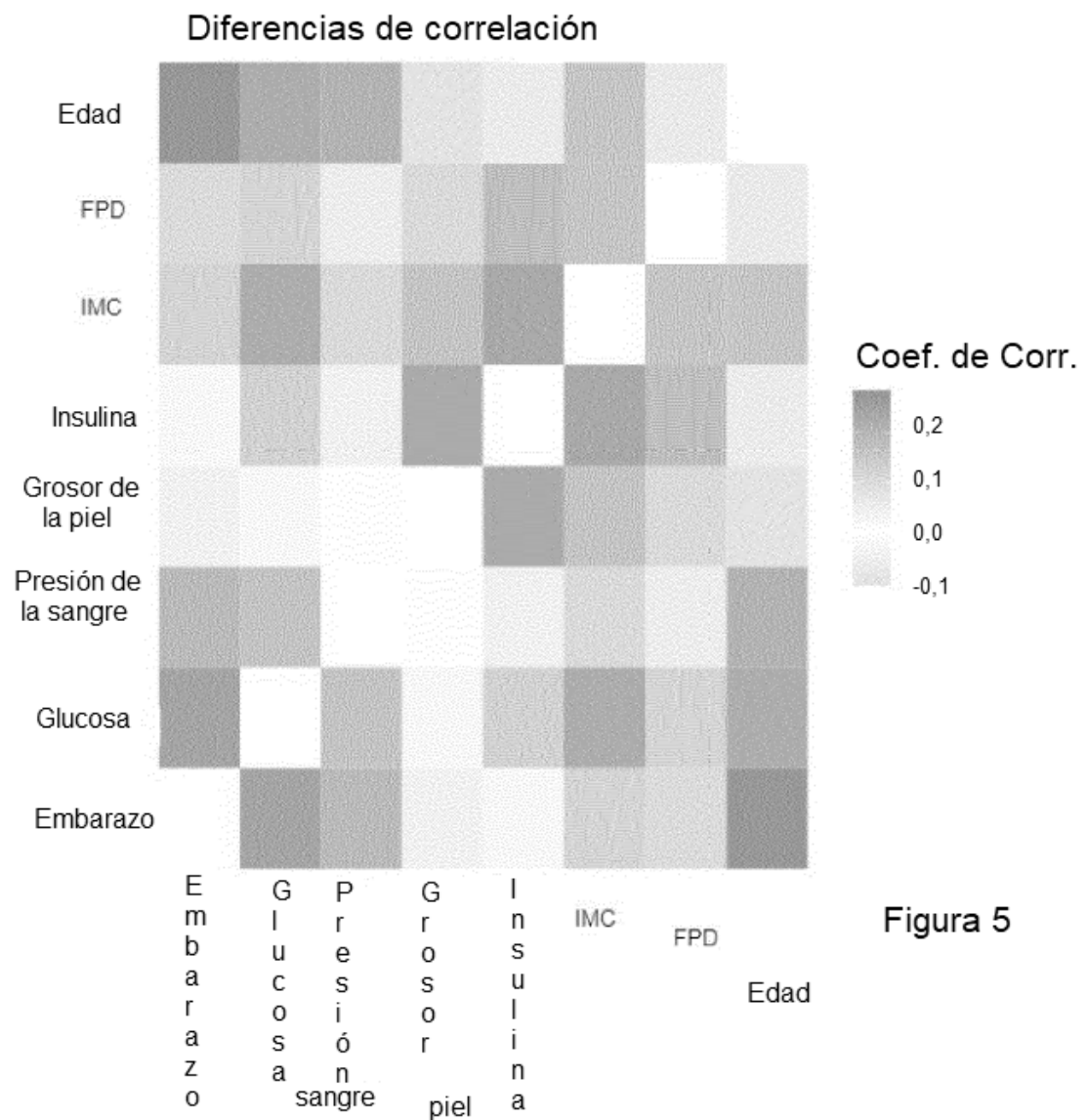
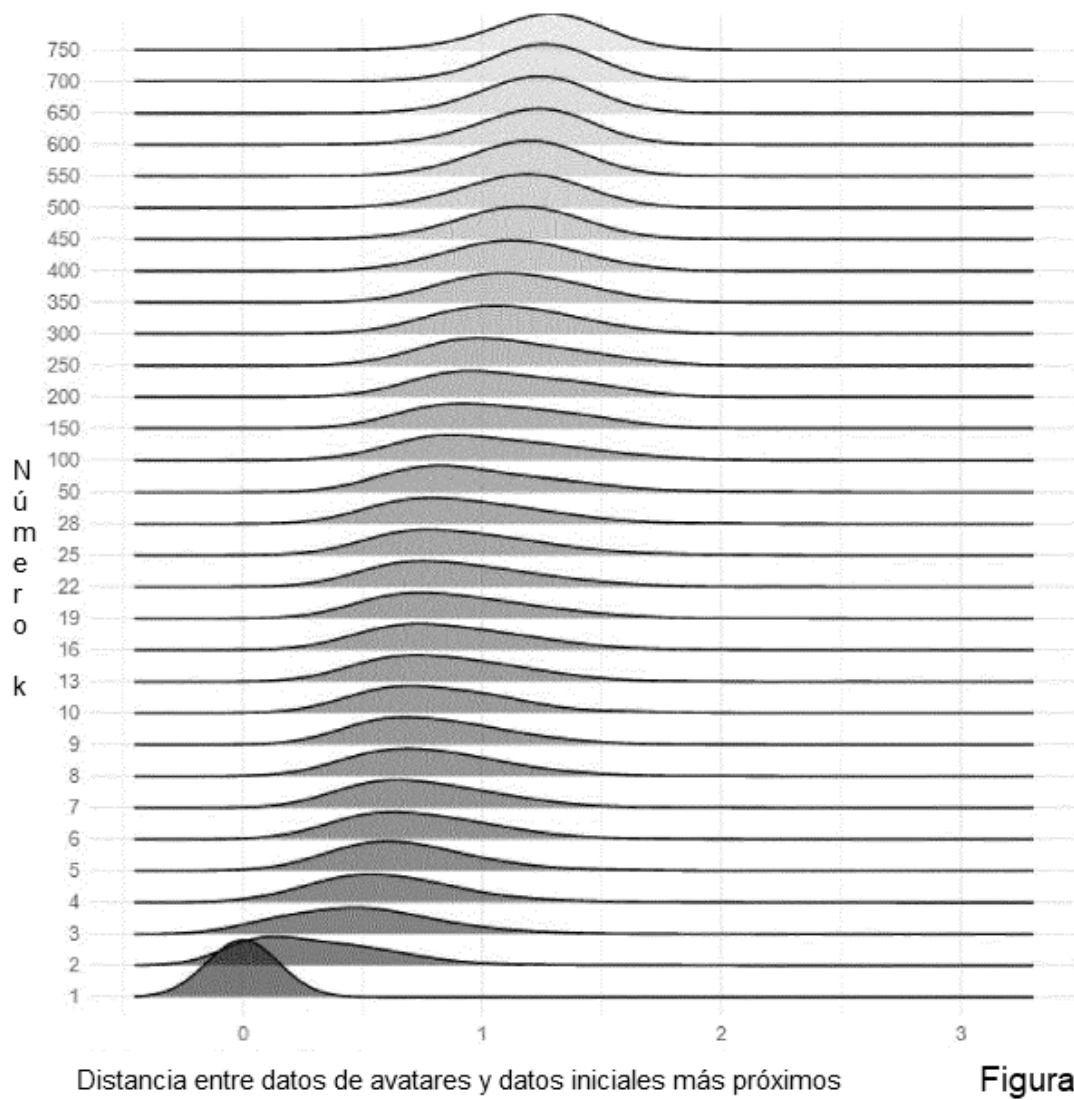


Figura 3







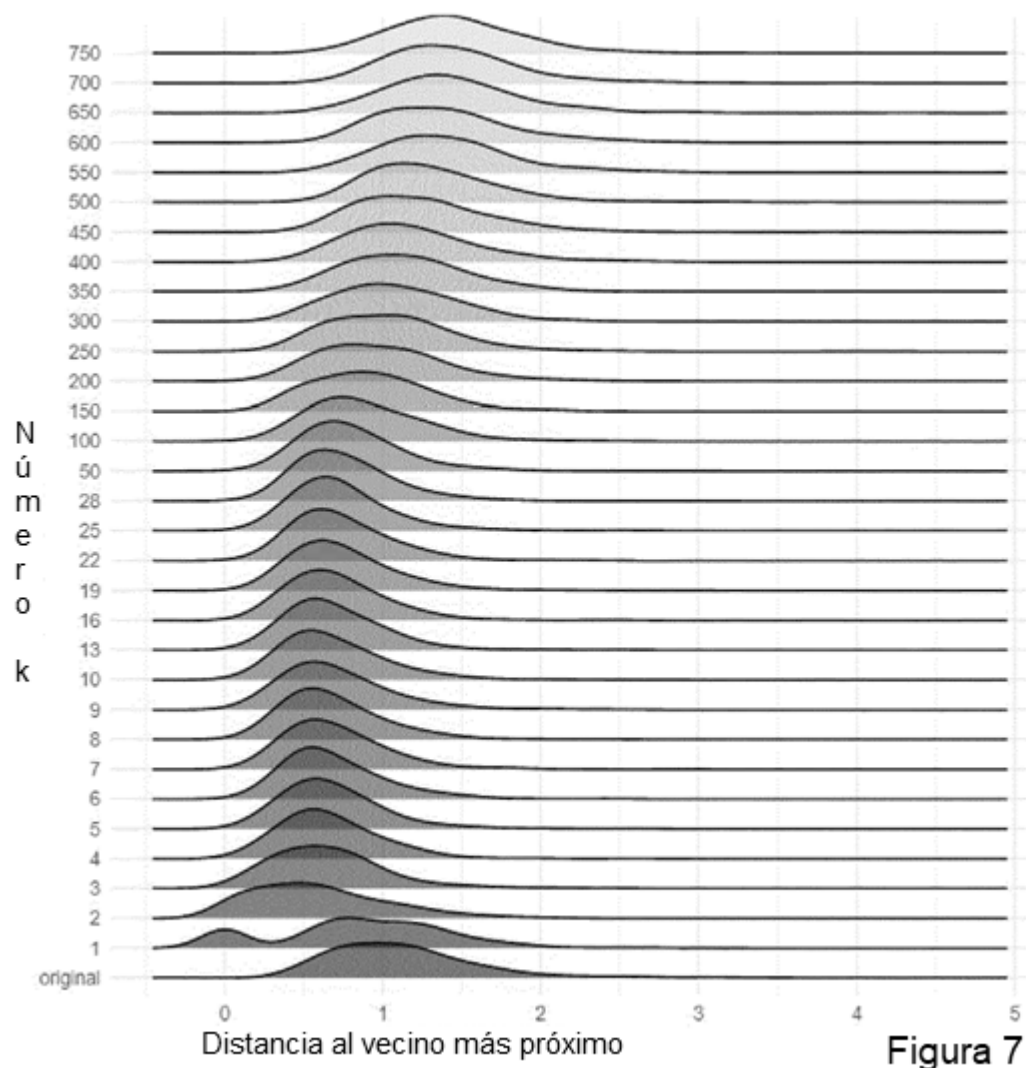


Figura 7