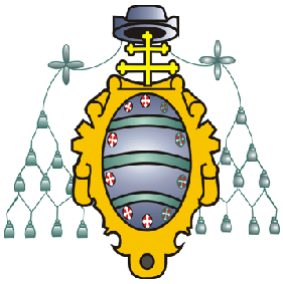

Universidad de Oviedo
Facultad de Ciencias

# Protección de la Privacidad en Microdatos

por

Pelayo Quirós Cueto

Trabajo Fin de Máster
bajo la dirección de los profesores
Pedro Alonso Velázquez
Irene Díaz Rodríguez y
Susana Montes Rodríguez

Julio, 2013

# Índice general

# Introducción

La difusión de información es quizá la actividad más frecuente en la era de la Sociedad de la Información, no sólo porque los avances informáticos han permitido el manejo de gran cantidad de bases de datos, sino también porque las plataformas para la difusión de dicha información han aumentado, así como la accesibilidad de la población a las mismas, de modo que se puede acceder casi instantáneamente a cantidades ingentes de datos, que posteriormente pueden ser utilizadas según convenga.

En este ámbito, muchos organismos públicos y privados divulgan datos de censos y encuestas, conocidos como microdatos (datos que no están resumidos en estadísticas sino que están directamente relacionados con individuos), mostrando así información sin conocimiento de los interesados. En muchas ocasiones la información publicada es irrelevante, pero si ésta se cruza con alguna otra fuente de información puede dejar al descubierto datos sensibles de un determinado individuo.

Por otro lado, la legislación de muchos países, en concreto la española (Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal), recoge el derecho de los ciudadanos a la privacidad y el control sobre los datos personales que obran en poder de terceros. La normativa más relevante en este aspecto es la *United States Healthcare Information Portability and Accountability Act* y la directiva de la Unión Europea 95/46/EC.

En este contexto, es necesario encontrar una solución de compromiso entre el derecho reconocido a la privacidad, es decir, el derecho al anonimato y el beneficio que produce la difusión de datos, que es obligatoria para

algunas administraciones públicas. Con el propósito de evitar que los individuos cuya información está publicada sean identificados unívocamente, las organizaciones habitualmente eliminan identificadores explícitos, como su número de identificación (el DNI en España), o el número de la Seguridad Social. Sin embargo, aunque algunas veces la información publicada parece anónima, la privacidad de los usuarios puede verse comprometida de forma involuntaria, sobre todo en presencia de usuarios maliciosos, que crucen datos procedentes de diversas fuentes.

La teoría del Control de la Revelación de Datos (*Statistical Disclosure Control* en inglés) se centra precisamente en buscar soluciones para este problema (ver [12]). Es decir, trata de proteger la confidencialidad de los posibles riesgos a los que está expuesta la información publicada sobre individuos. Entre los diferentes riesgos se encuentra el hecho de que cierta información sensible (y protegida, al menos teóricamente) se vea comprometida si se utiliza adecuadamente (o maliciosamente) cierta información pública. Para evitar este tipo de situaciones se han desarrollado distintas estrategias.

Tradicionalmente, los paradigmas para preservar el anonimato agrupan los individuos en clases de equivalencia utilizando diversas estrategias. Las técnicas más habitualmente utilizadas son: $k$-anonymity, $l$-diversity y $t$-closeness [1], siendo éstas las que se han considerado en este trabajo.

En un primer lugar se desarrolla la $k$-anonymity (ver [30] y [32]), donde para cada individuo existen al menos $k-1$ individuos indistinguibles en los datos. Otros métodos más complejos se han desarrollado para alcanzar tablas $k$-anónimas (ver [24]).

Dicha técnica tiene ciertos problemas con ataques de homogeneidad, por lo que se desarrolla la $l$-diversity (ver [16]), garantizando la existencia de al menos $l$ valores bien representados en cada grupo de $k$ individuos indistinguibles.

La $l$-diversity también presenta algunos problemas, como es el caso de

[1] Utilizaremos la denominación de estas téćnicas en inglés, por ser la que habitualmente aparece en la literatura, no existiendo ninguna referencia conocida en la que aparezcan traducidas.

la similitud de valores. Recientemente, se desarrolló la técnica $t$-closeness (ver [23]), la cual establece que la distribución del atributo sensible en cada clase de equivalencia ha de ser similar a la de los datos completos.

Todas las técnicas descritas pueden sufrir ataques que lleven a identificar a alguno de los individuos de la base de datos. El objetivo de este Trabajo Fin de Máster es usar conjuntos difusos como una vía para mejorar la protección de la privacidad de los microdatos, aprovechando las propiedades que este tipo de conjuntos presenta.

El trabajo está estructurado del siguiente modo: en el Capítulo 1 presentaremos el estado del arte, para a continuación, en el Capítulo 2, describir los conceptos básicos, tanto sobre las técnicas de privacidad, como sobre los conjuntos difusos, sus propiedades y las distintas formas de cardinalidad que tenemos en dichos conjuntos. En el Capítulo 3, incorporamos los conjuntos difusos a la privacidad, mejorando dicha protección. En el Capítulo 4, se presenta la experimentación de las estrategias propuestas llevada a cabo sobre bases de datos reales. Finalmente, se presentan las conclusiones alcanzadas, incluyendo las diferentes líneas abiertas en el tema, que permitirían continuar con la investigación en esta dirección.

# Capítulo 1

# Estado del Arte

Se han desarrollado numerosas técnicas para valorar el riesgo de comprometer el anonimato al publicar datos. Como se ha mencionado en la introducción, los primeros paradigmas usados para proteger la confidencialidad se basan en agrupar registros (técnica conocida como $k$-anonymity) o valores sensibles ($l$-diversity) (ver [16], [30] y [32]). De hecho hay muchos trabajos que se centran en producir tablas de datos $k$-anónimas de una forma eficiente. Samarati y Sweeney (ver [30] y [32]) definen el concepto de $k$-anonymity como la propiedad que hace un registro (un individuo) indistinguible de al menos otros $k-1$ individuos. Existen una gran cantidad de técnicas destinadas a obtener tablas $k$-anónimas de forma eficiente; desde las iniciales propuestas (ver [30] y [32]) a otras más elaboradas. Por ejemplo, Matatov en [24] presenta una técnica basada en un algoritmo de descomposición que utiliza a su vez un algoritmo genérico para buscar particiones óptimas.

La $k$-anonymity es capaz de prevenir y cuantificar el riesgo de revelar la identidad de un usuario, pero las tablas $k$-anónimas están desprotegidas frente al riesgo de averiguar características concretas de determinados grupos de usuarios mediante ataques de homogeneidad porque puede producir grupos de individuos indistinguibles, pero con el mismo valor para los atributos sensibles (ver [16]). Un ejemplo claro sería el siguiente: imaginemos que el Servicio de Salud del Principado de Asturias publica la relación de

usuarios atendidos por especialidades durante el año 2012, y que para evitar conocer la identidad de los enfermos, enmascara de algún modo los datos publicados, asegurándose que haya al menos $k$ individuos indistinguibles porque sus características (edad, sexo, valores de analíticas, etc.) coinciden. Si hay un sólo grupo de $k$ individuos indistinguibles que tengan la misma enfermedad (diabetes, por ejemplo), habremos adivinado un valor sensible para los $k$ individuos que forman el grupo, vulnerando así el derecho de los enfermos a mantener en la esfera de su intimidad la dolencia que padecen. Las tablas $k$-anónimas tampoco están exentas de ataques basados en conocimiento a priori.

Para intentar atajar estas carencias, en [16] los autores proponen otra técnica, la $l$-diversity. Esta técnica exige que la distribución de los atributos sensibles en cada grupo de $k$ elementos indistinguibles, tenga al menos $l$ valores diferentes. Esta técnica ha presentado con posterioridad diferentes variantes, como la *diversidad distinta*, que exige que al menos $l$ valores distintos de los datos ocurran en cada clase de equivalencia o la *diversidad basada en la entropía*, que requiere que la entropía de la distribución de los valores sensibles sea mayor o igual que $log(n)$. También conviene considerar la *diversidad recursiva*, que impone que el valor del atributo sensible más frecuente, no aparezca muy frecuentemente, ni el menos frecuente demasiado poco.

Sin embargo, la $l$-diversity también presenta carencias. La más relevante es que no soporta ataques basados en sesgos y similitudes. En efecto, cuando la distribución del atributo sensible es sesgada, el hecho de que un individuo pertenezca a una clase de equivalencia puede hacerlo más vulnerable que antes. Por otro lado, cuando se tienen en cuenta atributos numéricos, la técnica $l$-diversity no tiene en cuenta que hay valores que, aunque diferentes, pueden ser similares.

Más recientemente, en [23] se introduce la necesidad de considerar la similitud semántica que puede existir entre determinados valores para solventar los puntos débiles que presentan los métodos precedentes. De este modo introduce la técnica denominada $t$-closeness, que establece que la distribución de un atributo sensible en cualquier clase de equivalencia es próxima a la distribución del atributo en los datos de partida. Esta proximidad

se mide mediante la distancia *Earth Mover's Distance* (EMD), que tiene en cuenta sólo el orden de los valores y no la similitud entre los mismos. Por ejemplo, si tenemos dos individuos de 23 y 25 años, respectivamente, se consideran igual de próximos en edad que dos que tengan 40 y 60, siempre y cuando el número de individuos con edades comprendidas entre ellos sea el mismo. Además, la métrica EMD no se puede aplicar en todos los casos, como por ejemplo ocurre cuando se trabaja con información estructurada, ya que en este caso no es fácil establecer una ordenación.

Las herramientas citadas anteriormente son las más utilizadas para proteger y estudiar la privacidad de los usuarios cuando se publica información que les afecta. Sin embargo, existen otras muchas aproximaciones. Por ejemplo en [21] los autores introdujeron el concepto de $(l, \alpha)$-diversity, que requiere que el peso total de los valores sensibles en una clase de equivalencia, sea al menos $\alpha$, estando este peso controlado por un sistema recursivo.

Malin establece en [25] un modelo formal de protección de la información llamado $k$-unlinkability, para prevenir la reidentificación de al menos $k$ datos. Sin embargo, no representa un gran avance dado que un conjunto de datos $k$-anónimos satisfacen la propiedad de $k$-unlinkability, pero el recíproco no es cierto. Otros trabajos destacados en este ámbito se pueden ver en [6], donde se introduce una estrategia de supresión de datos para obtener privacidad, o en el trabajo desarrollado por Zhong en [35], que estudia diversos aspectos de privacidad relacionados con elementos frecuentes. También se pueden encontrar aproximaciones relacionadas en [11], [3], [31] o [28].

Como se puede observar, todas las técnicas mencionadas presentan algún punto débil: o no tienen en cuenta cuestiones de homogeneidad (por ejemplo la $k$-anonymity), o de similitud de valores (por ejemplo la $l$-diversity), o incluso no son intuitivas o generalizables para cualquier tipo de atributos (por ejemplo la $t$-closeness), por lo que la investigación en este campo es bastante prometedora.

La teoría de los conjuntos difusos proporciona un marco natural para analizar similitudes entre datos generalizados mediante conjuntos difusos (ver [8]). Dado que un método para publicar datos protegiendo la privacidad es precisamente agrupar individuos en clases mediante la generalización de los atributos que los caracterizan, nos planteamos en este trabajo utilizar

conjuntos difusos como herramienta de generalización, ya que proporcionan un modo natural de expresar dicha generalización. El procedimiento no es una simple generalización, puesto que el trabajar con datos difusos requiere adaptar la cardinalidad de los conjuntos a dicho caso (ver [10]).

Es necesario analizar las distintas definiciones de cardinalidad de un conjunto difuso que existen (ver [33]). Dichas definiciones se pueden clasificar en escalares (ver [14] y [34]), que asocian a cada conjunto difuso una cantidad, y difusas (ver [4], [19] y [27]), que asocian a cada conjunto difuso una función sobre los números naturales en el intervalo unidad. La aproximación utilizada en este trabajo será precisamente esta última.

# Capítulo 2

# Privacidad y Conjuntos Difusos

En este capítulo introduciremos aquellos conceptos, en la mayoría de los casos ya conocidos, que son necesarios para el desarrollo del resto del trabajo. Dichos conceptos se estructurarán en dos secciones, de acuerdo con los dos temas que vamos a aunar en el siguiente capítulo: privacidad y conjuntos difusos. Las definiciones básicas sobre privacidad que ahora presentamos han sido obtenidas, fundamentalmente, de [7], [22] y [16]. En el caso de los conjuntos difusos, una versión ampliada de los conceptos aquí expuestos puede verse, por ejemplo, en [13] o [15].

## 2.1. Privacidad

Tal como hemos señalado, nuestro objetivo será proteger la información de posibles filtraciones (ataques), de manera que los datos de aquellos individuos con los que se han elaborado los resultados, sean privados o, lo que es lo mismo en este contexto, estén protegidos. Comenzaremos con los conceptos básicos y a continuación revisaremos algunas de las principales técnicas que permiten alcanzar el objetivo planteado, así como los inconvenientes asociados a las mismas.

**Definición 2.1** *Sea* $T = \{t_1, \ldots, t_n\}$*, con atributos* $A_1, \ldots, A_m$*, la tabla*

*original de los datos a considerar. En este contexto se supone que $T$ es un subconjunto de una población mayor $\Omega$ y que cada tupla representa un individuo de la población. Los atributos a proteger los llamaremos* **atributos sensibles***, y los denotaremos por $S$, mientras que el resto, serán los* **atributos no sensibles***, y los denotaremos por $Q$.*

**Ejemplo 2.2** *Consideremos la Tabla 2.1 que recoge la Enfermedad que padecen los individuos de una muestra, incluyendo sobre cada uno de ellos, además de la descripción de dicha Enfermedad, los siguientes atributos: el código ZIP, la Edad y el Salario. En este caso los atributos no sensibles serían precisamente estos tres: el código ZIP, la Edad y el Salario, mientras que el atributo sensible sería, la Enfermedad.*

| Individuo | ZIP | Edad | Salario(k) | Enfermedad |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3.5 | úlcera gástrica |
| 2 | 47602 | 22 | 3.5 | gastritis |
| 3 | 47678 | 27 | 5 | cáncer de estómago |
| 4 | 47905 | 43 | 7 | gastritis |
| 5 | 47979 | 52 | 7 | gripe |
| 6 | 47906 | 47 | 8 | bronquitis |
| 7 | 47973 | 36 | 9 | neumonía |
| 8 | 47607 | 32 | 10 | cáncer de estómago |
| 9 | 47906 | 55 | 9 | ataque al corazón |
| 10 | 47925 | 56 | 9 | ataque al corazón |
| 11 | 47923 | 61 | 10 | angina de pecho |
| 12 | 47923 | 67 | 15 | neumonía |

Tabla 2.1: Tabla T original.

*Formalmente, $S = \{Enfermedad\}$ y $Q = \{ZIP,\ Edad,\ Salario\}$.*

Dado el ejemplo previo, podemos observar como si se conoce a alguien que ha participado en dicha encuesta, y esta persona tiene por ejemplo, 29 años, se puede directamente identificar que es el primer individuo, y así saber que padece úlcera gástrica. Lógicamente esto no resulta conveniente, por lo que es necesario enmascarar los datos, siendo éste el objetivo primordial de todas las técnicas que vamos a ir analizando en esta sección.

**Definición 2.3** *Dado un conjunto de atributos $QI = \{Q_1, \ldots, Q_r\}$, se dice que es un* **cuasi-identificador** *de $Q$ si, y sólo si, es un subconjunto de atributos no sensibles, es decir, $QI \subseteq Q$.*

**Ejemplo 2.4** *Si volvemos a considerar los datos del Ejemplo 2.2, un cuasi-identificador sería, por ejemplo, el conjunto formado por Edad y Salario:*

$$QI = \{Edad, Salario\}.$$

Un concepto fundamental a la hora de enmascarar los datos, será el de partición de un atributo, entendida ésta como partición de los valores posibles del mismo. Vamos a ver a continuación la definición formal de dicho concepto.

**Definición 2.5** *Dado $A$ un atributo de la tabla $T$ y sea $D(A)$ el conjunto de los valores que puede tomar dicho atributo $A$. Llamamos* **partición** *de $D(A)$ a una familia de conjuntos $\{A_i \,/\, i \in I\}$ tal que:*

1. $A_i \cap A_j = \emptyset, \forall i \neq j$,
2. $\bigcup_{i \in I} A_i = D(A)$.

**Ejemplo 2.6** *Con los datos de la Tabla 2.1, podemos hacer particiones del conjunto de valores que toma cada atributo no sensible, como sigue:*

- *Para el código ZIP, podemos tomar los dos conjuntos 476** y 479**, que agrupan a todos los códigos cuyas tres primeras cifras son las especificadas, y formar con ellos la partición $\{476^{**}, 479^{**}\}$.*
- *Para la Edad, podemos tomar una partición por intervalos, por ejemplo, $\{[0, 30], (30, 40], (40, 50], (50, \infty]\}$.*
- *Para el Salario, al igual que con la Edad, tomamos una partición por intervalos, por ejemplo, $\{[0, 5], (5, 7], (7, 9], (9, \infty]\}$.*

**Definición 2.7** *Dada una tabla original $T$, y dadas unas particiones de los distintos atributos verificando las condiciones de la Definición 2.5, se denomina* **tabla privada***, y se denota $PT$ (del inglés* Private Table*) a la tabla obtenida al sustituir para cada individuo y cada atributo en la tabla $T$ el valor de dicho atributo por el conjunto de la partición al que pertenece.*

**Ejemplo 2.8** *La tabla privada $PT$ que se muestra en la Tabla 2.2 es la que obtenemos para la tabla $T$ del Ejemplo 2.2 con las particiones consideradas en el Ejemplo 2.6.*

| Individuo | ZIP | Edad | Salario(k) | Enfermedad |
|---|---|---|---|---|
| 1 | 476** | [0,30] | [0,5] | úlcera gástrica |
| 2 | 476** | [0,30] | [0,5] | gastritis |
| 3 | 476** | [0,30] | [0,5] | cáncer de estómago |
| 4 | 479** | (40,50] | (5,7] | gastritis |
| 5 | 479** | (50,$\infty$] | (5,7] | gripe |
| 6 | 479** | (40,50] | (7,9] | bronquitis |
| 7 | 479** | (30,40] | (7,9] | neumonía |
| 8 | 476** | (30,40] | (9,$\infty$] | cáncer de estómago |
| 9 | 479** | (50,$\infty$] | (7,9] | ataque al corazón |
| 10 | 479** | (50,$\infty$] | (9,$\infty$] | ataque al corazón |
| 11 | 479** | (50,$\infty$] | (9,$\infty$] | angina de pecho |
| 12 | 479** | (50,$\infty$] | (9,$\infty$] | neumonía |

Tabla 2.2: Tabla privada obtenida a partir de T.

Tras estas particiones, y dada la tabla privada, vemos como el simple hecho de saber que uno de los participantes en la encuesta tiene 29 años, no nos indica, exactamente, cual es el problema que tiene, ya que entre las distintas opciones nos encontramos con úlcera gástrica, gastritis y cáncer de estómago.

Sin embargo, si la información de la que disponemos es que el individuo tiene 42 años, y conocemos que su Salario ronda los 6 000 euros, podemos identificarlo inmediatamente como el individuo 4, aún con la generalización que hemos hecho, y determinar, por tanto, que padece gastritis. Para tratar

de solucionar esto, se desarrolló la técnica llamada $k$-anonymity (véase [7] y [22]).

### 2.1.1. $k$-anonymity

**Definición 2.9** *Dada una tabla cualquiera $T$ se dice que* **satisface la condición k-anonymity** *si para cada tupla $t \in T$, existen otras $k-1$ tuplas $t_{i_1}, \ldots, t_{i_{k-1}} \in T$ tales que $t[C] = t_{i_1}[C] = \cdots = t_{i_{k-1}}[C]$, para todo cuasi-identificador $C \in QI$, donde $t[C]$ denota los valores que toma la tupla $t$ para el cuasi-identificador $C$.*

En otras palabras, una tabla cumple la condición $k$-anonymity si para cada tupla de dicha tabla, ésta es indistinguible entre otras $k-1$ tuplas respecto a los atributos no sensibles.

**Ejemplo 2.10** *Consideremos de nuevo la tabla $T$ del Ejemplo 2.2. Busquemos una tabla privada que cumpla la condición $k$-anonymity para algún $k \geq 2$ (evidentemente, dicha condición se verifica para $k = 1$ en cualquier tabla).*

*Si consideramos la tabla privada del Ejemplo 2.8, ésta no tendría la propiedad de $k$-anonymity para ningún $k \geq 2$, puesto que, como ya comentamos, la tupla $t_4$ correspondiente al cuarto individuo es distinguible de cualquiera de las otras.*

*El ejemplo buscado lo encontramos si se elimina el atributo Salario (o se generaliza a un único valor), se separa el código ZIP como hemos hecho en el Ejemplo 2.8, y para el atributo Edad se considera la partición $\{\leq 32, (32, 52], (52, \infty]\}$. Con esto, la tabla $T^*$ obtenida es la que se presenta en la Tabla 2.3.*

*Podemos ver, como con estas generalizaciones, nos encontramos con que cada tupla es indistinguible respecto a otras 3, es decir, la tabla obtenida cumple 4-anonymity.*

Tras aplicar el cuasi-identificador correspondiente a nuestra tabla $T$, será la tabla privada $PT$ la que será publicada.

| Individuo | ZIP | Edad | Enfermedad |
|---|---|---|---|
| 1 | 476** | $\leq 32$ | úlcera gástrica |
| 2 | 476** | $\leq 32$ | gastritis |
| 3 | 476** | $\leq 32$ | cáncer de estómago |
| 8 | 476** | $\leq 32$ | cáncer de estómago |
| 4 | 479** | $(32, 52]$ | gastritis |
| 5 | 479** | $(32, 52]$ | gripe |
| 6 | 479** | $(32, 52]$ | bronquitis |
| 7 | 479** | $(32, 52]$ | neumonía |
| 9 | 479** | $(52, \infty]$ | ataque al corazón |
| 10 | 479** | $(52, \infty]$ | ataque al corazón |
| 11 | 479** | $(52, \infty]$ | angina de pecho |
| 12 | 479** | $(52, \infty]$ | neumonía |

Tabla 2.3: Tabla generalizada cumpliendo 4-anoymity.

Utilizando la técnica anterior, el problema que se nos presentaba con las generalizaciones que hemos realizado en ejemplos anteriores desaparece. Sin embargo, pueden surgir otros problemas distintos, tales como:

- Encontrarnos en algún grupo de $k$ tuplas con el mismo valor para el atributo sensible. En el ejemplo anterior este problema se presentaría si para el grupo con los cuatro primeros individuos, todos tuviesen la misma enfermedad, es decir, si tuviéramos, por ejemplo, el caso mostrado en la Tabla 2.4.

| Individuo | ZIP | Edad | Enfermedad |
|---|---|---|---|
| 1 | 476** | $\leq 32$ | cáncer de estómago |
| 2 | 476** | $\leq 32$ | cáncer de estómago |
| 3 | 476** | $\leq 32$ | cáncer de estómago |
| 8 | 476** | $\leq 32$ | cáncer de estómago |

Tabla 2.4: Bloque cuyo valor sensible es igual para todas las tuplas indistinguibles.

  En este caso nos encontraríamos que tener la información de que un individuo tiene un código ZIP que empieza por 476, y que su Edad no

supera los 32 años, nos serviría para asegurar que dicho individuo ha sido diagnosticado de cáncer de estómago.

- Información adicional que pueda tener el atacante que intenta obtener datos. Si dicha persona quiere saber la enfermedad de alguien cuyo código ZIP empieza por 479 y tiene una edad superior a 52 años, existen tres opciones posibles, ataque al corazón, angina de pecho o neumonía. Sin embargo, si tiene un dato adicional relativo a que esa persona no tiene ningún problema relacionado con el corazón, descartaría inmediatamente tanto el ataque al corazón como la angina de pecho, y obtendría que la enfermedad que padece es neumonía. Para intentar paliar dichos problemas, surge la llamada $l$-diversity (ver [16]).

### 2.1.2. $l$-diversity

Recordemos que habíamos representado por $Q$ al conjunto de todos los atributos no sensibles. Además, supondremos la existencia de una distribución de probabilidad que mide el conocimiento del que dispone el atacante que intenta extraer la información acerca de otro individuo que haya participado en la encuesta con la que se ha realizado la tabla $T$.

**Definición 2.11** *El* **conocimiento a priori** *de que el individuo tenga como valor del atributo sensible $s$, supuesto que tiene como valor para el atributo no sensible $q$, se denota por $\alpha_{(q,s)}$ y viene dado por:*

$$\alpha_{(q,s)} = P(t[S] = s | t[Q] = q).$$

**Definición 2.12** *El* **conocimiento a posteriori**, *a través de la generalización $T^*$, de que el individuo tenga como valor del atributo sensible $s$, supuesto que tiene valor para el atributo no sensible $q$, se denota por $\beta_{(q,s,T^*)}$ y viene dado por:*

$$\beta_{(q,s,T^*)} = P_f(t[S] = s | t[Q] = q \wedge \exists t^* \in T^*, t \to t^*).$$

La siguiente definición proporciona una manera de calcular la información a posteriori de la que dispone el atacante.

**Definición 2.13** *Sea $q$ un valor del atributo no sensible $Q$ de la tabla $T$, y $q^*$ el valor generalizado de la tabla $T^*$. Sea $s$ un posible valor del atributo sensible, y $n_{(q^*,s')}$ el número de tuplas $t^* \in T^*$, donde $t^*[Q] = q^*$ y $t^*[S] = s'$. Además, sea $f(s'|q^*)$ la probabilidad condicionada de que el atributo sensible sea $s'$ condicionado a que el atributo no sensible tenga como generalización $q^*$. Entonces, siendo $\beta_{(q,s,T^*)}$ el conocimiento a posteriori dado que el individuo tenga un atributo sensible $s$ y un atributo no sensible $q$, a través de la generalización $T^*$, se tiene que:*

$$\beta_{(q,s,T^*)} = \frac{n_{(q^*,s)} \dfrac{f(s|q)}{f(s|q^*)}}{\displaystyle\sum_{s' \in S} n_{(q^*,s')} \frac{f(s'|q)}{f(s'|q^*)}}. \tag{2.1}$$

Los problemas asociados a la $k$-anonymity, descritos a continuación del Ejemplo 2.10, pueden ser formalizados utilizando los conceptos *"positive disclosure"* y *"negative disclosure"*, que se introducen a continuación.

**Definición 2.14** *Dada la tabla $T^*$ obtenida a partir de la generalización de $T$, decimos que:*

1. *Se tiene* **"positive disclosure"** *si el atacante puede identificar el valor del atributo sensible con una alta probabilidad.*
2. *Se tiene* **"negative disclosure"** *si el atacante puede eliminar algunos posibles valores del atributo sensible con una alta probabilidad.*

**Ejemplo 2.15** *Supongamos que los datos de los que dispone el atacante le lleva a las 3 tuplas de la Tabla 2.5.*

*Si el atacante sabe que el individuo lleva una muñeca vendada, el conocimiento que tiene le permite saber con alta probabilidad que el atributo*

| Individuo | ZIP | Edad | Enfermedad |
|---|---|---|---|
| 1 | 476** | $\leq 32$ | fractura de muñeca |
| 2 | 476** | $\leq 32$ | gastritis |
| 3 | 476** | $\leq 32$ | neumonía |

Tabla 2.5: Ejemplo para disclosure.

*sensible es fractura de muñeca, y por tanto estamos ante un caso de "positive disclosure". Si lo que sabe es que dicho individuo sale a correr todas las mañanas, puede eliminar la opción de neumonía con alta probabilidad, y por tanto, nos encontramos ante "negative disclosure".*

Es inmediato observar que, generalmente, la "positive disclosure" es más peligrosa que la "negative disclosure", ya que ésta última puede presentarse cuando simplemente se eliminan opciones poco importantes. Véase, por ejemplo, enfermedades que sean muy poco comunes en ciertas zonas del mundo, como puede ser el ébola en Europa. Si el individuo del que se pretende obtener información vive en una zona alejada de los núcleos del ébola, y no ha tenido contacto con dichos lugares, es lógico que esta opción sea descartada.

De todo lo descrito anteriormente, surge el siguiente principio, que es básico en la búsqueda de la privacidad ideal.

**Principio 2.16 (Principio desinformativo)** *La tabla publicada debe proporcionar al atacante poca información adicional a la información que el mismo tenga a priori. Es decir, no debe haber gran diferencia entre la información a priori y la información a posteriori.*

Sin embargo, nos vamos a seguir encontrando con diversos problemas:

- desconocimiento de la distribución completa de los atributos en la población,
- desconocimiento de la información de la que el atacante dispone,

- información no modelable probabilísticamente,
- existencia de varios atacantes con diferentes niveles de información.

En la formalización de los posibles problemas que pueden presentarse es necesario el siguiente concepto.

**Definición 2.17** *Dada una tabla $T$ y una tabla generalizada asociada a la misma $T^*$, se denomina $q^*$-**bloque** al grupo de tuplas de $T^*$ cuyo valor sensible es $q^*$.*

Imaginemos que el atacante pretende obtener el valor del atributo sensible mediante "positive disclosure", es decir, que $t[S] = s$ con una alta probabilidad. Según la ecuación (2.1), esto sólo ocurre cuando:

$$\exists s \text{ tal que } \forall s' \neq s, n_{(q^*,s')} \frac{f(s'|q)}{f(s'|q^*)} \ll n_{(q^*,s)} \frac{f(s|q)}{f(s|q^*)}. \tag{2.2}$$

Este problema puede presentarse por dos factores: una falta de diversidad en el atributo sensible del $q^*$-bloque y/o una gran información adicional del atacante:

- **Falta de diversidad**: la falta de diversidad en el atributo sensible se manifiesta del siguiente modo:

  $$\forall s' \neq s, n_{(q^*,s')} \ll n_{(q^*,s)}. \tag{2.3}$$

  Es decir, casi todas las tuplas tienen un mismo valor sensible $s$, y por tanto $\beta_{(q,s,T^*)} \approx 1$.

  Una forma de defenderse frente a estos ataques es conseguir que las generalizaciones realizadas para llegar a la tabla $T^*$ permitan que en cada $q^*$-bloque haya $l \geq 1$ opciones distintas del atributo sensible, y que estén lo suficientemente representadas. En este caso, se dice que el $q^*$-bloque está bien representado por $l$ valores sensibles. Conviene notar que al requerir lo anterior, nos podemos encontrar con tener que generalizar demasiado.

- **Gran cantidad de información adicional**: incluso cuando tenemos que el $q^*$-bloque tiene $l$ valores sensibles bien representados, con la información adicional del atacante nos podemos encontrar con que puede eliminar opciones del atributo, cuando ocurre lo siguiente:

$$\exists s', \frac{f(s'|q)}{f(s'|q^*)} \approx 0. \tag{2.4}$$

  Es decir, para el valor no sensible $q$, el valor $s'$ es mucho menos probable que otro en el $q^*$-bloque.

Lógicamente, si tenemos $l$ valores bien representados en el $q^*$-bloque, el atacante necesitará eliminar $l-1$ valores del atributo sensible para llegar a "positive disclosure", por lo que a mayor $l$, mayor protección ante este tipo de ataques. Así, una buena representación de los valores es importante, evitando casos claramente descartables, como el ejemplo dado anteriormente sobre el ébola.

En base a lo anterior, se establece un segundo principio, básico en la descripción de la $l$-diversity.

**Principio 2.18 (Principio de $l$-diversity)** *Un $q^*$-bloque es $l$-diverso si contiene al menos $l$ valores bien representados para el atributo sensible $S$. Una tabla es $l$-diversa si cada $q^*$-bloque es $l$-diverso.*

**Ejemplo 2.19** *Si consideramos la tabla obtenida en el Ejemplo 2.10, se tiene que cumple la condición de $l$-diversity, para $l = 3$, ya que en los tres $q^*$-bloques obtenidos se tiene al menos 3 valores sensibles bien representados en cada uno. En el caso de que nos encontremos con problemas con la información adicional del atacante, esto se podría solucionar considerando una $l$-diversity para valores mayores de $l$, aunque esto conllevaría una pérdida de información.*

La definición de $l$-diversity no se ha limitado al caso en el que se considera un único atributo sensible, sino que ha sido generalizada al caso de varios atributos sensibles.

**Definición 2.20 ($l$-diversity para varios atributos sensibles)** *Sea $T$ una tabla con $Q_1, \ldots, Q_m$ atributos no sensibles y $S_1, \ldots, S_r$ atributos sensibles. Decimos que $T$ es* **$l$-diversa** *si $\forall\ i = 1, \ldots, r$, la tabla es $T$ diversa cuando $S_i$ es tratado como el único atributo sensible y $\{Q_1, \ldots, Q_m\}$ junto a $\{S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_r\}$ son tratados como el cuasi-identificador.*

**Ejemplo 2.21** *Si para los atributos de los ejemplos que estamos tratando (ver Tabla 2.1), consideramos que el atributo Salario debe ser tratado como sensible, nos encontraríamos que los atributos no sensibles serían el código ZIP y la Edad, mientras que los sensibles serían el Salario y la Enfermedad.*

A pesar de todo lo anterior, la $l$-diversity no permite asegurar la privacidad cuando el atacante tenga información sobre la población, o en caso de similitud en los valores del atributo sensible. Con el objetivo de solucionar este problema, surge otro método para intentar proteger los datos, la $t$-closeness.

### 2.1.3. $t$-closeness

En la $t$-closeness, se separa la información de la que dispone el atacante en dos: la relativa a la población general, y la que dispone acerca de individuos específicos (ver [23]).

El atacante, tiene un conocimiento a priori sobre el atributo sensible del individuo, que lo vamos a denotar como $B_0$. Tras haber sido liberada la tabla generalizada, dicho conocimiento se ve influenciado por la distribución del atributo sensible en la tabla completa, que denotaremos por $Q$, y pasa a ser $B_1$. Al identificar los valores del cuasi-identificador en dicha tabla, el atacante identificará el bloque al que pertenece, y tendrá la distribución del atributo sensible en dicho bloque, que lo denotaremos por $P$, y su conocimiento pasa a ser $B_2$.

$$\boxed{B_0 \xrightarrow{Q} B_1 \xrightarrow{P} B_2}$$

Utilizando el esquema descrito, la $l$-diversity intenta que la diferencia entre $B_0$ y $B_2$ se limite lo más posible, requiriendo que $P$ tenga un cierto nivel de diversidad. En la $t$-closeness lo que intentaremos limitar será la diferencia entre $B_1$ y $B_2$, dando por supuesto que $Q$ es información pública.

Para conseguir dicha limitación, se actúa a través de $P$ y $Q$. De manera intuitiva, podemos ver que si $P = Q$, entonces $B_1$ y $B_2$ deberían ser la misma. Del mismo modo, si $P$ y $Q$ son cercanos, debería ocurrir lo mismo con $B_1$ y $B_2$.

**Principio 2.22 ($t$-closeness)** *Un bloque se dice que cumple la* **condición $t$-closeness** *si la diferencia entre la distribución del atributo sensible en dicho bloque y la distribución del atributo sensible en la tabla completa, no es mayor de un límite $t$. Una tabla se dice que cumple la condición $t$-closeness si todos los bloques la cumplen.*

Lógicamente, al intentar que $P$ y $Q$ sean parecidas, se limita la cantidad de información útil que se libera, ya que estaríamos limitando la relación entre el cuasi-identificador y el atributo sensible. Ésta es una consecuencia esperada, ya que de no hacerlo, el atacante obtendrá información suficiente como para poder llevar a cabo alguno de los ataques que hemos visto en los ejemplos anteriores. El parámetro $t$ nos permite elegir entre utilidad y privacidad, según el valor tomado, ya que a menor valor de $t$, tenemos mayor privacidad, pero se pierde información en los datos publicados.

Puesto que en este método es necesario medir la diferencia entre ambas distribuciones de probabilidad, se necesita disponer de una distancia. Existen varios tipos de distancias, entre ellas están la distancia variacional y la distancia de Kullback-Leibler, pero es la *Earth Mover's Distance* (EMD) la que mejor se adapta a las necesidades de este método y, por tanto, la que ha sido elegida para cuantificar la diferencia entre las dos distribuciones de probabilidad.

El uso de la EMD, nos proporciona dos importantes propiedades de la $t$-closeness:

**Propiedad 2.23 (Propiedad de generalización)** *Sea $T$ una tabla, y $A$ y $B$ dos generalizaciones de dicha tabla, tal que $A$ es más general que $B$. Si $T$*

*satisface la t-closeness para la generalización B, entonces también satisface la t-closeness para la generalización A.*

**Propiedad 2.24 (Propiedad de subconjuntos)** *Sea $T$ una tabla y $C$ un conjunto de atributos de $T$. Si $T$ satisface la t-closeness respecto a $C$, entonces $T$ también satisface la t-closeness respecto a cualquier subconjunto de atributos $D \subset C$.*

Para utilizar la distancia EMD, debemos distinguir los casos en que el atributo sensible sea cuantitativo o cualitativo.

- **EMD para atributos cuantitativos:** Supongamos que $\{v_1, \ldots, v_m\}$ son los valores numéricos, ordenados de forma creciente, que puede tomar el atributo sensible. Supongamos también que las distribuciones $P$ y $Q$, correspondientes a las probabilidades de $v_i, \quad \forall i \in \{1, \ldots, m\}$, vienen dadas del siguiente modo: $P = (p_1, \ldots, p_m)$ y $Q = (q_1, \ldots, q_m)$. Si denotamos por $r_i$, $\forall i \in \{1, \ldots, m\}$, la diferencia $r_i = p_i - q_i$, la distancia entre $P$ y $Q$ quedará definida por:

$$\begin{aligned} D[P,Q] &= \frac{1}{m-1}(|r_1| + |r_1 + r_2| + \cdots + |r_1 + \cdots + r_m|) \\ &= \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i} r_j\right|. \end{aligned} \tag{2.5}$$

- **EMD para atributos cualitativos:** En este caso, no tendremos, en general, una relación de orden entre los atributos, tal como ocurría en el caso numérico, por lo que la distancia se obtiene como:

$$D[P,Q] = \frac{1}{2}\sum_{i=1}^{m}|p_i - q_i| = \sum_{p_i \geq q_i}(p_i - q_i) = -\sum_{p_i < q_i}(p_i - q_i). \tag{2.6}$$

**Ejemplo 2.25** *Considerando la Tabla 2.6, en la Tabla 2.7 podemos ver su tabla 3-diversa asociada.*

| Individuo | ZIP | Edad | Salario(k) | Enfermedad |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3 | úlcera gástrica |
| 2 | 47602 | 22 | 4 | gastritis |
| 3 | 47678 | 27 | 5 | cáncer de estómago |
| 4 | 47905 | 43 | 6 | gastritis |
| 5 | 47909 | 52 | 11 | gripe |
| 6 | 47906 | 47 | 8 | bronquitis |
| 7 | 47605 | 30 | 7 | bronquitis |
| 8 | 47673 | 36 | 9 | neumonía |
| 9 | 47607 | 32 | 10 | cáncer de estómago |

Tabla 2.6: Tabla original antes de aplicar $t$-closeness.

| Individuo | ZIP | Edad | Salario(k) | Enfermedad |
|---|---|---|---|---|
| 1 | 476** | 2* | 3 | úlcera gástrica |
| 2 | 476** | 2* | 4 | gastritis |
| 3 | 476** | 2* | 5 | cáncer de estómago |
| 4 | 479** | $\geq 40$ | 6 | gastritis |
| 5 | 479** | $\geq 40$ | 11 | gripe |
| 6 | 479** | $\geq 40$ | 8 | bronquitis |
| 7 | 476** | 3* | 7 | bronquitis |
| 8 | 476** | 3* | 9 | neumonía |
| 9 | 476** | 3* | 10 | cáncer de estómago |

Tabla 2.7: Tabla 3-diversa.

*En este caso,* $Q = \{3, 4, 5, 6, 7, 8, 9, 10, 11\}$ *son los valores del atributo sensible Salario, ordenados de menor a mayor, que son equiprobables. Así, puesto que* $P_1 = \{3, 4, 5\}$, $P_2 = \{6, 8, 11\}$ *y* $P_3 = \{7, 9, 10\}$, *si utilizamos la EMD para medir las distancias* $D[P_1, Q]$ *y* $D[P_2, Q]$, *se tiene que*

$$D[P_1, Q] = 0.375 \text{ y } D[P_2, Q] = 0.167.$$

### 2.1.4. Tipos de ataques

En este subapartado, detallaremos los distintos tipos de ataques que hemos analizado hasta ahora.

- **Ataques de homogeneidad.** Este tipo de ataque radica en la aparición de un único valor para el atributo sensible en algún bloque de la tabla. Este problema lo podemos encontrar en tablas $k$-anónimas, donde aunque los $k$ individuos son indistinguibles respecto a los valores no sensibles, el valor sensible puede coincidir y romper la privacidad. Para solucionar esto, hemos utilizado la $l$-diversity, donde al exigir que haya $l$ valores sensibles que queden suficientemente representados, no nos encontraremos con este tipo de ataque.

- **Información adicional.** Es un tipo de ataque difícil de abordar, ya que no disponemos de la información que tiene el atacante (o atacantes). Dicha información puede hacer, a quien intenta atacar la privacidad, eliminar ciertos valores del atributo sensible ("negative disclosure") o dar con uno concreto ("positive disclosure"). El primero de ellos se puede proteger con tablas $l$-diversas con un valor suficientemente grande de $l$, ya que el atacante necesitará eliminar un total de $l - 1$ valores para poder dar con el atributo sensible.

- **Similitud de valores.** Este problema radica en que los valores que toma el atributo sensible sean muy similares entre sí, y un posible atacante pueda llegar a ciertas conclusiones bastante certeras. Este problema lo encontramos en la $l$-diversity, que si bien requiere a los bloques una representación de los atributos sensibles, no logra tener en cuenta las similitudes entre esos valores.

  Puede aparecer tanto en atributos sensibles cualitativos (por ejemplo, enfermedades de un tipo similar puede dar una conclusión al atacante indeseada por el individuo), como cuantitativos (por ejemplo, el salario de una persona, si obtenemos que los valores posibles están en un intervalo 1000-1200 euros, tendremos una buena aproximación). Para abordarlo, tenemos la $t$-closeness, que busca que la distribución del atributo sensible en cada bloque sea similar a la de los datos completos.

Ninguna de las técnicas presentadas hasta ahora asegura completamente la protección de los datos sensibles, por lo que en este trabajo se ha intentado obtener métodos alternativos que generalicen y mejoren a los anteriores. Dichos métodos consistirán, básicamente, en la utilización de los conjuntos difusos en la representación de los atributos de las tablas generalizadas. Para facilitar una mejor comprensión de dichos métodos, en la siguiente sección se presentan los fundamentos de la teoría de conjuntos difusos.

## 2.2. Conjuntos difusos

La palabra fuzzy viene del inglés fuzz (tamo, pelusa, vello) y se traduce por difuso o borroso. Muchos conceptos que manejamos los humanos a menudo, no tienen una definición clara: ¿qué es una persona alta? ¿a partir de qué edad una persona deja de ser joven? Para dar respuesta a este tipo de preguntas surgen los conjuntos difusos (en inglés *fuzzy sets*), los cuales se definen basándose en la existencia de distintos grados de pertenencia al conjunto.

### 2.2.1. Definiciones básicas de conjuntos difusos

Los conceptos que desarrollaremos en este apartado se pueden encontrar con mayor detalle en [13], [14], [15], [29] y [33].

**Definición 2.26** *Dado un conjunto referencial $X$ al que llamaremos* **universo***, se dice que $A$ es un subconjunto* **difuso** *(o* fuzzy*) de $X$, cuando a cada elemento $x \in X$ se le asigna un grado de pertenencia a dicho conjunto $A$.*

**Definición 2.27** *El grado de pertenencia en $A$ de los elementos de $X$ viene dado por la* **función de pertenencia** *de $A$, que denotaremos por $\mu_A$, y verifica que $0 \leq \mu_A(x) \leq 1, \forall x \in X$. Al conjunto $A$ lo denotaremos, en*

*función de $X$ y $\mu_A$, del siguiente modo:*

$$A = \{x, \mu_A(x)\}, x \in X.$$

**Nota 2.28** *Un subconjunto clásico de $X$, $B \subset X$, se puede ver como un caso particular de un subconjunto difuso de $X$, cuya función de pertenencia asociada es:*

$$\mu_B(x) = \begin{cases} 1 & si\ x \in B, \\ 0 & si\ x \notin B. \end{cases}$$

**Definición 2.29** *Se dice que un conjunto difuso $A$ es* **vacío** *si su función de pertenencia asociada es tal que $\mu_A(x) = 0, \forall x \in X$.*

**Ejemplo 2.30** *La representación gráfica de una función de pertenencia, correspondiente al conjunto difuso $A$ con función de pertenencia:*

$$\mu_A(x) = \begin{cases} x & si\ x \in [0,1], \\ 2-x & si\ x \in (1,2], \\ 0 & en\ el\ resto, \end{cases}$$

*viene dada en la Figura 2.1.*

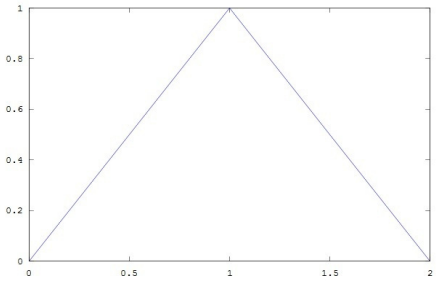


Figura 2.1: Ejemplo de una función de pertenencia.

**Definición 2.31** *Decimos que un conjunto difuso $A$ es* **normal** *si:*

$$\max_{x \in X} \mu_A(x) = 1.$$

**Definición 2.32** *Decimos que un conjunto difuso $A$ es* **convexo** *cuando se verifica que:*

$$\mu_A(\lambda x + (1-\lambda)y) \geq min\{\mu_A(x), \mu_A(y)\}, \quad \forall x, y \in X, \forall \lambda \in [0,1].$$

A continuación definiremos las operaciones, relaciones lógicas y leyes básicas que nos encontraremos entre conjuntos difusos:

- **Unión.** Sean $A$ y $B$ dos conjuntos difusos, la función de pertenencia de $A \cup B$ viene dada por:

$$\mu_{A\cup B}(x) = max\{\mu_A(x), \mu_B(x)\}, \ \forall x \in X.$$

- **Intersección.** Sean $A$ y $B$ dos conjuntos difusos, la función de pertenencia de $A \cap B$ viene dada por:

$$\mu_{A\cap B}(x) = min\{\mu_A(x), \mu_B(x)\}, \ \forall x \in X.$$

- **Complemento.** Sea $A$ un conjunto difuso, definimos el complemento como el conjunto difuso $\overline{A}$ con función de pertenencia:

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x), \ \forall x \in X.$$

- **Relación de equivalencia.** Se dice que dos conjuntos difusos $A$ y $B$ definidos sobre un mismo universo son equivalentes si sus funciones de pertenencia son iguales, es decir,

$$\mu_A(x) = \mu_B(x), \ \forall x \in X.$$

- **Relación de inclusión.** Dados dos conjuntos difusos $A$ y $B$ definidos sobre el mismo universo $X$, se dice que $A$ está contenido en $B$, y se representa por $A \subset B$, si y sólo si sus funciones de pertenencia verifican:

$$\mu_A(x) \leq \mu_B(x), \forall x \in X.$$

En lo que respecta a las operaciones algebraicas más comunes, se tiene:

- **Producto algebraico.** Dados dos conjuntos difusos $A$ y $B$ definidos sobre el mismo universo $X$, se define el producto algebraico de $A$ y $B$, y se denota por $A \cdot B$, como el conjunto difuso con función de pertenencia:
$$\mu_{A \cdot B}(x) = \mu_A(x) \cdot \mu_B(x), \forall x \in X.$$
- **Suma algebraica.** Dados dos conjuntos difusos $A$ y $B$ definidos sobre el mismo universo $X$, se define la suma algebraica de $A$ y $B$, y se denota por $A + B$, como el conjunto difuso con función de pertenencia:
$$\mu_{A+B}(x) = \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x), \forall x \in X.$$

A continuación definimos los conceptos de $\lambda$-complemento y $\alpha$-corte de un conjunto difuso:

**Definición 2.33** *Dado un conjunto difuso cualquiera $A$ del universo $X$ y dado un número real cualquiera $\lambda$ en el intervalo $(-1, \infty)$, se define el* **$\lambda$-complemento** *de $A$, y se denota por $\overline{A}^{\lambda}$, como el conjunto difuso cuya función de pertenencia viene dada por:*

$$\mu_{\overline{A}^{\lambda}}(x) = \frac{1 - \mu_A(x)}{1 + \lambda \mu_A(x)}, \quad \forall x \in X.$$

Este concepto nos da un complemento gradual del conjunto difuso $A$ cumpliendo las siguientes propiedades:

- $\overline{A}^0 = \overline{A}$,
- $\overline{A}^i \subseteq \overline{A}^j$ si $i \geq j$,
- $\lim_{\lambda \to -1} A^{\lambda} = X$,
- $\lim_{\lambda \to \infty} A^{\lambda} = \emptyset$.

**Definición 2.34** *Dado un conjunto difuso cualquiera $A$ del universo $X$, podemos asociar a él dos conjuntos clásicos:*

- **$\alpha$-corte** *o* **$\alpha$-corte fuerte***: es el conjunto clásico formado por los elementos de $X$ que tienen grado de pertenencia a $A$ mayor o igual que $\alpha$, es decir,*

$$A_\alpha = \{x | \mu_A(x) \geq \alpha\}, \alpha \in (0, 1].$$

- **$\alpha$-corte débil***: es el conjunto clásico formado por los elementos de $X$ que tienen grado de pertenencia a $A$ mayor que $\alpha$, es decir,*

$$A_\alpha = \{x | \mu_A(x) > \alpha\}, \alpha \in [0, 1).$$

**Nota 2.35** *En el caso de que la función de pertenencia sea una función continua, no hay diferencia entre ambos $\alpha$-cortes (fuerte y débil). En tal caso, el $\alpha$-corte verifica las siguientes propiedades:*

- $A_0 = A$,
- $A_1 = \{x, \mu_A(x) = 1\}$*, conjunto que recibe el nombre de* **núcleo**.

A continuación se definen tres tipos de conjuntos difusos, especialmente útiles en este trabajo.

- **Triangulares de extremos finitos.** Estos conjuntos difusos se denotarán por $A = (a; b; c)$, con $-\infty < a < b < c < \infty$ y vienen definidos por partes, puesto que su función de pertenencia viene dada por las rectas entre los puntos $(a, 0)$ y $(b, 1)$ en el primer tramo y $(b, 1)$ y $(c, 0)$ en el segundo. Es decir:

$$\mu_A(x) = \begin{cases} \dfrac{x-a}{b-a} & \text{si } x \in [a, b], \\ \dfrac{x-c}{b-c} & \text{si } x \in (b, c], \\ 0 & \text{en el resto.} \end{cases}$$

  Un ejemplo de difuso triangular de extremo finito puede verse en la Figura 2.1, donde se ha representado el conjunto $A = (0; 1; 2)$.

- **Triangulares con algún extremo infinito.** Estos conjuntos difusos se denotarán por $A = (-\infty; a; b)$ (o bien $(a; b; \infty)$, cuya definición y razonamiento es análogo), con $a < b$. Su función de pertenencia queda determinada por la recta constante en 1 hasta el punto $a$, y la recta entre los puntos $(a, 1)$ y $(b, 0)$. Es decir:

$$\mu_A(x) = \begin{cases} 1 & \text{si } x \leq a, \\ \dfrac{x-a}{b-a} & \text{si } x \in (a, b], \\ 0 & \text{en el resto.} \end{cases}$$

  Un ejemplo de conjunto difuso triangular con extremo infinito es $A = (-\infty; 1; 2)$, cuya función de pertenencia aparece representada en la Figura 2.2.

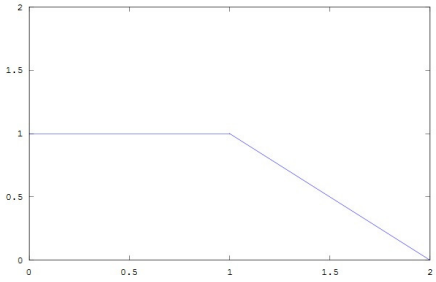


Figura 2.2: Función de pertenencia de $A = (-\infty; 1; 2)$.

- **Trapezoidales.** Este tipo de conjunto difuso se denotará usualmente por $A = (a; b; c; d)$, con $-\infty < a < b < c < d < \infty$ y su función de pertenencia viene definida a trozos por las rectas que unen los puntos: $(a, 0)$ y $(b, 1)$, $(b, 1)$ y $(c, 1)$ y $(c, 1)$ y $(d, 0)$. Es decir:

$$\mu_A(x) = \begin{cases} \dfrac{x-a}{b-a} & \text{si } x \in [a, b], \\ 1 & x \in (b, c], \\ \dfrac{x-d}{c-d} & \text{si } x \in (c, d], \\ 0 & \text{en el resto.} \end{cases}$$

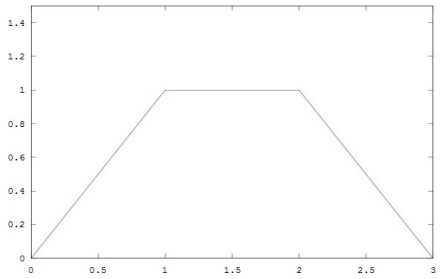


Figura 2.3: Función de pertenencia de la trapezoidal (0;1;2;3).

Un ejemplo de conjunto difuso trapezoidal es $A = (0; 1; 2; 3)$, cuya función de pertenencia aparece representada en la Figura 2.3.

Si algún extremo es infinito, estaríamos en el caso de las triangulares con extremo infinito.

### 2.2.2. Cardinalidad en conjuntos difusos

En este apartado, desarrollaremos los distintos métodos utilizados para obtener la cardinalidad de los conjuntos difusos. Para información más detallada sobre los mismos ver [9] y [27].

Existen diferentes formas de obtener la cardinalidad de un conjunto difuso, de manera que las distintas formas de contar los elementos de un conjunto pueden proporcionar soluciones distintas a un mismo problema. Por ejemplo, supongamos que en un grupo de personas, se quiere saber cuántas son jóvenes, podemos considerar que hay tres respuestas posibles:

**(i)** sobre 5,

**(ii)** 5,

**(iii)** 4.53.

Parece que la tercera respuesta posible no es la más adecuada, mientras que las dos primeras parecen más lógicas. La primera corresponderá a la

obtenida mediante un concepto de cardinalidad difusa, mientras que las otras dos están asociadas a conceptos de cardinalidad no difusa. Vamos a desarrollar y formalizar, a continuación, dichos conceptos.

**Definición 2.36** *Sea $X = \{x_1, \ldots, x_n\}$ el universo y $A$ un conjunto difuso de él con función de pertenencia $\mu_A$. La* **cardinalidad** *del conjunto difuso $A$, que será llamada* **$\sigma$-count**, *se define como:*

$$|A|_\sigma = \sum_{x \in X} \mu_A(x). \tag{2.7}$$

**Ejemplo 2.37** *Sea $X = \{1, \ldots, 5\}$, y $A$ un conjunto difuso cuya función de pertenencia viene dada por:*

$$\left( \begin{array}{c|ccccc} x & 1 & 2 & 3 & 4 & 5 \\ \mu_A(x) & 0.1 & 0.35 & 0.4 & 0 & 0.3 \end{array} \right).$$

*Entonces, la cardinalidad $\sigma$-count de $A$ será:*

$$|A|_\sigma = \sum_{x \in X} \mu_A(x) = 0.1 + 0.35 + 0.4 + 0 + 0.3 = 1.15.$$

Esta forma de contar los elementos de $A$, se correspondería con la opción (iii) del ejemplo planteado al inicio del apartado y, como ya comentamos en su momento, no parece ser la respuesta más lógica.

Para intentar dar una respuesta más ajustada a la realidad, se desarrolla otra técnica dentro de la cardinalidad difusa basada en el siguiente teorema (ver [26]):

**Teorema 2.38** *Dado un conjunto difuso $A$ de un universo $X = \{x_1, \ldots, x_n\}$, la* **cardinalidad difusa** *de $A$ viene dada por:*

$$|A|_f(k) = min\{\mu_{(k)}, 1 - \mu_{(k+1)}\}, \quad k = 0, 1, \ldots, n, \tag{2.8}$$

*donde $\mu_{(1)}, \ldots, \mu_{(n)}$ se corresponden con los valores $\mu_A(x_1), \ldots, \mu_A(x_n)$ ordenados de forma decreciente, siendo $\mu_{(0)} = 1$ y $\mu_{(n+1)} = 0$.*

Obsérvese que hemos denotado por $|A|_f(k)$ a la posibilidad de que el cardinal de $A$ sea exactamente $k$, es decir, $k$ elementos pertenezcan a $A$, y $n-k$ no pertenezcan. Esto es, $|A|_f(k) = Poss(|A| = k)$.

En [26] se concluye que la cardinalidad difusa $|A|_f$ es un conjunto difuso convexo y la posibilidad (Poss) de que $A$ tenga al menos $k$ elementos se puede expresar como:

$$Poss(|A|_f \geq k) = \begin{cases} \mu_{(k)} & \text{si } k \geq j, \\ max\{1-\mu_{(j)}, \mu_{(j)}\} & \text{si } k < j, \end{cases} \tag{2.9}$$

donde

$$j = \begin{cases} max\{1 \leq s \leq n | \mu_{(s-1)} + \mu_{(s)} > 1\} & \text{si } A \neq \emptyset, \\ 0 & \text{si } A = \emptyset. \end{cases} \tag{2.10}$$

Teniendo en cuenta lo anterior, definimos la **cardinalidad no difusa** del conjunto difuso $A$ como:

$$|A|_{nf} = \{0 \leq k \leq n | max_{0 \leq s \leq n} |A|_f(s) = |A|_f(k)\}. \tag{2.11}$$

**Nota 2.39** *Si $A$ es un subconjunto clásico (no difuso) de $X$ con $r$ elementos, entonces los valores ordenados de $\mu_A(x_1), \ldots, \mu_A(x_n)$ serán $1 \geq \cdots \geq 1 \geq 0 \geq \cdots \geq 0$, donde hay $r$ 1's y $n-r$ 0's. Por lo que:*

$$|A|_f(k) = \begin{cases} 1 & si\ k = r, \\ 0 & si\ k \neq r. \end{cases}$$

**Ejemplo 2.40** *Sea $A$ un conjunto difuso definido como sigue:*

$$A = \left( \begin{array}{c|cccc} x & x_1 & x_2 & x_3 & x_4 \\ \mu_A(x) & 0.7 & 0.6 & 0.3 & 0 \end{array} \right).$$

*Entonces:*

$$\begin{aligned} |A|_f(0) &= min\{\mu_{(0)}, 1-\mu_{(1)}\} = min\{1, 1-0.7\} = 0.3, \\ |A|_f(1) &= min\{\mu_{(1)}, 1-\mu_{(2)}\} = min\{0.7, 1-0.6\} = 0.4, \\ |A|_f(2) &= min\{\mu_{(2)}, 1-\mu_{(3)}\} = min\{0.6, 1-0.3\} = 0.6, \\ |A|_f(3) &= min\{\mu_{(3)}, 1-\mu_{(4)}\} = min\{0.3, 1-0\} = 0.3, \\ |A|_f(4) &= min\{\mu_{(4)}, 1-\mu_{(5)}\} = min\{0, 1-0\} = 0, \end{aligned}$$

*de donde se obtiene que:*

$$|A|_f = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.3 & 0.4 & 0.6 & 0.3 & 0 \end{pmatrix}.$$

*Por lo tanto, como* $|A|_{nf} = m$ *tal que* $|A|_f(m) = max_{0 \leq k \leq 4}|A|_f(k)$*, y tenemos que:*

$$max_{0 \leq k \leq 4}|A|_f(k) = max\{0.3, 0.4, 0.6, 0.3, 0\} = 0.6,$$

*que se corresponde a 2 en* $|A|_f$*, por tanto:*

$$|A|_{nf} = 2.$$

Tenemos una forma alternativa de calcular la cardinalidad no difusa, que viene dado junto a la ecuación (2.8), con la siguiente expresión:

$$|A|_{nf} = \begin{cases} 0 & \text{si } A = \emptyset, \\ j & \text{si } A \neq \emptyset \text{ y } \mu_{(j)} \geq 0.5, \\ j-1 & \text{si } A \neq \emptyset \text{ y } \mu_{(j)} < 0.5. \end{cases} \tag{2.12}$$

**Ejemplo 2.41** *Con los mismos datos anteriores, calculamos* $j$ *según (2.8) como sigue:*

*Como* $A \neq \emptyset$*, se tiene que:*

$$j = max\{1 \leq s \leq 4 | \mu_{(s-1)} + \mu_{(s)} > 1\} = 2,$$

*y como tenemos que* $\mu_{(2)} = 0.6 \geq 0.5$*, con la expresión de (2.10), resulta que:*

$$|A|_{nf} = 2.$$

*Comparándolo con la* $\sigma$*-count, tenemos que:*

$$|A|_\sigma = \sum_{x \in X} \mu_A(x) = 0.7 + 0.6 + 0.3 + 0 = 1.6.$$

# Capítulo 3

# Nueva técnicas de privacidad basadas en conjuntos difusos

En este apartado desarrollaremos las técnicas de privacidad anteriormente estudiadas, $k$-anonymity, $l$-diversity y $t$-closeness, considerando los conjuntos difusos como forma de describir los atributos.

Como explicamos en los conceptos básicos de privacidad, dispondremos de una tabla original de datos, no generalizados, en los que a lo más, se han eliminado los datos que puedan relacionar directamente a cada individuo, como pueden ser el nombre y los apellidos o el DNI. Hemos visto como las generalizaciones que hemos ido haciendo a las tablas originales han sido siempre considerando intervalos (en la Edad o en el Salario lo hemos hecho en varios ejemplos), o eliminando cifras (para el código ZIP, hemos pasado de 5 cifras, a sólo las 3 ó 4 primeras).

Para ir desarrollando nuestro trabajo y facilitar su comprensión, utilizaremos un ejemplo concreto a lo largo de todo este capítulo, el cual corresponderá a los datos presentados en la Tabla 3.1, que es el mismo que ya consideramos a lo largo del capítulo anterior.

Antes de abordar los nuevos métodos, necesitamos presentar algunas definiciones complementarias. Así, comenzaremos con la definición de partición difusa de un atributo.

| Individuo | ZIP | Edad | Salario (k) | Enfermedad |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3.5 | úlcera gástrica |
| 2 | 47602 | 22 | 3.5 | gastritis |
| 3 | 47678 | 27 | 5 | cáncer de estómago |
| 4 | 47905 | 43 | 7 | gastritis |
| 5 | 47929 | 52 | 7 | gripe |
| 6 | 47906 | 47 | 8 | bronquitis |
| 7 | 47973 | 36 | 9 | neumonía |
| 8 | 47607 | 32 | 10 | cáncer de estómago |
| 9 | 47906 | 55 | 9 | ataque al corazón |
| 10 | 47925 | 56 | 9 | ataque al corazón |
| 11 | 47923 | 61 | 10 | angina de pecho |
| 12 | 47923 | 67 | 15 | neumonía |

Tabla 3.1: Tabla original a generalizar utilizando conjuntos difusos.

**Definición 3.1** *Sea $A$ un atributo de la tabla $T$ y $D(A)$ los valores nítidos que puede tomar el atributo $A$. Una* **partición difusa** *de $D(A)$ viene dada por $\{A_1, \ldots, A_n\}$, siendo $A_i$ conjuntos difusos $\forall i \in \{1, \ldots, n\}$, con funciones de pertenencia asociadas $\mu_{A_1}, \ldots, \mu_{A_n}$, tales que:*

$$\forall x \in D(A), \sum_{i \in \{1,\ldots,n\}} \mu_{A_i}(x) = 1. \tag{3.1}$$

**Definición 3.2** *Decimos que una partición difusa $\{A_1, \ldots, A_n\}$ es* **robusta** *si $\forall x \in X$, $\exists k$ tal que $\mu_{A_k}(x) \geq 0.5$.*

**Definición 3.3** *Decimos que una partición difusa $\{A_1, \ldots, A_n\}$ es* **normal** *si $A_k$ es normal, para todo $k \in \{1, 2, \ldots, n\}$, es decir,*

$$max_{x \in X} \mu_{A_k}(x) = 1, \quad \forall k = 1, \ldots, n.$$

Es evidente que si una partición difusa es normal, también es robusta. De aquí en adelante, consideraremos sólo particiones difusas normales, aunque no se diga explícitamente.

**Proposición 3.4** *Sea* $\{A_1, \ldots, A_n\}$ *una partición difusa definida sobre un universo* $X$ *finito, entonces*

$$\sum_{k=1}^{n} |A_k|_\sigma = \sum_{k=1}^{n} |A_k|_{nf} = |X|. \tag{3.2}$$

La demostración de dicha proposición la podemos ver en [9].

**Ejemplo 3.5** *Consideremos dos particiones difusas de los atributos Edad y Salario de la Tabla 3.1. Utilizaremos los tipos de conjuntos difusos ya descritos, en particular, los dos tipos de conjuntos triangulares, tanto con extremos finitos como con un extremo infinito.*

*Para el atributo Salario, tomaremos la siguiente partición:*

$$\textit{Bajo}= (0; 2; 5),\ \textit{Medio}= (2; 5; 8),\ \textit{Alto}= (5; 8; 11)\ \textit{y Muy Alto}= (8; 11; \infty).$$

*El conjunto difuso Bajo lo trataremos como de extremo izquierdo infinito, ya que el mínimo posible de Salario es* 0*. Las funciones de pertenencia las podemos ver en la Figura 3.1.*

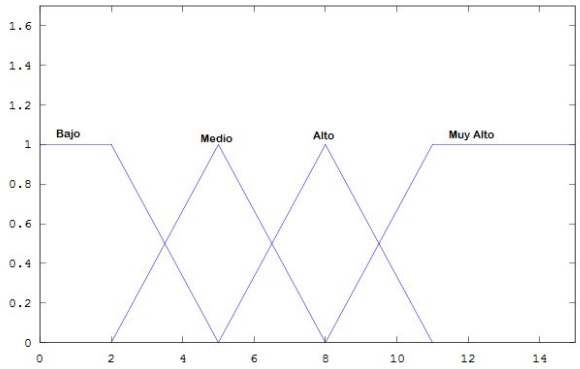


Figura 3.1: Función de pertenencia de cada conjunto difuso de la partición hecha al atributo Salario.

*De modo similar procedemos con el atributo Edad, obteniendo la siguiente partición:*

$$Baja = (0; 30; 42),\ Media = (30; 42, 54),$$
$$Alta = (42; 54; 66),\ Muy\ Alta = (54; 66; \infty).$$

*La representación de las funciones de pertenencia serían similares al caso del Salario.*

*Por último, haríamos la partición del código ZIP, que en este caso consiste en dividirlo en dos grupos,* $476^{**}$ *y* $479^{**}$*. Es claro que tenemos una partición no difusa, ya que un individuo o está o no está en cada conjunto, por lo que sería también una partición difusa para funciones de pertenencia que sólo toman valores* 1 *ó* 0*, de modo que se cumple la restricción (3.1).*

*Así, la tabla que obtendremos con dicha partición será la Tabla 3.2.*

| ZIP | Edad | Salario | Enfermedad |
|---|---|---|---|
| 476** | Baja | Bajo | úlcera gástrica<br>gastritis<br>cáncer de estómago |
| 476** | Baja | Medio | úlcera gástrica<br>gastritis<br>cáncer de estómago |
| 479** | Media | Alto | gastritis<br>bronquitis<br>neumonía |
| 479** | Alta | Alto | gripe<br>ataque al corazón |
| 479** | Muy Alta | Muy Alto | angina de pecho<br>neumonía |

Tabla 3.2: Tabla obtenida tras las particiones difusas hechas a los tres atributos no sensibles.

*El resto de las combinaciones posibles no las hemos incluido, ya que son casi despreciables respecto a las funciones de pertenencia que se obtienen en ellas. Por simplicidad, denotaremos a cada clase de forma abreviada como sigue:*

$Q_1 = 476^{**} \times Baja \times Bajo,$

$Q_2$=*476*** $\times$ *Baja* $\times$ *Medio,*
$Q_3$=*479*** $\times$ *Media* $\times$ *Alto,*
$Q_4$=*479*** $\times$ *Alta* $\times$ *Alto,*
$Q_5$=*479*** $\times$ *Muy Alta* $\times$ *Muy Alto.*

También conviene notar que es necesario calcular la función de pertenencia de un elemento a un conjunto difuso, pero en este caso, cuando el conjunto es un producto de conjuntos difusos.

**Definición 3.6** *Sean $A_1, \ldots, A_n$ conjuntos difusos, con sus correspondientes funciones de pertenencia. Entonces se define el* **producto cartesiano** *de ellos como el conjunto difuso con función de pertenencia:*

$$\mu_{A_1 \times \cdots \times A_n}(x) = \mu_{A_1}(x) \cdots \mu_{A_n}(x), \quad \forall x. \tag{3.3}$$

En las Tablas 3.3 y 3.4 se presentan los valores para cada individuo en las particiones difusas del Salario y de la Edad.

| Individuo | Bajo | Medio | Alto | Muy Alto |
|---|---|---|---|---|
| 1 | 0.5 | 0.5 | 0 | 0 |
| 2 | 0.5 | 0.5 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0.33 | 0.67 | 0 |
| 5 | 0 | 0.33 | 0.67 | 0 |
| 6 | 0 | 0 | 1 | 0 |
| 7 | 0 | 0 | 0.67 | 0.33 |
| 8 | 0 | 0 | 0.33 | 0.67 |
| 9 | 0 | 0 | 0.67 | 0.33 |
| 10 | 0 | 0 | 0.67 | 0.33 |
| 11 | 0 | 0 | 0.33 | 0.67 |
| 12 | 0 | 0 | 0 | 1 |

Tabla 3.3: Valores de las funciones de pertenencia de los conjuntos difusos de la partición del Salario para los 12 individuos de la tabla original.

| Individuo | Baja | Media | Alta | Muy Alta |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 |
| 4 | 0 | 0.92 | 0.08 | 0 |
| 5 | 0 | 0.17 | 0.83 | 0 |
| 6 | 0 | 0.58 | 0.42 | 0 |
| 7 | 0.25 | 0.75 | 0 | 0 |
| 8 | 0.75 | 0.25 | 0 | 0 |
| 9 | 0 | 0 | 0.92 | 0.08 |
| 10 | 0 | 0 | 0.83 | 0.17 |
| 11 | 0 | 0 | 0.42 | 0.58 |
| 12 | 0 | 0 | 0 | 1 |

Tabla 3.4: Valores de las funciones de pertenencia de los conjuntos difusos de la partición del Edad para los 12 individuos de la tabla original.

**Ejemplo 3.7** *Consideremos que $x$ denota al individuo 1 de la Tabla 3.1, entonces se tiene que:*

$$\mu_{476**\times Baja\times Bajo}(x) = \mu_{476**}(x) \cdot \mu_{Baja}(x) \cdot \mu_{Bajo}(x) = 1 \cdot 1 \cdot 0.5 = 0.5.$$

Las siguientes subsecciones están dedicadas a la adaptación de las técnicas de $k$-anonymity, $l$-diversity y $t$-closeness para el caso de conjuntos difusos.

## 3.1. Extensión de la $k$-anonymity

A diferencias de la $l$-diversity o la $t$-closeness, que veremos más adelante como generalizarlas al caso en el que se estén utilizando particiones difusas, la $k$-anonimity carece de sentido en dicho caso. Por esta razón se desarrollará a continuación un método específico cuyos principios serán similares a los considerados en la $k$-anonymity.

**Definición 3.8 (Q-anonymity)** *Sea $T$ una tabla con atributos $A_1, \ldots, A_n$, y sea QIF un cuasi-identificador asociado a ésta, codificado mediante conjuntos difusos. La Q-anonymity de $T$ respecto a QIF viene dada por:*

$$Poss(|T|_f \geq Q) = \mathcal{T}(\beta_{f_1}, \ldots, \beta_{f_s}), \tag{3.4}$$

*donde:*

- $T[QIF]_1, \ldots, T[QIF]_s$ *son los diferentes conjuntos difusos,*
- $\beta_{f_i}$ *la posibilidad de que $T[QIF]_i$ tenga al menos $Q$ elementos,*
- $\mathcal{T}$ *un operador de agregación, es decir, una aplicación* $\mathcal{T} : \bigcup_{n\in\mathbb{N}} [0,1]^n \to [0,1]$ *verificando las condiciones:*
  - *de acotación:*

    $$\mathcal{T}(0, \ldots, 0) = 0, \mathcal{T}(1, \ldots, 1) = 1,$$

  - *de monotonía:*

    $$\forall n \in \mathbb{N} : x_1 \leq y_1, \ldots, x_n \leq y_n \Rightarrow \mathcal{T}(x_1, \ldots, x_n) \leq \mathcal{T}(y_1, \ldots, y_n).$$

**Ejemplo 3.9** *Calculemos la Q-anonymity de los datos correspondientes a la Tabla 3.1. Utilizaremos las particiones dadas en el Ejemplo 3.5. La última fila de la Tabla 3.5 nos muestra la Q-anonymity para la Tabla 3.2 difusa, mientras que las anteriores filas nos dan la posibilidad de tener k elementos para los diferentes cuasi-identificadores difusos.*

**Ejemplo 3.10** *Comparemos dos generalizaciones, una para la k-anonymity y otra para la Q-anonymity, y veamos cuales son las diferencias que nos podemos encontrar. Partimos de los datos de la Tabla 3.6, donde el atributo sensible será la enfermedad.*

| | k=1 | k=2 | k=3 |
|---|---|---|---|
| Poss(476** × Baja × Bajo) | 0.75 | 0.5 | 0.33 |
| Poss(476** × Baja × Medio) | 0.75 | 0.5 | 0.33 |
| Poss(479** × Media × Alto) | 0.67 | 0.67 | 0.58 |
| Poss(479** × Alta × Alto) | 0.67 | 0.67 | 0.67 |
| Poss(479** × Muy Alta × Muy Alto) | 1 | 0.58 | 0.58 |
| Poss($\|T\|_f \geq k$) | 0.77 | 0.58 | 0.49 |

Tabla 3.5: $Q$-anonymity obtenida.

| Individuo | Edad | Enfermedad |
|---|---|---|
| 1 | 21 | gastritis |
| 2 | 24 | gripe |
| 3 | 32 | úlcera de estómago |
| 4 | 36 | gastritis |
| 5 | 45 | neumonía |
| 6 | 56 | gripe |
| 7 | 58 | ataque al corazón |
| 8 | 62 | ataque al corazón |
| 9 | 65 | neumonía |

Tabla 3.6: Tabla de datos para el ejemplo comparativo de $k$-anonymity y $Q$-anonymity.

*Por un lado, haremos una partición no difusa del atributo Edad, como sigue:*

$$(0, 35], \quad (35, 57], \quad (57, \infty).$$

*Con dicha partición, obtenemos la Tabla 3.7, cumpliendo 3-anonymity.*

*Por el otro lado, realizamos otra partición difusa del atributo Edad, como sigue:*

$$Joven = (-\infty; 30; 36), \quad Adulto = (30; 36; 54; 60),$$
$$EdadAvanzada = (54; 60; \infty).$$

*Con esta partición obtenemos la Tabla 3.8, y en la Tabla 3.9 podemos ver la Q-anonymity obtenida.*

| Individuo | Edad | Enfermedad |
|---|---|---|
| 1 | (0,35] | gastritis |
| 2 | (0,35] | gripe |
| 3 | (0,35] | úlcera de estómago |
| 4 | (35,57] | gastritis |
| 5 | (35,57] | neumonía |
| 6 | (35,57] | gripe |
| 7 | $(57, \infty)$ | ataque al corazón |
| 8 | $(57, \infty)$ | ataque al corazón |
| 9 | $(57, \infty)$ | neumonía |

Tabla 3.7: Tabla 3-anónima.

| Edad | Enfermedad |
|---|---|
| Joven | gastritis<br>gripe<br>úlcera de estómago |
| Adulto | gastritis<br>neumonía<br>gripe |
| Edad Avanzada | ataque al corazón<br>neumonía |

Tabla 3.8: Tabla generalizada con una partición difusa.

*La principal ventaja que podemos ver en la Q-anonymity sobre la k-anonymity es la privacidad que proporciona el hecho de que los conjuntos de la partición no se den explicítamente, sino que con conceptos orientativos como son Joven, Adulto o Edad Avanzada, de tal modo que no sea posible identificar al individuo directamente con total seguridad, lo cual si sucede en el caso de conjuntos convencionales, al ser los intervalos conocidos.*

En la generalización a los conjuntos difusos que hemos hecho, nos vamos a encontrar con unas limitaciones similares al caso de conjuntos clásicos: los ataques por homogeneidad. Es decir, aunque una tabla cumpla la condición de $Q$-anonymity, puede suceder que para algún bloque, todos los atributos sensibles sean los mismos, o con una alta posibilidad respecto al resto.

| | k=1 | k=2 | k=3 | k=4 |
|---|---|---|---|---|
| Poss(Joven) | 0.67 | 0.67 | 0.67 | 0 |
| Poss(Adulto) | 0.67 | 0.67 | 0.67 | 0 |
| Poss(Edad Avanzada) | 0.67 | 0.67 | 0.67 | 0.33 |
| Poss($\|T\|_f \geq k$) | 0.67 | 0.67 | 0.67 | 0.11 |

Tabla 3.9: $Q$-anonymity obtenida.

**Ejemplo 3.11** *Podemos ver que la Tabla 3.10 es idéntica a la Tabla 3.1 con la que hemos trabajado hasta el momento, pero cambiando el atributo sensible del individuo 5 de gripe a ataque al corazón.*

| Individuo | ZIP | Edad | Salario (k) | Enfermedad |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3.5 | úlcera gástrica |
| 2 | 47602 | 22 | 3.5 | gastritis |
| 3 | 47678 | 27 | 5 | cáncer de estómago |
| 4 | 47905 | 43 | 7 | gastritis |
| 5 | 47929 | 52 | 7 | ataque al corazón |
| 6 | 47906 | 47 | 8 | bronquitis |
| 7 | 47973 | 36 | 9 | neumonía |
| 8 | 47607 | 32 | 10 | cáncer de estómago |
| 9 | 47906 | 55 | 9 | ataque al corazón |
| 10 | 47925 | 56 | 9 | ataque al corazón |
| 11 | 47923 | 61 | 10 | angina de pecho |
| 12 | 47923 | 67 | 15 | neumonía |

Tabla 3.10: Tabla original de datos.

*Si procedemos con la misma partición que en el Ejemplo 3.5, obtendremos la Tabla 3.11, y aplicando la Definición 3.8, obtendremos los mismos resultados que en el Ejemplo 3.9, dados en la Tabla 3.5.*

*Así, vemos como aun con lo obtenido en dicha tabla, si el atacante sabe que un individuo que ha participado en dichos datos tiene como valores no sensibles código ZIP* 479$^{**}$*, Edad media, y Salario alto, podrá relacionarlo con una alta posibilidad al atributo sensible "ataque al corazón".*

| ZIP | Edad | Salario | Enfermedad |
|---|---|---|---|
| 476** | Baja | Bajo | úlcera gástrica<br>gastritis<br>cáncer de estómago |
| 476** | Baja | Medio | úlcera gástrica<br>gastritis<br>cáncer de estómago |
| 479** | Media | Alto | gastritis<br>bronquitis<br>neumonía |
| 479** | Alta | Alto | ataque al corazón |
| 479** | Muy Alta | Muy Alto | angina de pecho<br>neumonía |

Tabla 3.11: Tabla obtenida tras las particiones difusas hechas a los tres atributos no sensibles.

Lo anterior nos plantea la necesidad de mejorar este aspecto, generalizando al caso difuso la $l$-diversity.

## 3.2. Extensión de la $l$-diversity

Nuestro objetivo a partir de este punto, será adaptar la Definición 2.13 al caso en el que trabajamos con conjuntos y particiones difusas, intentando generalizar la expresión que aparece en él, esto es:

$$\beta_{(q,s,T^*)} = \frac{n_{(q^*,s)} \dfrac{f(s|q)}{f(s|q^*)}}{\displaystyle\sum_{s' \in S} n_{(q^*,s')} \frac{f(s'|q)}{f(s'|q^*)}}.$$

Analizaremos la adaptación al caso difuso de cada uno de los términos que conforman la expresión previa.

En el caso difuso no tendremos el elemento $q^*$ del que hablamos en

la Definición 2.1, ya que cada individuo no tiene una generalización en la tabla $T^*$. Esto lo supliremos con las funciones de pertenencia relativas a cada partición.

**Definición 3.12** *Sea $s \in S$ y $q \in Q$, donde $S$ y $Q$ denotan los atributos sensibles y no sensibles respectivamente. Supongamos que la partición difusa de $Q$ sea la dada por los conjuntos $\{Q_1, \ldots, Q_n\}$ y denotemos por $n(s, q, T^*)$ al número de elementos en la tabla generalizada $T^*$ cuyos valores sean $s$ para el atributo sensible y $q$ para los no sensibles. Entonces:*

$$n(s, q, T^*) = \sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \#(Q_i \cap s). \tag{3.5}$$

*Donde $\#(Q_i \cap s)$ denota el número de elementos que están en $Q_i$ y a la vez tienen como atributo sensible $s$.*

Para poder obtener la expresión (3.5) es necesario calcular $\#(Q_i \cap s)$, o de manera más simple, $\#A$, con $A$ un conjunto difuso cualquiera. Para ello, recurriremos a la cardinalidad no difusa.

**Ejemplo 3.13** *Sea $q = 47905 \times 43 \times 7$ y $s =$gastritis. A continuación, calcularemos el valor de $n(s, q, T^*)$ para la Tabla 3.2.*

*Tenemos que $Q =$ZIP $\times$ Edad $\times$ Salario y $S =$Enfermedad, con lo que los valores de $\mu_{Q_i}(q)$ resultan ser:*

$$\begin{aligned}
\mu_{Q_1}(q) &= \mu_{476**}(q) \cdot \mu_{Baja}(q) \cdot \mu_{Bajo}(q) = 0 \cdot 0 \cdot 0 = 0,\\
\mu_{Q_2}(q) &= \mu_{476**}(q) \cdot \mu_{Baja}(q) \cdot \mu_{Medio}(q) = 0 \cdot 0 \cdot 0.33 = 0,\\
\mu_{Q_3}(q) &= \mu_{479**}(q) \cdot \mu_{Media}(q) \cdot \mu_{Alto}(q) = 1 \cdot 0.92 \cdot 0.67 = 0.6164,\\
\mu_{Q_4}(q) &= \mu_{479**}(q) \cdot \mu_{Alta}(q) \cdot \mu_{Alto}(q) = 1 \cdot 0.08 \cdot 0.67 = 0.0536,\\
\mu_{Q_5}(q) &= \mu_{478**}(q) \cdot \mu_{MuyAlta}(q) \cdot \mu_{MuyAlto}(q) = 1 \cdot 0 \cdot 0 = 0.
\end{aligned}$$

*Ahora, necesitamos obtener $\#(Q_i \cap s)$ para cada $i$.*

*Primero tendremos en cuenta que el atributo $s =$gastritis (sensible) lo trataremos como un atributo difuso con función de pertenencia 1 ó 0 en*

*razón de si el individuo tiene o no la enfermedad correspondiente. En este caso, tendremos que:*

$$\mu_{(Q_1 \cap s)} = \mu_{Q_1} \cdot \mu_s.$$

*Debemos calcular el valor de $j$ como en (2.10), para el conjunto $Q_1 \cap s$, para lo cual, primero debemos obtener la función de pertenencia de cada individuo y ordenarlos de mayor a menor. Por simplicidad, eliminaremos aquellos cuya función de pertenencia sea 0, ya que serán intrascendentes.*

*Para el caso $i = 1$, sólo los individuos 2 y 4 cumplen el requisito de tener como atributo sensible gastritis, por lo que serán los únicos a tratar:*

$$\begin{aligned}
\mu_{(Q_1 \cap s)}(2) &= \mu_{Q_1}(2) \cdot \mu_s(2) = \mu_{Q_1}(2) = \mu_{476**}(2) \cdot \mu_{Baja}(2) \cdot \mu_{Bajo}(2) \\
&= 1 \cdot 1 \cdot 0.5 = 0.5, \\
\mu_{(Q_1 \cap s)}(4) &= \mu_{Q_1}(4) \cdot \mu_s(4) = \mu_{Q_1}(4) = \mu_{476**}(4) \cdot \mu_{Baja}(4) \cdot \mu_{Bajo}(4) \\
&= 0 \cdot 0 \cdot 0 = 0.
\end{aligned}$$

*Considerando que $\mu_{(Q_1 \cap s)}(4) = 0$, sólo nos queda el individuo 2, esto es, $x_1 = 2$. Así,*

$$j = max\{1 \leq s \leq 12 | \mu_{(s-1)} + \mu_{(s)} > 1\} = 1.$$

*Como $\mu_{(1)} = 0.5$ y $\mu_{(0)} = 1$, la suma es mayor que 1, y desde ese término en adelante, serán ceros y no se obtendrá de nuevo la suma mayor que uno. Además, tenemos que $\mu_{(1)} = \mu_{(Q_1 \cap s)}(2) = 0.5 \geq 0.5$, por lo que por (2.12), se tiene que $\#(Q_1 \cap s) = j = 1$.*

*De modo análogo, calcularíamos este valor para $i = 2, 3, 4, 5$, y obtendríamos:*

$$\#(Q_1 \cap s) = 1, \quad \#(Q_2 \cap s) = 1, \quad \#(Q_3 \cap s) = 1,$$

$$\#(Q_4 \cap s) = 0, \quad \#(Q_5 \cap s) = 0.$$

*Por tanto, podemos concluir que*

$$n(s, q, T^*) = \sum_{i=1}^{5} \mu_{Q_i}(q) \cdot \#(Q_i \cap s) = 1 \cdot 0.6164 = 0.6164.$$

A continuación desarrollaremos la adaptación realizada para el término $\frac{f(s|q)}{f(s|q^*)}$, para todo $s$. El término $f(s|q)$ lo obtendremos de modo análogo al caso clásico, por lo que nos centraremos en el término $f(s|q^*)$, para lo que seguiremos un procedimiento similar al realizado en el término $n(s, q, T^*)$.

**Definición 3.14** *Sea $s \in S$ y $q \in Q$, donde $S$ y $Q$ denotan los atributos sensibles y no sensibles respectivamente. Supongamos que la partición difusa de $Q$ sea la dada por los conjuntos $\{Q_1, \ldots, Q_n\}$, entonces*

$$f(s|q^*) = \sum_{i=1}^{r} \mu_{Q_i}(q) \cdot f(s|Q_i), \tag{3.6}$$

*donde*

$$f(s|Q_i) = \frac{f(s \cap Q_i)}{f(Q_i)} = \frac{\#(s \cap Q_i)}{\#Q_i} \quad \forall i \in \{1, \ldots, n\}. \tag{3.7}$$

**Teorema 3.15** *Sea $i$ un número natural cualquiera entre 1 y $n$, se tiene que la función $f(\cdot|Q_i)$ definida en (3.7) es una probabilidad.*

**Demostración:** Fijado un valor de $i$ cualquiera, veamos que $f(\cdot|Q_i)$ es una probabilidad. Para ello, comprobaremos que cumple los tres axiomas de Kolmogorov.

- $f(s|Q_i) \geq 0$: Obvio por construcción, ya que $\#Q_i$ siempre es mayor que 0, y el numerador ha de ser mayor o igual que 0.

- $f(S|Q_i) = 1$:
$$f(S|Q_i) = \frac{\#(S \cap Q_i)}{\#Q_i} = \frac{\#Q_i}{\#Q_i} = 1.$$

- Sea $(A_n)_n \subset S$, disjuntos dos a dos, veamos que $f(\cup_n A_n|Q_i) =$

$\sum_n f(A_n|Q_i)$, puesto que

$$\begin{aligned} f(\cup_n A_n|Q_i) &= \frac{\#(\bigcup_n A_n \cap Q_i)}{\#Q_i} = \frac{\#(\bigcup_n (A_n \cap Q_i))}{\#Q_i} \\ &= \frac{\sum_n \#(A_n \cap Q_i)}{\#Q_i} = \sum_n \frac{\#(A_n \cap Q_i)}{\#Q_i} = \sum_n f(A_n|Q_i). \end{aligned}$$

Tal como queríamos demostrar. ■

**Ejemplo 3.16** *Sigamos con los datos del Ejemplo 3.13, donde* $q = 47905 \times 43 \times 7$ *y* $s =$*gastritis. Calculemos*

$$\sum_{i=1}^{5} \mu_{Q_i}(q) \cdot f(s|Q_i) = \sum_{i=1}^{5} \mu_{Q_i}(q) \cdot \frac{\#(s \cap Q_i)}{\#Q_i},$$

*para ello necesitaremos* $\mu_{Q_i}(q)$, $\#(s \cap Q_i)$ *y* $\#Q_i$, $\forall i \in \{1, \ldots, 5\}$.

*En el Ejemplo 3.13 tenemos los dos primeros términos para cada* $i$, *así que sólo necesitamos obtener* $\#Q_i$, *lo que haremos de forma análoga a* $\#(s \cap Q_i)$. *Por ejemplo, para el caso* $i = 1$, *obtendremos lo siguiente:*

$$\begin{aligned} \mu_{Q_1}(1) &= \mu_{476**}(1) \cdot \mu_{Baja}(1) \cdot \mu_{Bajo}(1) = 1 \cdot 1 \cdot 0.5 = 0.5, \\ \mu_{Q_1}(2) &= \mu_{476**}(2) \cdot \mu_{Baja}(2) \cdot \mu_{Bajo}(2) = 1 \cdot 1 \cdot 0.5 = 0.5, \\ \mu_{Q_1}(j) &= 0, \quad \forall j = \{3, \ldots, 12\}. \end{aligned}$$

*Como podemos ver, serán los individuos* 1 *y* 2 *los únicos que no obtengan una función de pertenencia* 0, *por lo que serán los únicos a tratar.*

*A continuación, procedemos a ordenarlos de forma decreciente en función de la pertenencia a* $Q_1$, *lo que coincide con el orden original, es decir,* $x_1 = 1$ *y* $x_2 = 2$.

*Con esto procedemos a calcular* $j$ *como en (2.10):*

$$j = max\{1 \leq s \leq 12 | \mu_{(s-1)} + \mu_{(s)} > 1\} = 1.$$

*Finalmente, como* $\mu_{(j)} = \mu_{(1)} = \mu_{Q_1}(x_1) = \mu_{Q_1}(1) = 0.5 \geq 0.5$*, se tiene que* $\#Q_1 = 1$.

*De manera análoga, obtenemos:*

$$\#Q_2 = 2, \quad \#Q_3 = 3, \quad \#Q_4 = 3, \quad \#Q_5 = 1,$$

*con lo que:*

$$\sum_{i=1}^{5} \mu_{Q_i}(q) \cdot \frac{\#(s \cap Q_i)}{\#Q_i} = 0.5 \cdot \frac{1}{1} + 0.5 \cdot \frac{1}{2} + 0 = 0.5 + 0.25 = 0.75.$$

Teniendo en cuenta los resultados previos podemos definir las probabilidades a posteriori en el caso difuso como sigue:

**Definición 3.17** *Sea* $s \in S$ *y* $q \in Q$*, donde* $S$ *y* $Q$ *denotan el atributo sensible y los atributos no sensibles respectivamente. Sea* $T^*$ *la tabla obtenida por la partición difusa de* $Q$ *dada por los conjuntos* $\{Q_1, \ldots, Q_n\}$*. Entonces:*

$$\beta_{(q,s,T^*)} = \frac{(\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \#(Q_i \cap s)) \dfrac{f(s|q)}{\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \dfrac{\#(Q_i \cap s)}{\#Q_i}}}{\sum_{s' \in S} \{(\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \#(Q_i \cap s')) \dfrac{f(s'|q)}{\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \dfrac{\#(Q_i \cap s')}{\#Q_i}}\}}. \quad (3.8)$$

**Ejemplo 3.18** *Continuemos con el mismo ejemplo, donde* $q = 47905 \times 43 \times 7$*, y* $s =$*gastritis. Vamos a buscar el valor de* $\beta_{(q,s,T^*)}$*, para lo que calcularemos el término:*

$$g(s) := \left( \sum_{i=1}^{5} \mu_{Q_i}(q) \cdot \#(Q_i \cap s) \right) \frac{f(s|q)}{\sum_{i=1}^{5} \mu_{Q_i}(q) \cdot \dfrac{\#(Q_i \cap s)}{\#Q_i}},$$

*para cada* $s \in S$.

*Para* $s =$*gastritis, por el Ejemplo 3.13, tenemos que:*

$$\sum_{i=1}^{5} \mu_{Q_i}(q) \cdot \#(Q_i \cap s') = 0.6164,$$

*mientras que por el Ejemplo 3.16 se obtiene:*

$$\sum_{i=1}^{5} \mu_{Q_i}(q) \cdot \frac{\#(s \cap Q_i)}{\#Q_i} = 0.75.$$

*Además, por la Tabla 3.1, podemos obtener que* $f(s|q) = 1$.

*Por tanto, podemos concluir que:*

$$g(s) = 0.6164 \cdot \frac{1}{0.75} = 0.82187.$$

*Además, por los datos de la Tabla 3.1, tenemos que* $f(s'|q) = 0$*, para todo* $s'$ *distinto de gastritis y así* $g(s') = 0$*, de donde:*

$$\beta_{(q,s,T^*)} = \frac{g(s)}{\sum_{s' \in S} g(s')} = \frac{g(s)}{g(s)} = 1.$$

El hecho de que, por simplicidad, se haya considerado un ejemplo con un número de individuos en el estudio no demasiado elevado, es el motivo por el que se ha obtenido que $\beta_{(q,s,T^*)} = 1$.

Con el fin de comprobar como influye las condiciones del problema en la aplicación de la Definición 3.17, a continuación presentamos un ejemplo donde tenemos dos atributos, uno no sensible (Edad) y otro sensible (Enfermedad) y donde el número de individuos es mayor que en los ejemplos previos.

**Ejemplo 3.19** *Supongamos que los datos de la Tabla 3.12 son de un grupo de individuos de edades entre* 21 *y* 24 *años con enfermedades relacionadas con el estómago.*

| | Nº individuos | Edad | Enfermedad |
|---|---|---|---|
| 1 | 10 | 21 | gastritis |
| 2 | 15 | 21 | úlcera gástrica |
| 3 | 25 | 21 | cáncer de estómago |
| 4 | 5 | 22 | gastritis |
| 5 | 5 | 22 | úlcera gástrica |
| 6 | 15 | 22 | cáncer de estómago |
| 7 | 15 | 23 | gastritis |
| 8 | 5 | 23 | úlcera gástrica |
| 9 | 5 | 23 | cáncer de estómago |
| 10 | 10 | 24 | gastritis |
| 11 | 10 | 24 | úlcera gástrica |
| 12 | 30 | 24 | cáncer de estómago |
| 13 | 10 | 25 | gastritis |
| 14 | 5 | 25 | úlcera gástrica |
| 15 | 10 | 25 | cáncer de estómago |

Tabla 3.12: Tabla con los datos correspondientes a individuos con enfermedades del estómago.

*Sea la partición difusa de la Edad la formada por los conjuntos difusos triangulares:* $Q_1 = (-\infty, 20, 23)$, $Q_2 = (20, 23, 26)$, $Q_3 = (23, 26, \infty)$. *Buscaremos el valor de* $\beta(q, s, T^*)$ *para* $q = 21$ *y* $s = s_g$ =*gastritis.*

$$f(s_g|q) = \frac{10}{50} = \frac{1}{5}, \quad f(s_u|q) = \frac{15}{50} = \frac{3}{10}, \quad f(s_c|q) = \frac{25}{50} = \frac{1}{2},$$

$$\mu_{Q_1}(q) = 0.667, \quad \mu_{Q_2}(q) = 0.333, \quad \mu_{Q_3}(q) = 0.$$

*Nos queda por obtener* $\#Q_i$ *y* $\#(Q_i \cap s)$, $\forall i$ *y* $\forall s$.

$\#Q_1$ :

$$j = max\{1 \leq s \leq 175 | \mu_{Q_1}(x_{s-1}) + \mu_{Q_1}(x_s) > 1\} = 50,$$

$$\mu_{Q_1}(x_{50}) = 0.667 \geq 0.5 \Rightarrow \#Q_1 = j = 50.$$

$\#Q_2$ :

$$j = max\{1 \leq s \leq 175 | \mu_{Q_2}(x_{s-1}) + \mu_{Q_2}(x_s) > 1\} = 100,$$
$$\mu_{Q_2}(x_{100}) = 0.667 \geq 0.5 \Rightarrow \#Q_2 = j = 100.$$

$\#Q_3$ :

$$j = max\{1 \leq s \leq 175 | \mu_{Q_3}(x_{s-1}) + \mu_{Q_3}(x_s) > 1\} = 25,$$
$$\mu_{Q_3}(x_1) = 0.667 \geq 0.5 \Rightarrow \#Q_3 = j = 25.$$

$\#(Q_1 \cap s_g)$ :

$$j = max\{1 \leq s \leq 175 | \mu_{(Q_1 \cap s_g)}(x_{s-1}) + \mu_{(Q_1 \cap s_g)}(x_s) > 1\} = 10,$$
$$\mu_{Q_1}(x_{10}) = 0.667 \geq 0.5 \Rightarrow \#(Q_1 \cap s_g) = j = 10.$$

*De modo análogo:*

$$\begin{array}{lll} \#(Q_1 \cap s_g) = 10, & \#(Q_1 \cap s_u) = 15, & \#(Q_1 \cap s_c) = 25, \\ \#(Q_2 \cap s_g) = 30, & \#(Q_2 \cap s_u) = 20, & \#(Q_2 \cap s_c) = 50, \\ \#(Q_3 \cap s_g) = 10, & \#(Q_3 \cap s_u) = 5, & \#(Q_3 \cap s_c) = 10. \end{array}$$

*Así,*

$$\beta_{(q,s,T^*)} = \frac{g(s_g)}{g(s_g) + g(s_u) + g(s_c)},$$

*donde*

$$g(s) = \left( \sum_{i=1}^{3} \mu_{Q_i}(q) \cdot \#(Q_i \cap s') \right) \frac{f(s'|q)}{\sum_{i=1}^{3} \mu_{Q_i}(q) \cdot \frac{\#(Q_i \cap s')}{\#Q_i}}.$$

*Calculemos $g(s_g)$, $g(s_u)$ y $g(s_c)$:*

$$\begin{aligned} g(s_g) &= \left(\frac{2}{3}\cdot 10+\frac{1}{3}\cdot 30+0\cdot 10\right)\cdot \frac{\frac{1}{5}}{\left(\frac{2}{3}\cdot\frac{10}{50}+\frac{1}{3}\cdot\frac{30}{100}+0\cdot\frac{10}{25}\right)} = \frac{100}{7}, \\ g(s_u) &= \left(\frac{2}{3}\cdot 15+\frac{1}{3}\cdot 20+0\cdot 5\right)\cdot \frac{\frac{3}{10}}{\left(\frac{2}{3}\cdot\frac{15}{50}+\frac{1}{3}\cdot\frac{20}{100}+0\cdot\frac{5}{25}\right)} = \frac{75}{4}, \\ g(s_c) &= \left(\frac{2}{3}\cdot 25+\frac{1}{3}\cdot 50+0\cdot 10\right)\cdot \frac{\frac{1}{2}}{\left(\frac{2}{3}\cdot\frac{25}{50}+\frac{1}{3}\cdot\frac{50}{100}+0\cdot\frac{10}{25}\right)} = \frac{100}{3}. \end{aligned}$$

*Finalmente, se tiene que:*

$$\beta_{(q,s,T^*)} = \frac{g(s_g)}{g(s_g)+g(s_u)+g(s_c)} = \frac{\frac{100}{7}}{\frac{100}{7}+\frac{75}{4}+\frac{100}{3}} = 0.2152.$$

Tras este ejemplo con unos datos más completos, continuaremos con el desarrollo de la $l$-diversity para los conjuntos difusos, de manera similar al caso con conjuntos convencionales, adaptándolo con lo que hemos obtenido hasta el momento.

Imaginemos que el atacante pretende obtener el valor del atributo sensible mediante "positive disclosure", es decir, que $t[S] = s$ con una alta probabilidad. Por (3.8), esto sólo ocurre cuando:

$$\exists s, \forall s' \neq s, g(s') \ll g(s). \tag{3.9}$$

Lo ocurrido en (3.9), puede deberse a dos factores, una falta de diversidad en el atributo sensible y/o una gran información adicional del atacante, los cuales analizaremos a continuación.

- **Falta de diversidad**: la falta de diversidad en el atributo sensible se manifiesta del siguiente modo,

$$\forall s' \neq s,\ \sum_{i=1}^{n} \mu_{Q_i}(q)\cdot \#(Q_i \cap s') \ll \sum_{i=1}^{n} \mu_{Q_i}(q)\cdot \#(Q_i \cap s). \tag{3.10}$$

  Es decir, casi todas las tuplas tienen un mismo valor sensible $s$, y por tanto $\beta_{(q,s,T^*)} \approx 1$.

Una forma de solucionar esto es, haciendo que las generalizaciones realizadas permitan que haya $l \geq 1$ opciones de atributo sensible en cada conjunto de la partición difusa realizada, y que estén lo suficientemente representadas, aunque al requerir esto, nos podemos encontrar con clases demasiado heterogéneas.

- **Gran cantidad de información adicional**: incluso cuando tenemos la premisa anterior, con la información adicional del atacante nos podemos encontrar con que puede eliminar opciones del atributo cuando ocurre lo siguiente:

$$\exists s', \frac{f(s'|q)}{\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \frac{\#(Q_i \cap s')}{\#Q_i}} \approx 0. \tag{3.11}$$

Lógicamente, si tenemos $l$ valores bien representados, el atacante necesitará eliminar $l-1$ valores del atributo sensible para llegar a "positive disclosure", por lo que a mayor $l$, mayor protección ante este tipo de ataques.

Para finalizar este apartado, presentamos el principio de $l$-diversity para particiones difusas:

**Principio 3.20** *Decimos que una tabla $T$ cuyos bloques de atributos no sensibles $\{Q_1, \ldots, Q_n\}$ vienen determinados mediante particiones difusas es $l$-diversa si para cada $Q_i$, existen al menos $l$ valores del atributo sensible suficientemente representados.*

Para solucionar los problemas que sigue presentando la $l$-diversity frente a ataques de similitud de valores en el atributo sensible, procedemos a definir la técnica $t$-closeness al caso en el que utilizamos particiones difusas.

## 3.3. Extensión de la $t$-closeness

Al igual que en el caso de los conjuntos clásicos, la $t$-closeness para conjuntos difusos mejora la protección que nos da la $l$-diversity en ciertos

aspectos, como puede suceder en los casos en los que los diferentes atributos sensibles de un mismo bloque, aun siendo distintos, sean similares, y el atacante puede obtener una información relevante.

Como hemos visto previamente, la $t$-closeness busca minimizar la distancia entre la distribución del atributo sensible en la tabla completa y la distribución del atributo sensible en el bloque al que pertenece el individuo a investigar. Sin embargo, en el caso difuso, dicho elemento no pertenecerá a un bloque, sino que tendrá una pertenencia relacionada a cada conjunto difuso de la partición.

Para poder solucionar este problema, se ha desarrollado el siguiente teorema:

**Teorema 3.21** *Sean $A_1, \ldots, A_n$ los distintos atributos no sensibles y $S$ el atributo sensible de una tabla de datos, $S = \{s_1, \ldots, s_m\}$ los diferentes valores del atributo sensible, $Q_1, \ldots, Q_r$ una partición difusa de los atributos no sensibles, $Q = (q_1, \ldots, q_m)$ la distribución del atributo sensible en la tabla completa, $P_1, \ldots, P_r$ la distribución del atributo sensible en los conjuntos $Q_1, \ldots, Q_r$ respectivamente, donde $P_i = (p_1^i, \ldots, p_m^i)$, $\forall i = 1, \ldots, r$, y $\mu_{Q_i}$ la función de pertenencia de $Q_i$ para cada $i$. Entonces, para todo $x$, individuo de la tabla de datos, se tiene que:*

$$P := \sum_{i=1}^{r} \mu_{Q_i}(x) \cdot P_i, \tag{3.12}$$

*es una distribución de probabilidad.*

**Demostración:**

Para demostrar el resultado, comprobemos que $p_i \geq 0$, $\forall i = 1, \ldots, m$ y $\sum_{i=1}^{m} p_i = 1$.

- $p_i \geq 0, \quad \forall i = 1, \ldots, m$

Por definición de función de pertenencia, $\mu_{Q_i}$ es no negativa para cada $i$, y además $p_i^j \geq 0$ para todo $i$, $j$ por ser probabilidades. Por tanto,

$$p_j = \sum_{i=1}^{r} \mu_{Q_i}(x) \cdot p_j^i \geq 0, \quad \forall j = 1, \ldots, m.$$

- $\sum_{i=1}^{m} p_i = 1$

  Por ser $P_j$ distribuciones de probabilidad, sabemos que $\sum_{i=1}^{m} p_i^j = 1$ y por la definición de partición difusa, tenemos que para todo $x$ se tiene que

  $$\sum_{i=1}^{r} \mu_{Q_i}(x) = 1.$$

  Así

  $$\begin{aligned} \sum_{j=1}^{m} p_j &= \sum_{j=1}^{m}\sum_{i=1}^{r} \mu_{Q_i}(x) \cdot p_j^i = \sum_{i=1}^{r}\sum_{j=1}^{m} \mu_{Q_i}(x) \cdot p_j^i = \\ &= \sum_{i=1}^{r} \mu_{Q_i}(x) \cdot (\sum_{j=1}^{m} p_j^i) = \sum_{i=1}^{r} \mu_{Q_i}(x) = 1. \end{aligned}$$

Por tanto, tenemos que $P$ es una distribución de probabilidad. ■

El valor $P$ obtenido en el teorema anterior será el que utilizaremos para aplicar la misma técnica que en el caso de conjuntos clásicos, con la distancia EMD.

**Ejemplo 3.22** *Sean los datos dados por la Tabla 3.13.*

| Individuo | Edad | Salario(k) | Enfermedad |
|---|---|---|---|
| 1 | 29 | 3 | úlcera gástrica |
| 2 | 22 | 4 | gastritis |
| 3 | 27 | 5 | cáncer de estómago |
| 4 | 43 | 6 | gastritis |
| 5 | 52 | 11 | gripe |
| 6 | 47 | 8 | bronquitis |
| 7 | 30 | 7 | bronquitis |
| 8 | 36 | 9 | pneumonía |
| 9 | 32 | 10 | cáncer de estómago |

Tabla 3.13: Tabla de datos para el ejemplo de t-closeness difusa.

*En este ejemplo, trataremos como atributo no sensible la edad, mientras que los atributos sensibles serán el salario y la enfermedad. Para la Edad, una partición difusa viene dada por los siguientes conjuntos:*

$$Q_1 = (-\infty, 25, 30),\ Q_2 = (25, 30, 35, 40),\ Q_3 = (35, 40, \infty).$$

*Con esto, obtenemos la generalización dada por la Tabla 3.14.*

| Edad | Salario(k) | Enfermedad |
|---|---|---|
| $Q_1$ | 4 | gastritis |
| | 5 | cáncer de estómago |
| $Q_2$ | 3 | úlcera gástrica |
| | 7 | cáncer de estómago |
| | 9 | bronquitis |
| | 10 | pneumonía |
| $Q_3$ | 6 | gastritis |
| | 8 | gripe |
| | 11 | bronquitis |

Tabla 3.14: Tabla de datos generalizados para el ejemplo de t-closeness difusa.

*En primer lugar, haremos el ejemplo para el atributo sensible Salario,*

*el cual es cuantitativo, abordando a continuación el de Enfermedad, el cual es cualitativo.*

*Los valores que toma el atributo Salario en la tabla son* $S = \{3, 4, 5, 6, 7, 8, 9, 10, 11\}$.

*Los valores que corresponden a cada bloque son* $S_1 = \{4, 5\}$, $S_2 = \{3, 7, 9, 10\}$ *y* $S_3 = \{6, 8, 11\}$.

*Tanto en* $S$ *como en* $S_i$ *tenemos que cada valor es equiprobable, y de ahí obtenemos las distribuciones* $Q$ *y* $P_i$.

*Ahora, debemos calcular* $P$ *como en (3.12) para cada individuo de la tabla, en este caso, tenemos* 9 *individuos.*

$$\begin{aligned} P_{x_1} &= \sum_{i=1}^{3} \mu_{Q_i}(x_1) \cdot P_i = \mu_{Q_1}(x_1) \cdot P_1 + \mu_{Q_2}(x_1) \cdot P_2 + \mu_{Q_3}(x_1) \cdot P_3 = \\ &= 0.2 \cdot P_1 + 0.8 \cdot P_2 + 0 \cdot P_3 = 0.2 \cdot P_1 + 0.8 \cdot P_2. \end{aligned}$$

*De modo análogo obtenemos el resto:*

$$\begin{array}{llll} P_{x_1} = 0.2 \cdot P_1 + 0.8 \cdot P_2, & P_{x_2} = P_1, & P_{x_3} = 0.6 \cdot P_1 + 0.4 \cdot P_2, & P_{x_4} = P_3, \\ P_{x_5} = P_3, & P_{x_6} = P_3, & P_{x_7} = P_2, & P_{x_8} = 0.8 \cdot P_2 + 0.2 \cdot P_3, \quad P_{x_9} = P_2. \end{array}$$

*Ahora, tenemos que calcular* $D[P_i, Q]$ *para cada* $i$*, esto es:*

$$D[P_{x_1}, Q] = \frac{1}{8} \sum_{i=1}^{9} |\sum_{j=1}^{i} (p_j^1 - q_j)|.$$

*Realizando los cálculos correspondientes, obtenemos los siguientes resultados:*

$$\begin{array}{lll} D[P_{x_1}, Q] = 0.065278, & D[P_{x_2}, Q] = 0.34028, & D[P_{x_3}, Q] = 0.17778, \\ D[P_{x_4}, Q] = 0.16667, & D[P_{x_5}, Q] = 0.16667, & D[P_{x_6}, Q] = 0.16667, \\ D[P_{x_7}, Q] = 0.10069, & D[P_{x_8}, Q] = 0.091667, & D[P_{x_9}, Q] = 0.10069. \end{array}$$

*Como el mínimo de los valores es el obtenido en* $D[P_{x_1}, Q] = 0.065278$*, tenemos que la Tabla 3.14 tiene* $0.065278$*-closeness para el atributo Salario.*

*Ahora, procedamos para el atributo Enfermedad, que al ser un atributo cualitativo, nos obliga a modificar la estrategia del cálculo.*

*Las enfermedades que tenemos en la tabla son:* $E = \{$`úlcera gástrica, gastritis, cáncer de estómago, gripe, bronquitis, pneumonía`$\}$, *y que por simplicidad denotaremos como* $E = \{UG, GA, CE, GR, BR, PN\}$.

*Las enfermedades que corresponden a cada bloque son:* $E_1 = \{GA, CE\}$, $E_2 = \{UG, CE, BR, PN\}$ *y* $E_3 = \{GA, GR, BR\}$. *Las distribuciones asociadas serán equiprobables para los valores que toma.*

*Con las nuevas* $P_i$, *tenemos que* $P_{x_j}$ *son de la misma forma que las vistas para el caso del salario, cambiando las distribuciones correspondientes.*

*Ahora, para calcular* $D[P_i, Q]$ *utilizaremos la expresión:*

$$D[P_i, Q] = \frac{1}{2} \sum_{i=1}^{6} |p_i - q_i|,$$

*lo que nos permite obtener los siguientes resultados:*

$$\begin{array}{ccc} D[P_{x_1}, Q] = 0.23333, & D[P_{x_2}, Q] = 0.66667, & D[P_{x_3}, Q] = 0.36667, \\ D[P_{x_4}, Q] = 0.5, & D[P_{x_5}, Q] = 0.5, & D[P_{x_6}, Q] = 0.5, \\ D[P_{x_7}, Q] = 0.33333, & D[P_{x_8}, Q] = 0.2, & D[P_{x_9}, Q] = 0.33333. \end{array}$$

*Como el mínimo de los valores es el obtenido en* $D[P_{x_8}, Q] = 0.2$, *tenemos que la Tabla 3.14 tiene* 0.2*-closeness para el atributo Enfermedad.*

Con todo esto, hemos conseguido adaptar la técnica $t$-closeness que habíamos definido para conjuntos convencionales, al caso de conjuntos difusos.

La utilidad práctica de todas las técnicas desarrolladas en este capítulo, será puesta de manifiesto en el siguiente, en el que se analizará el uso de particiones difusas sobre una base de datos real.

# Capítulo 4

# Experimentación

En los capítulos anteriores, hemos desarrollado los fundamentos teóricos del trabajo, donde se han estudiado tanto las particiones clásicas (Capítulo 2), como las que hemos obtenido mediante la utilización de particiones difusas (Capítulo 3).

Como hemos podido ver, hemos comprobado que las definiciones de la $l$-diversity y la $t$-closeness son consistentes en el caso difuso, donde hemos observado que tiene sentido pensar en dichos criterios para el caso en el que utilizamos particiones difusas.

Además, tenemos una técnica nueva ($Q$-anonymity) que si bien no es directamente relacionable a la $k$-anonymity, si podemos utilizarla para comparar métodos de particiones difusas distintos.

Nuestro objetivo en este apartado será la experimentación sobre datos reales para comprobar que las particiones difusas previamente definidas, presentan una seria alternativa a las particiones clásicas conocidas a la hora de proteger la privacidad de los datos a enmascarar.

Para ello, sobre una base de datos real hemos llevado a cabo varias pruebas sobre distintos cuasi-identificadores. Para cada uno de éstos, hemos realizado distintos tipos de particiones, tanto nítidas como difusas, y aplicado varios criterios como los descritos a lo largo de los dos capítulos

anteriores. Con dichos criterios, hemos podido realizar las comparaciones para ver qué métodos son los que preservan la privacidad de una forma más eficaz.

Para la experimentación que hemos realizado, hemos utilizado dos programas distintos: *R* (*Versión 2.15.2*) y *Matlab* (*Versión R2010b*).

Los datos con la que hemos trabajado están en la base *CENSUS*, obtenida el 27 de julio de 2000, utilizando el sistema de extracción de datos del censo de Estados Unidos (*Data Extraction System of the U. S. Bureau of the Census*). Estos datos están disponibles en el paquete *sdcMicro* de *R* (ver [20]) bajo el nombre de *CASCrefmicrodata*.

En dichos datos, nos encontramos un total de 1080 individuos, donde a cada uno se le atribuye un valor para **13 atributos distintos**:

1. AFNLWGT: Peso final.
2. AGI: Ingresos brutos ajustados.
3. EMCONTRB: Cotización de los empresarios para el seguro de salud.
4. ERNVAL: Ganancias netas de negocio o finca.
5. FEDTAX: Impuestos federales.
6. FICA: Impuesto sobre la nómina de jubilación.
7. INTVAL: Cuantía de ingresos por intereses.
8. PEARNVAL: Ganancias totales persona.
9. POTHVAL: Ingresos totales de otras personas.
10. PTOTVAL: Ingresos totales.
11. STATETAX: Impuesto sobre la renta estatal.
12. TAXINC: Cantidad del ingreso imponible.
13. WALVAL: Salario total.

## 4.1. Métodos para particionar un conjunto de datos

A lo largo de la experimentación, hemos utilizado diversos métodos para particionar los datos. Distinguiremos **dos tipos**, los métodos para **particiones nítidas**, y los métodos para **particiones difusas**.

### 4.1.1. Métodos nítidos

Para las particiones nítidas, hemos utilizado dos métodos distintos. Por un lado, el método de las *k-medias*, y por el otro, el método *Global Recode*.

#### 4.1.1.1. Método de las k-medias

Se trata de un método iterativo, en el cual se debe proporcionar el número de conjuntos en el que queremos particionar los datos ($k$ conjuntos). Los pasos fundamentales del método de las *k-medias* (ver [2]) son:

- **Paso 1**: Seleccionar $k$ elementos aleatoriamente como centros de los conjuntos.
- **Paso 2**: Calcular la distancia de todos los individuos a cada centro de los conjuntos.
- **Paso 3**: Asignar cada individuo al conjunto cuya distancia sea menor.
- **Paso 4**: Recalcular los centros de los conjuntos.
- **Paso 5**: Repetir los pasos 2, 3 y 4 hasta que no haya movimiento de individuos entre los distintos conjuntos, en cuyo caso, será esa la partición de $k$ conjuntos obtenida.

Este método es muy común, y podemos encontrarlo implementado en varios programas distintos. Nosotros, hemos utilizado el que nos ofrece $Matlab$ con la función $kmeans$ (ver [5]).

Este programa necesita como datos de **entrada**:

- la matriz de datos con la que vamos a trabajar,
- el número de conjuntos que queremos que nuestra partición tenga.

Como **salida** obtenemos:

- un vector de longitud el número de individuos, que asigna a cada uno al conjunto al que pertenece,
- los centros de cada uno de los conjuntos generados.

**Ejemplo 4.1** *Veamos a continuación un ejemplo de partición con el método de las k-medias. Hemos utilizado unos datos de dos variables donde tenemos un total de 133 individuos, y hemos hecho la partición en 4 conjuntos.*

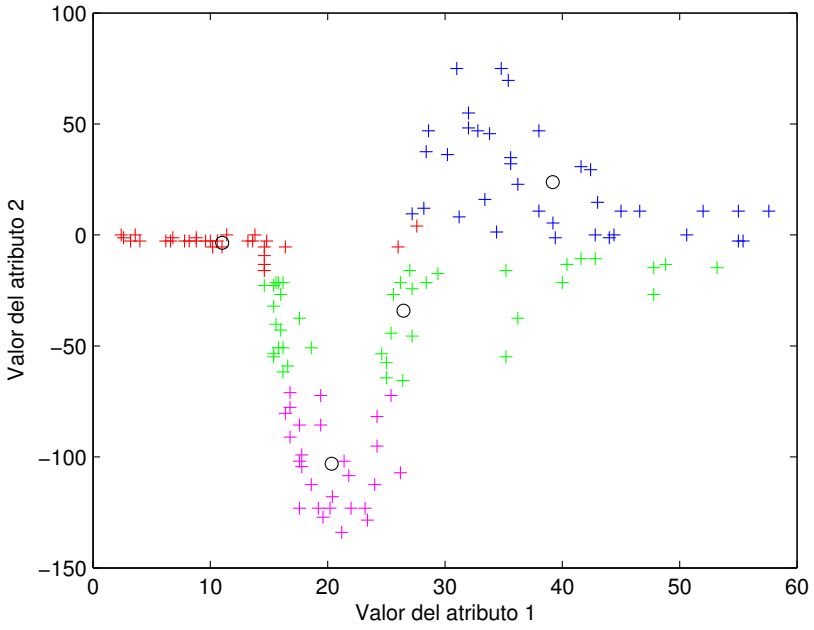


Figura 4.1: Ejemplo de partición con el método *k-medias*.

*Podemos ver lo obtenido en la Figura 4.1, con el atributo 1 representado en el eje $X$ y el atributo 2 en el eje $Y$, donde cada color representa los individuos pertenencientes a un mismo conjunto de la partición obtenida, y los 4 círculos negros, representan los centros de dichos conjuntos.*

#### 4.1.1.2. Método Global Recode

Este método construye intervalos, asignando a cada individuo al conjunto al que pertenence, como hemos hecho en los casos nítidos a lo largo del trabajo. Existen varias implementaciones de dicho método. Nosotros hemos elegido la existente en $R$, en el paquete *sdcMicro* (ver [20]). Dentro de dicho programas, tenemos tres posibles métodos a utilizar:

- *equidistant*: construye conjuntos de una misma longitud.
- *logEqui*: contruye conjuntos de una misma longitud para los datos transformados mediante logaritmo.
- *equalAmount*: construye los conjuntos de tal forma que haya aproximadamente un mismo número de individuos en cada uno de éstos.

Entre los tres posibles métodos, hemos seleccionado el tercero, *equalAmount*, el cual realiza dichas particiones de forma que los conjuntos tengan un número de individuos similar, habiendo visto previamente que dicho método mejora a los otros dos disponibles.

Como argumentos de **entrada** necesitamos:

- método utilizado (*equidistant, logEqui, equalAmount*),
- el vector de datos con el que queremos trabajar,
- el número de conjuntos que queremos que tenga la partición.

Por otro lado, como **salida** obtenemos:

- un vector de tamaño el número de individuos, que asigna a cada uno al conjunto al que pertenece, o bien si se ha eliminado dicho individuo de los datos por tratarse de un valor atípico (*outlier* en inglés), que pudiera suponer un gran cambio en la partición.

**Ejemplo 4.2** *Considerando los mismos datos como ejemplo para el caso del método de las k-medias, donde tenemos 133 individuos con dos atributos distintos. Realizamos la partición con el método Global Recode, y obtenemos la representación gráfica en la Figura 4.2, donde los distintos colores representan el conjunto al que se han asociado cada uno de los individuos.*

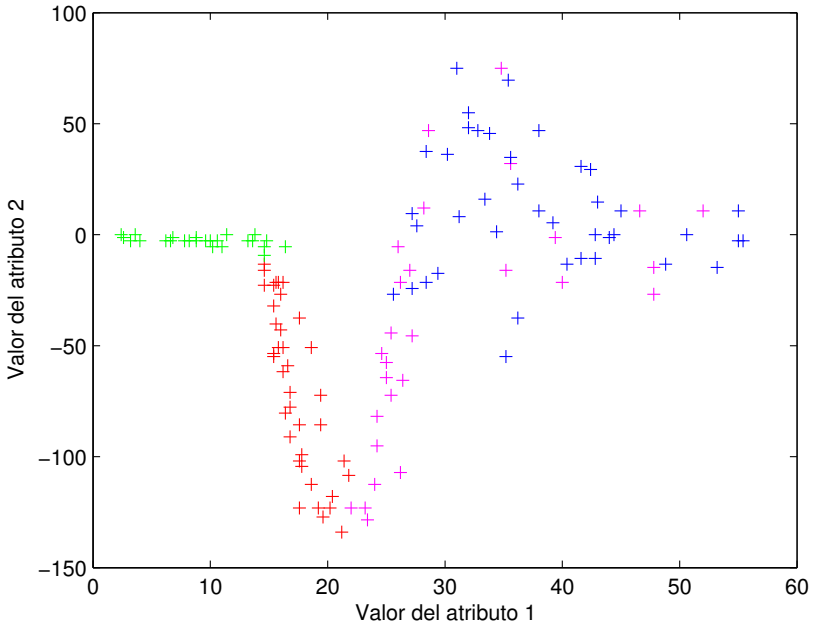


Figura 4.2: Ejemplo de partición con el método *Global Recode*.

*En este caso, a diferencia del método de las k-medias, el algoritmo ha eliminado uno de los individuos de los datos para obtener la partición óptima, ya que se trata de un individuo cuya utilización puede provocar un gran cambio en la partición obtenida. El individuo eliminado es el que tiene*

*mayor valor para el atributo 1, representado en el eje $X$. Se trata de un valor atípico (o outlier).*

Conviene observar que hemos obtenido distintas particiones por ambos métodos, procediendo el segundo de ellos (*Global Recode*) a la eliminación de un individuo al considerarlo un valor atípico.

### 4.1.2. Métodos difusos

Para obtener las particiones difusas, se han seleccionado dos métodos. Por un lado, el método de las *c-medias difuso*, y por otro, el método de *Gustafson-Kessel*.

#### 4.1.2.1. Método de las c-medias difuso

El método de las *c-medias difuso* (ver [17]) busca la minimización de la función objetivo, llamada *funcional de las c-medias*, definida del siguiente modo:

$$J(X;U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m||x_k - v_i||_A^2, \tag{4.1}$$

donde:

- $X$: el conjunto de datos,
- $U = [\mu_{ik}]$: la matriz de pertenencias de cada individuo a cada conjunto,
- $V = [v_1, \ldots, v_c]$: siendo $v_i$ los centros de los conjuntos,
- $m$: el exponente que determina la borrosidad (*fuzziness*) de los conjuntos de la partición, el cual debe ser un natural.

Finalmente, la norma utilizada viene dada por la siguiente expresión:

$$D_{ikA}^2 = ||x_k - v_i||_A^2 = (x_k - v_i)^T A(x_k - v_i), \tag{4.2}$$

donde la matriz que induce dicha norma en el programa que hemos utilizado es la matriz identidad, $A = I$.

La minimización de dicho funcional representa un problema de optimización no lineal.

El programa que hemos utilizado para resolver dicho problema, ha sido el algoritmo *FCMclust* del paquete *Fuzzy Clustering and Data Analysis Toolbox* de *Matlab* (ver [1]).

Este programa necesita como argumentos de **entrada**:

- la matriz de datos, previamente normalizada mediante el programa *clust_normalize* del mismo paquete (ver [1]), el cual realiza la normalización de los individuos en el intervalo $[0, 1]$,
- una serie de parámetros que son:
  - el número de conjuntos,
  - el exponente que determina la borrosidad de los conjuntos,
  - la tolerancia del método.

Como **salida** obtenemos:

- la matriz con las pertenencias de cada individuo a los distintos conjuntos generados,
- los centros de los conjuntos.

**Ejemplo 4.3** *Llevaremos a cabo un ejemplo del método de las c-medias difuso, con los mismos datos que hemos utilizado anteriormente.*

*Podemos ver la representación gráfica en la Figura 4.3, donde tenemos en el eje $X$ el valor del atributo 1 normalizado, y en el eje $Y$ el valor del atributo 2 normalizado.*

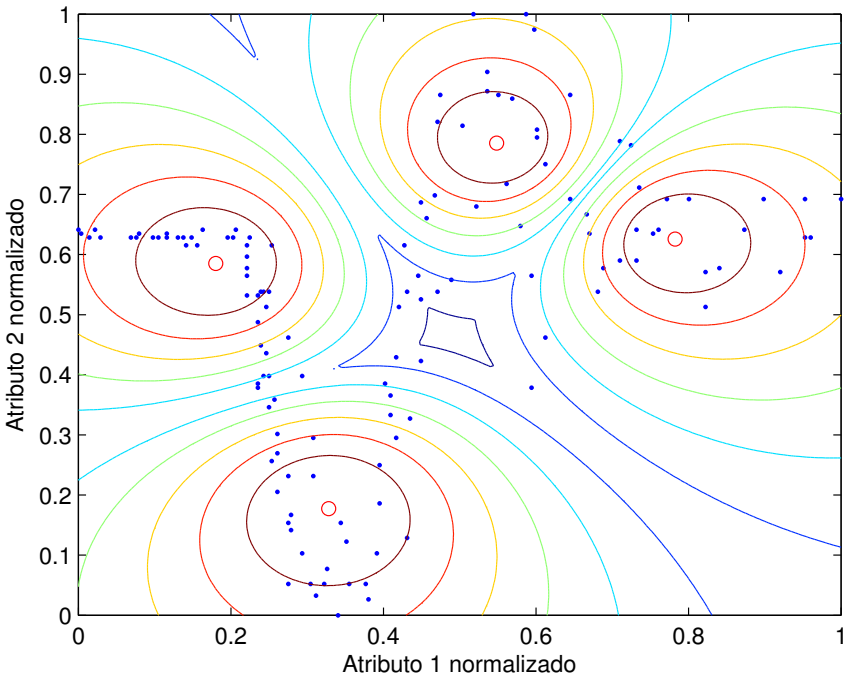


Figura 4.3: Ejemplo de partición con el método de las *c-medias difuso*.

*Los individuos están todos representados por puntos azules, mientras que los centros de los conjuntos por círculos rojos.*

*En este caso, la pertenencia viene dada por las curvas de contorno que rodean a cada uno de los centros, cuanto más cerca esté de uno de los centros, mayor será la pertenencia a dicho conjunto, mientras que si se encuentra a una distancia similar de varios, la pertenencia estará más repartida entre tales conjuntos.*

*Además, en la Figura 4.4 podemos ver la representación gráfica en tres dimensiones de las funciones de pertenencia generadas por dicho método a cada uno de los cuatro conjuntos generados, lo cual nos puede ayudar a ver la forma que dichos conjuntos presentan.*

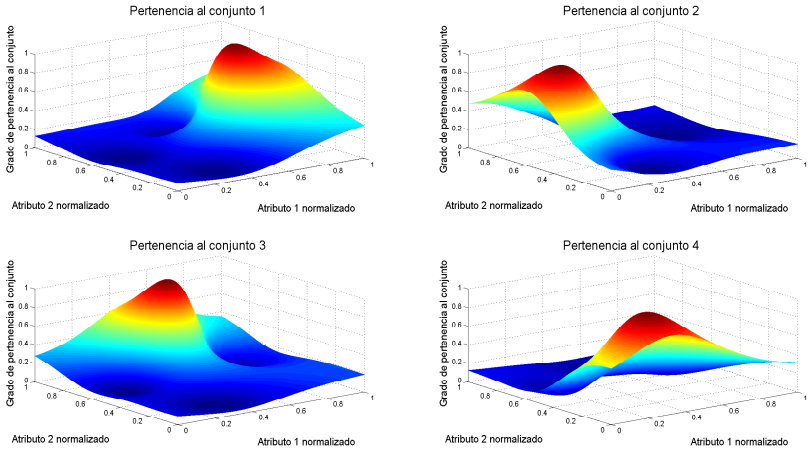


Figura 4.4: Pertenencias de los conjuntos obtenidos con el método de las *c-medias difuso*.

#### 4.1.2.2. Método Gustafson-Kessel

El método de *Gustafson-Kessel* (ver [17]) se diferencia del método de las *c-medias difuso* en la norma utilizada en el funcional a minimizar. En este caso, a cada conjunto generado se le asigna una matriz $A_i$, a partir de la cual se define la norma asociada a dicho conjunto del siguiente modo:

$$D^2_{ikA_i} = ||x_k - v_i||^2_{A_i} = (x_k - v_i)^T A_i (x_k - v_i), \tag{4.3}$$

donde las matrices $A_i$ se utilizan como variables de optimización a la hora de minimizar el funcional. Denotando por $A = (A_1, \ldots, A_c)$, tenemos que la función objetivo a minimizar en este caso es:

$$J(X; U, V, A) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D^2_{ikA_i}. \tag{4.4}$$

Para obtener las matrices $A_i$, se fija previamente el determinante de cada una como argumento de entrada, denotándolo por $det(A_i) = \rho_i$, $\forall i = 1 \ldots, c$. Una vez fijado, se obtienen las expresiones de $A_i$ del siguiente modo:

$$A_i = |\rho_i det(F_i)|^{1/n} F_i^{-1},$$

donde $F_i$ es la matriz de covarianzas difusa del conjunto i-ésimo definida como sigue:

$$F_i = \frac{\sum_{i=1}^{N}(\mu_{ik})^m(x_k - v_i)(x_k - v_i)^T}{\sum_{i=1}^{N}(\mu_{ik})^m}. \tag{4.5}$$

El algoritmo con el que obtener los resultados es *GKclust*, implementado también en el paquete de *Matlab* (ver [1]).

Los argumentos de **entrada** y de **salida** del programa *GKclust* son los mismos que los de *FCMclust*, salvo que en este caso, necesitamos añadir como parámetro de entrada el vector $\rho$ que representa los valores de los determinantes de las matrices $A_i$ anteriormente definidas. La salida, es análoga a la del otro método difuso que hemos analizado.

**Ejemplo 4.4** *Procedemos con un ejemplo del método de Gustafson-Kessel, con los mismos datos. Podemos ver la representación gráfica en la Figura 4.5, cuya estructura es análoga al caso del algoritmo de las c-medias difuso.*

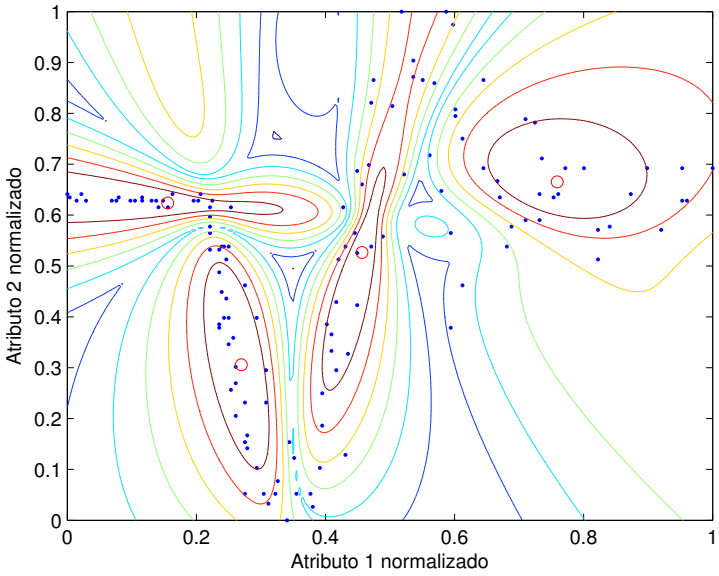


Figura 4.5: Ejemplo de partición con el método de *Gustafson-Kessel*.

*De la misma forma que el ejemplo anterior, los individuos están todos representados por puntos azules, los centros de los conjuntos por círculos rojos y la pertenencia viene dada por las curvas de contorno alrededor de cada uno de los centros.*

*Además, en la Figura 4.6 tenemos también la representación en tres dimensiones de las funciones de pertenencia de los cuatro conjuntos generados por este método.*

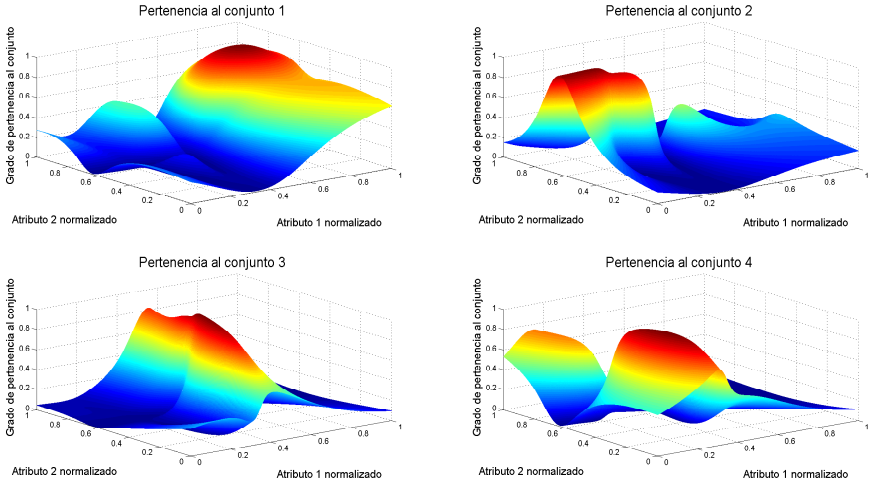


Figura 4.6: Pertenencias de los conjuntos obtenidos con el método de *Gustafson-Kessel*.

Si comparamos las dos gráficas obtenidas con ambos métodos de particiones difusas, vemos como las que corresponden al método de las *c-medias difuso* nos dan conjuntos difusos con formas más "suaves" que el de *Gustafson-Kessel*, cuyas curvas son más irregulares.

El hecho de utilizar distintas matrices a la hora de obtener la norma, hace que la mayor complejidad en el problema de optimización no lineal en el caso de *Gustafson-Kessel* en comparación con el caso del método de las *c-medias difuso* resulte en dicha diferencia entre los conjuntos generados.

Las gráficas obtenidas para este tipo de particiones, guardan cierta similitud con las curvas de nivel de los mapas topográficos, como el de la Figura 4.7. Cada línea que rodea a la cumbre une puntos de misma altura, del mismo modo que obtenemos nosotros uniendo puntos alrededor del centro del conjunto de igual pertenencia al conjunto correspondiente.

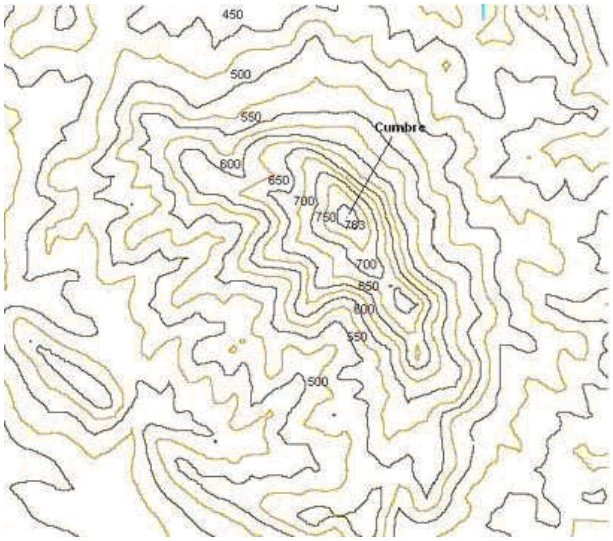


Figura 4.7: Curvas de nivel en un mapa topográfico.

## 4.2. Parámetros de la experimentación

Como ya hemos comentado, la base de datos que hemos utilizado, *CENSUS*, tiene 13 atributos distintos. Hemos seleccionado como posibles **atributos sensibles** el número 10 (*PTOTVAL*) y el número 12 (*TAXINC*), por tratarse de los que tienen un menor número de valores distintos, y así poder obtener resultados más ilustrativos que con otras elecciones.

Hemos considerado como posibles problemas a comparar las combinaciones dos a dos de los atributos de dicha base que no hemos tratado como atributos sensibles (*AFNLWGT, AGI, EMCONTRB, ERNVAL, FEDTAX, FICA, INTVAL, PEARNVAL, POTHVAL, STATETAX, WALVA*).

Tenemos 11 atributos a combinar dos a dos, lo que resulta un total de **55 cuasi-identificadores distintos**.

Los cuasi-identificadores que tenemos en cuenta están formados por, a lo sumo 2 atributos, ya que la utilización de más de dos simultáneamente puede llevar a situaciones más complejas, con la consiguiente imposibilidad de obtener una salida gráfica. Por otro lado, al ser una base de datos no demasiado grande (1080 individuos), tenemos que al aumentar el número de atributos en el cuasi-identificador, el valor de la $k$-anonymity tiende rápidamente a 1 en todos los casos.

Para cada uno de los casos de estudio, hemos obtenido los valores de distintos criterios:

- **$k$-anonymity**: lo obtenemos tanto en el caso de las particiones nítidas (Capítulo 2), como en el de las particiones difusas, donde lo hemos aplicado sobre el vector de etiquetas que determina a qué conjunto tiene cada individuo mayor pertenencia de la partición realizada,
- **$l$-diversity**: obtenido para el caso nítido como hemos visto en el Capítulo 2, y para el caso difuso, como hemos desarrollado en el Capítulo 3,
- **$t$-closeness**: al igual que la $l$-diversity, lo tenemos tanto en el caso nítido como en el difuso en los Capítulos 2 y 3 respectivamente,
- **$Q$-anonymity**: sólo la obtenemos en el caso de las particiones difusas (como hemos definido en el capítulo anterior), ya que carece de sentido en el caso nítido,
- **measure_risk**: número de individuos en riesgo que nos encontramos en cada caso a través de la función *measure_risk* del paquete *sdcMicro* de *R* (ver [20]).

Por tanto, para cada par de atributos como cuasi-identificador, obtendremos la $l$-diversity y la $t$-closeness tanto para el caso en el que tratamos al atributo *PTOTVAL* como sensible, como para el caso en el que el atributo sensible sea *TAXINC*.

Obtenemos la medida de los individuos en riesgo, **measure_risk**, del siguiente modo (ver [18]):

- Para cada individuo de la tabla publicada $i^* \in T^*$, se obtiene la probabilidad $\rho_i$ de que dicho individuo sea relaccionado con un individuo de la tabla original, conocida la tabla publicada tras la partición, $T^*$.

- Obtenemos $r_i$, **riesgo individual de re-identificación**, que representa la misma probabilidad que $\rho_i$, pero bajo la condición en la que el atacante trata de obtener los valores no sólo para el individuo $i^*$, sino para todos los posibles valores de la tabla publicada.

  Es decir, puede obtener más información por haber intentado atacar no sólo a un individuo de la tabla, sino a todos, lo que puede llevar al atacante a obtener información más relevante que si sólo lo hiciera con un individuo.

- El argumento de salida del programa *measure_risk* que tenemos en cuenta nos da el número de individuos de la tabla cuyo **riesgo individual de re-identificación** $r_i$ es mucho mayor que el resto de la tabla, y que por tanto, pueden suponer un problema para la privacidad de éstos.

Hemos dividido la experimentación en dos bloques, uno de ellos en el que realizamos la codificación del cuasi-identificador en conjunto, y otro en el que realizamos la codificación independiente para cada atributo del cuasi-identificador.

### 4.2.1. Codificación independiente de cada atributo

Este caso es el análogo al que hemos visto en los ejemplos ilustrativos de los capítulos anteriores. Para cada cuasi-identificador de par de atributos, tomamos cada uno por separado, y obtenemos las particiones respectivas. Finalmente, cruzamos ambas particiones y obtenemos los datos protegidos con la partición final.

En el caso de las particiones nítidas, tenemos las particiones obtenidas mediante los métodos de las *k-medias* y de *Global Recode*.

Para las particiones difusas, lo hemos realizado con los métodos de las *c-medias difuso* y de *Gustafson-Kessel*. Sin embargo, estos dos últimos métodos, al ser aplicados sobre un solo atributo cada vez, resultan en la misma partición.

A continuación, tenemos el teorema donde hemos demostrado que esto se cumple:

**Teorema 4.5** *Sea $X$ una matriz de datos de dimensión $N \times 1$, es decir, con un solo atributo. Consideramos los métodos de las c-medias difuso y el de Gustafson-Kessel con el mismo número de conjuntos en la partición, tomando para el último como vector de determinantes de las matrices inductoras de la norma $\rho = (1, \ldots, 1)$. Entonces, tenemos que ambos métodos coinciden.*

**Demostración:** Como tenemos que la dimensión de $X$ es $N \times 1$, tenemos que $x_k \in \mathbb{R}$, $\forall k = 1, \ldots, N$. Del mismo modo, tenemos que $v_i \in \mathbb{R}$, $\forall i = 1, \ldots, c$, donde $c$ es el número de conjuntos de las particiones. Veamos que las funciones a minimizar de ambos métodos coinciden.

Comencemos por el método de las *c-medias difuso*, donde tenemos que la función a mínimizar es (4.1), y aplicando lo explicado en el párrafo anterior junto a que $A = I$, tenemos que:

$$||x_k - v_i||_A^2 = (x_k - v_i)^T I (x_k - v_i) = (x_k - v_i)^2,$$

por lo que la función a minimizar será:

$$J(X; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m (x_k - v_i)^2.$$

En el caso del método de *Gustafson-Kessel*, como tenemos como hipótesis que $\rho = (1, \ldots, 1)$, sabemos que $det(A_i) = 1$, $\forall i = 1, \ldots, c$.

Además, sabemos que $A_i = det(F_i)^{1/n} F_i^{-1} = det(F_i) F_i^{-1}$, donde:

$$F_i = \frac{\sum_{i=1}^{N} (\mu_{ik})^m (x_k - v_i)^2}{\sum_{i=1}^{N} (\mu_{ik})^m},$$

sin embargo, resulta que $F_i \in \mathbb{R}$, por lo que $A_i = F_i F_i^{-1} = F_i \frac{1}{F_i} = 1$.

Así, la función a minimizar, dada por la expresión que tenemos en (4.4) será:

$$J(X; U, V, A) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m (x_k - v_i)^2,$$

y por tanto, ambos métodos coinciden en estas condiciones. ■

Por tanto, para las particiones de los atributos por separado, tenemos dos métodos distintos para el caso nítido, mientras que para el caso difuso, tenemos un solo método.

Hemos tomado tres conjuntos para cada una de las particiones de cada atributo por separado, obteniendo en total, tras unir ambos atributos, 9 conjuntos en la partición final.

Hemos decidido tomar un mismo número de conjuntos para todos los pares por ser de gran ayuda al comparar los resultados obtenidos. Además, debemos tener en cuenta que, al ser una base de datos con un número bajo de individuos, se presenta el mismo problema que hemos descrito a la hora de elegir cuasi-identificadores de tamaño a lo sumo 2, donde a mayor número de conjuntos, obtendremos que los valores de la $k$-anonymity serán muy pequeños, tendiendo a 1.

### 4.2.2. Codificación conjunta

En este caso, codificamos de forma conjunta todos los atributos del cuasi-identificador.

Para esta situación, en el caso nítido solo disponemos del método de las *k-medias*, ya que el método *Global Recode* no permite trabajar con más de un atributo simultáneamente.

Para las particiones difusas, a diferencia del caso en el que tomamos los atributos por separado, sí que disponemos de los dos métodos de las *c-medias difuso* y de *Gustafson-Kessel*, ya que ahora existe diferencia en la norma al haber más de un atributo sobre el que realizar la partición.

En total, tenemos un método para particiones nítidas y dos para las particiones difusas.

Del mismo modo que en la codificación de los conjuntos por separado, hemos decidido tomar un número uniforme de conjuntos por partición, siendo 3 conjuntos por cada caso lo elegido, por las mismas razones que hemos dado para el caso anterior.

## 4.3. Análisis de los resultados

Tras realizar la experimentación, hemos obtenido datos para 55 pares distintos.

Para cada uno de estos pares, tenemos dos tipos de resultados como hemos descrito en el apartado anterior, tanto codificando conjuntamente o de forma independiente, con 3 tipos de particiones en cada uno de ellos.

En total, tenemos **6 codificaciones distintas** realizadas para cada uno de los pares.

Los **criterios** que hemos obtenido para cada uno de los casos han sido los siguientes:

- $k$-anonymity,
- measure_risk.
- $Q$-anonymity (sólo en el caso de las particiones difusas),
- Tomando $PTOTVAL$ como atributo sensible:
  - $l$-diversity,

    - $t$-closeness,

- Tomando *TAXINC* como atributo sensible:

    - $l$-diversity,
    - $t$-closeness,

Compararemos los resultados dividiéndolos de nuevo dependiendo de la forma de partición hecha (tomando los atributos en conjunto o tratándolos uno a uno).

### 4.3.1. Análisis de la codificación independiente

Como hemos visto en la sección anterior, tenemos un total de tres métodos (dos nítidos y uno difuso) para el caso en el que tratamos los atributos uno a uno:

- *k-medias*,
- *Global Recode*,
- *Gustafson-Kessel* (recordemos que era análogo en este caso a *c-medias difuso*).

**Ejemplo 4.6** *Veamos un ejemplo de lo obtenido para estas tres particiones en el caso de utilizar como cuasi-identificador los atributos AFNLWGT y AGI.*

*Comenzamos por la partición obtenida por el método de las* **k-medias** *en la Figura 4.8. Todas las gráficas tienen en el eje $X$ los valores del atributo AFNLWGT y en el eje $Y$ los del atributo AGI.*

*En las dos gráficas superiores, podemos ver las particiones que hemos obtenido para cada atributo por separado: a la izquierda en el caso de aplicar la partición obtenida para el atributo AFNLWGT, y a la derecha para el atributo AGI.*

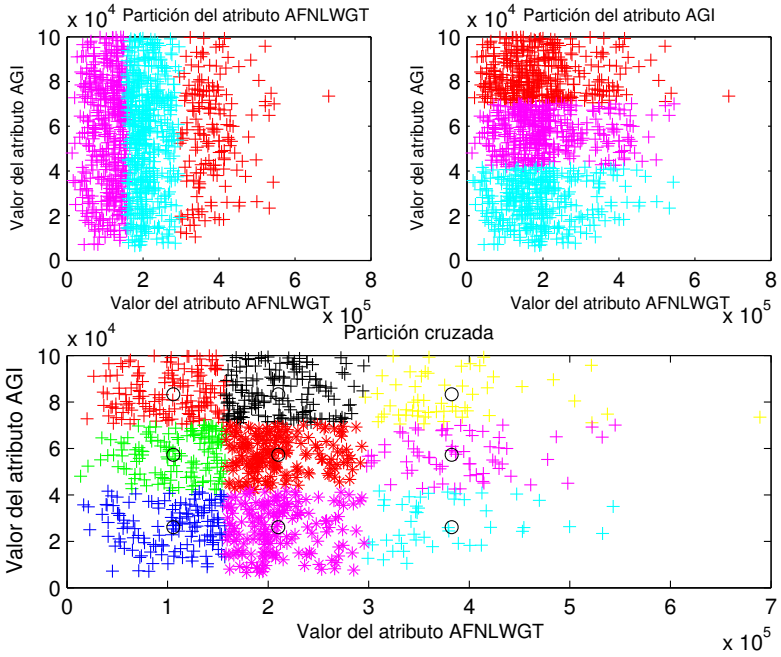


Figura 4.8: Partición de los atributos *AFNLWGT* y *AGI* por el método *k-medias* tratándolos uno a uno.

*En la gráfica inferior, tenemos la partición resultante de cruzar ambas particiones de los dos atributos individualmente, obteniendo la partición final de 9 conjuntos.*

*Por otro lado, obtenemos la correspondiente al método* **Global Recode** *(ver Figura 4.9). Del mismo modo a las gráficas obtenidas con el método de las k-medias, tenemos en las dos gráficas superiores las particiones individuales para cada uno de los atributos, y en la inferior la partición final obtenida al cruzarlas.*

*Si comparamos la gráfica de la partición final de ambos métodos, podemos ver como en el caso de la obtenida con las* **k-medias** *distinguimos claramente los conjuntos generados, mientras que en el caso del método* **Global Recode**, *no podemos decir lo mismo, ya que son conjuntos difíciles de apreciar gráficamente.*

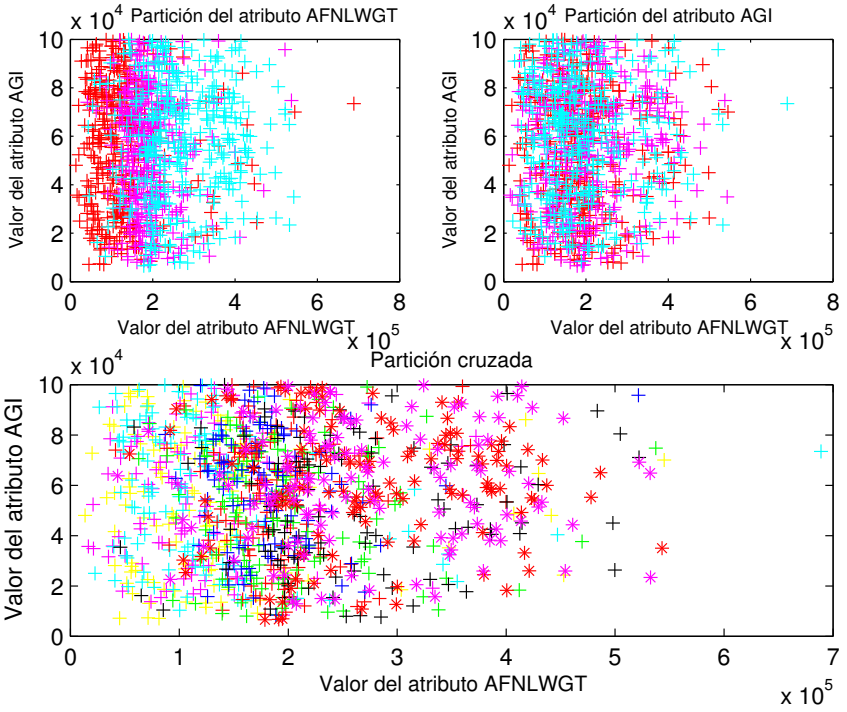


Figura 4.9: Partición de los atributos *AFNLWGT* y *AGI* por el método *Global Recode* tratándolos uno a uno.

*Esto sucede ya que el método* **Global Recode** *agrupa individuos que no necesariamente tienen similitud en los atributos, a diferencia del caso de las* **k-medias**, *que realiza dichas particiones por intervalos.*

*Por último veamos la partición obtenida por el método* **difuso**, *con los valores de los atributos normalizados en los ejes (ver Figura 4.10).*

*Como podemos observar, tenemos 9 centros (círculos rojos), y a su alrededor, quedan representadas las curvas que determinan la pertenencia a cada conjunto.*

*Ahora, procedemos a analizar los resultados obtenidos para estas tres particiones en función de los criterios anteriormente descritos.*

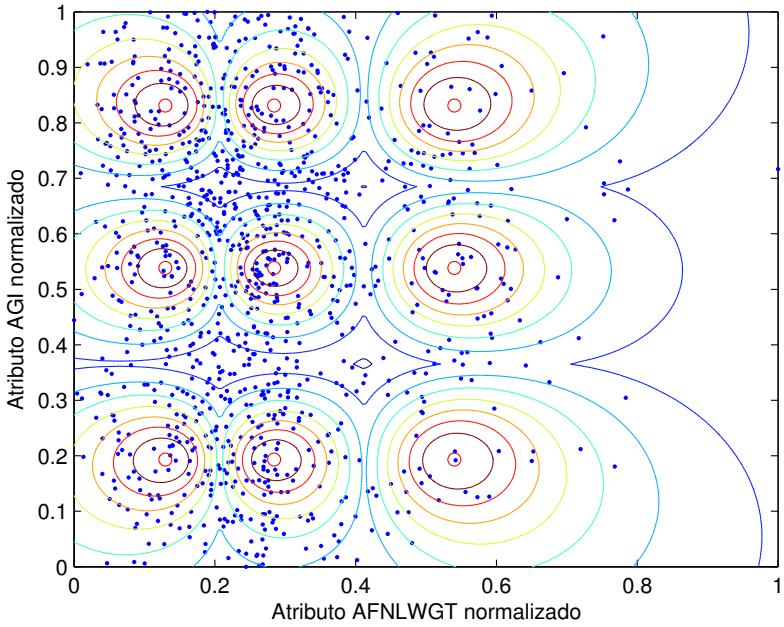


Figura 4.10: Partición de los atributos *AFNLWGT* y *AGI* por el método difuso tratándolos uno a uno.

*En la Tabla 4.1 tenemos los datos con los que podemos comparar los tres métodos (todos los resultados en el Anexo).*

| Criterio | | *k-medias* | *Global Recode* | *Difuso* |
|---|---|---|---|---|
| $k$-anonymity | | 45 | 102 | 44 |
| *measure_risk* | | 165 | 313 | 170 |
| *PTOTVAL* | $l$-diversity | 34 | 73 | 29 |
| | $t$-closeness | 0.0645 | 0.019 | 0.0305 |
| *TAXINC* | $l$-diversity | 34 | 71 | 29 |
| | $t$-closeness | 0.0629 | 0.0255 | 0.0279 |

Tabla 4.1: Tabla comparativa para la codificación independiente de los atributos *AFNLWGT* y *AGI*.

*Cada una de las celdas de la tabla representa el valor obtenido para*

*el criterio dado (por filas) con el método elegido (por columnas), y en el caso de la l-diversity y t-closeness, distinguimos los dos casos en función del atributo sensible tratado.*

*Analicemos los distintos criterios en función de los métodos:*

- ***k*-anonymity y measure_risk***:*
  - ***k*-anonymity***: cuanto mayor sea el valor, mejor será la protección de los datos. En nuestro caso, tenemos que el método* **Global Recode** *tiene un valor de 102 frente a 45 y 44 del método de las* **k-medias** *y el método* **difuso** *respectivamente, es decir,* **Global Recode** *presenta una mejor protección respecto a este criterio.*
  - **measure_risk***: representa el número de individuos que tienen una mayor probabilidad respecto al resto de individuos de ser re-identificados, por lo que cuanto menor sea este valor, mejor protección tendremos. En este caso, el método* **Global Recode** *obtiene el peor valor (313) frente a dos valores similares (165 y 170) de los métodos de las* **k-medias** *y* **difuso** *respectivamente, los cuales son mejores para este criterio.*

  *El número de individuos en riesgo que tiene el método* **Global Recode** *es mucho más alto frente a los otros dos métodos, lo cual parece suficiente como para no poder elegir un método preferente aún con los mejores resultados en la k-anonymity.*

- **Atributo sensible** *PTOTVAL:*
  - ***l*-diversity***: al igual que la k-anonymity, a mayor valor, mejor protección. Obsevamos como el método* **Global Recode** *obtiene un valor 73 mayor al 34 y 29 correspondientes a los otros dos métodos, lo cual lo hace el preferible respecto a esta técnica.*
  - ***t*-closeness***: en este caso, buscamos valores pequeños, ya que a menor valor de t, menor será la diferencia de las distribuciones en el atributo sensible antes y después de la partición. Como podemos observar, vuelve a ser el método* **Global Recode**

*(0.019) el que obtiene el mejor valor frente a los otros dos métodos (0.0629 y 0.0279).*

*Al considerar el atributo PTOTVAL como sensible, es el método* **Global Recode** *el que presenta mejores resultados ante la l-diversity y la t-closeness.*

- **Atributo sensible** *TAXINC:*

  - *l*-**diversity***: de forma similar al caso en el que tratabamos como atributo sensible PTOTVAL, tenemos que es el método* **Global Recode** *el mejor.*
  - *t*-**closeness***: al igual que con la l-diversity, obtenemos resultados parecidos al anterior, volviendo a obtener como mejor método el* **Global Recode***.*

  *De modo similiar al caso anterior, vemos como es el método* **Global Recode** *el preferible sobre este atributo sensible.*

*Como podemos observar, en todos los criterios menos en measure_risk,* **Global Recode** *es superior a los otros dos. Sin embargo, debemos que tener en cuenta que el número de individuos en riesgo en este caso es mucho más grande que el de los otros dos métodos.*

*Cabe destacar que el hecho de tener un mayor valor en measure_risk que los otros dos, es un hecho puntual, ya que en la gran mayoría de casos es éste el que obtiene el mejor valor, como veremos a continuación.*

En la Tabla 4.2 tenemos un resumen de los datos obtenidos para todos los cuasi-identificadores tratados, donde tenemos para cada criterio de los anteriormente descritos (por filas) y cada uno de los tres métodos (por columnas) el número de casos en el que su valor ha sido el mejor respecto al resto de métodos.

En algunos casos el mejor valor no es único, por lo que el número de casos válidos no siempre coincide con la suma de los casos favorables para cada criterio.

| Criterio | | *k-medias* | *Global Recode* | *Difuso* |
|---|---|---|---|---|
| $k$-anonymity | | 8 | 42 | 7 |
| *measure_risk* | | 8 | 37 | 9 |
| *PTOTVAL* | $l$-diversity | 7 | 48 | 1 |
| | $t$-closeness | 2 | 42 | 10 |
| *TAXINC* | $l$-diversity | 7 | 48 | 1 |
| | $t$-closeness | 2 | 42 | 10 |

Tabla 4.2: Tabla resumen de los resultados obtenidos para los 54 cuasi-identificadores respecto a los criterios utilizados.

Además, para este caso, tenemos un total de 54 cuasi-identificadores, ya que en uno de los pares utilizados, los resultados no alcanzan la $k$-anonymity superior a 1 en ninguna partición, y por tanto, lo obviamos.

En la Tabla 4.3, obtendremos además: las **medias** para cada uno de los métodos para los criterios $k$-anonymity y *measure_risk*, que denotamos por $\overline{X}$, y las **medias de las distancias al valor óptimo** en los casos en el que cada método no haya sido el mejor valor, que denotaremos por $D$.

| | *k-medias* | | *Global Recode* | | *Difuso* | |
|---|---|---|---|---|---|---|
| Criterio | $\overline{X}$ | $D$ | $\overline{X}$ | $D$ | $\overline{X}$ | $D$ |
| $k$-anonymity | 15.24 | 46.89 | 52.67 | 8.17 | 15.23 | 45.89 |
| *measure_risk* | 229.53 | 151.72 | 124.33 | 79.11 | 248.49 | 177.7 |

Tabla 4.3: Medias y distancias a los óptimos de los criterios $k$-anonymity y $t$-closeness para los tres métodos.

Para el resto de criterios ($l$-diversity y $t$-closeness) no los obtenemos ya que tenemos varios casos en los que la $k$-anonymity o la $Q$-anonymity tienen valor 1 y no tiene lógica su estudio.

Analicemos ahora los datos de las Tablas 4.2 y 4.3:

- **$k$-anonymity**: tenemos que de los 54 casos válidos, el método **Global Recode** resulta ser el método que obtiene mejores resultados en 42 casos, con los valores más grandes de la $k$-anonymity. Los otros dos

métodos, son favorables en 8 casos para el método de las **k-medias** y en 7 casos para el método **difuso** estudiado.

En la Tabla 4.3 vemos como es la media del método **Global Recode** la mayor, además de ser el método con menor distancia a los valores óptimos en los casos en que no es éste el preferible, con una amplia diferencia respecto a los otros dos métodos.

- **measure_risk**: en 37 de los 54 posibles cuasi-identificadores es de nuevo el método **Global Recode** el que obtiene mejores resultados, presentando un menor número de individuos en riesgo en dichos casos. El método de las **k-medias** es el que obtiene los mejores resultados en 8 casos, mientras que el método **difuso** en 9 casos.

  En la Tabla 4.3 nos encontramos con que la media menor es la del método **Global Recode**, y al igual que en el caso de la $k$-anonymity, el valor de $D$ es el más pequeño respecto a los otros métodos.

- **$l$-diversity**: de los 54 casos válidos, tenemos que en 48 es el método **Global Recode** el que tiene un valor mayor de la $l$-diversity. Para el método de las **k-medias** tenemos 7 casos favorables y 1 para el caso del método **difuso**. Tanto tratando *PTOTVAL* como *TAXINC* como atributo sensible, tenemos los mismos resultados.

- **$t$-closeness**: el método **Global Recode** es el que obtiene valores más pequeños de $t$, con 42 casos favorables. El método **k-medias** obtiene el valor más pequeño de $t$ en 2 ocasiones, mientras que el método **difuso** en 10. Del mismo modo que con la $l$-diversity, tenemos los mismos resultados para ambos atributos sensibles.

Como podemos observar, tanto para el número de individuos en riesgo como para los tres criterios $k$-anonymity, $l$-diversity y $t$-closeness, obtenemos el mayor número de casos favorables para el método **Global Recode**.

Es decir, podemos concluir que en el caso en el que realizamos la codificación independiente, y posteriormente la cruzamos para obtener la partición final, es claramente el método **Global Recode** el mejor respecto a los otros dos métodos analizados respecto a los criterios que hemos utilizado para compararlos.

### 4.3.2. Análisis de la codificación conjunta

Para el caso en el que realizamos la codificación conjunta, comparamos entre tres métodos, donde tenemos que el primero es un método nítido, mientras que los otros dos se tratan de métodos difusos:

- *k-medias*,
- *Gustafson-Kessel*,
- *c-medias difuso*.

**Ejemplo 4.7** *Veamos las particiones conjuntas que obtenemos para los atributos AFNLWGT y AGI como hemos hecho en el apartado anterior.*

*Empecemos por el método nítido,* **k-medias**, *el único método nítido que utilizaremos en este caso (ver Figura 4.11).*

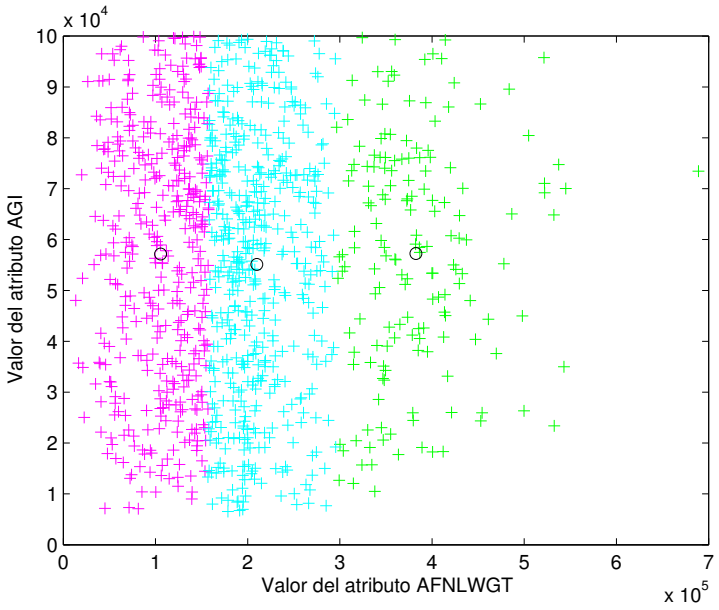


Figura 4.11: Partición de los atributos *AFNLWGT* y *AGI* por el método *k-medias* tratándolos en conjunto.

*Podemos ver como los conjuntos de la partición son fácilmente reconocibles, al ser intervalos lo que la conforman.*

*Ahora, obtendremos la partición generada por el primero de los dos métodos difusos que utilizaremos, el método de* **Gustafson-Kessel**, *que podemos ver en la Figura 4.12, donde queda representado el atributo AFNLWGT normalizado en el eje X y el atributo AGI normalizado en el eje Y.*

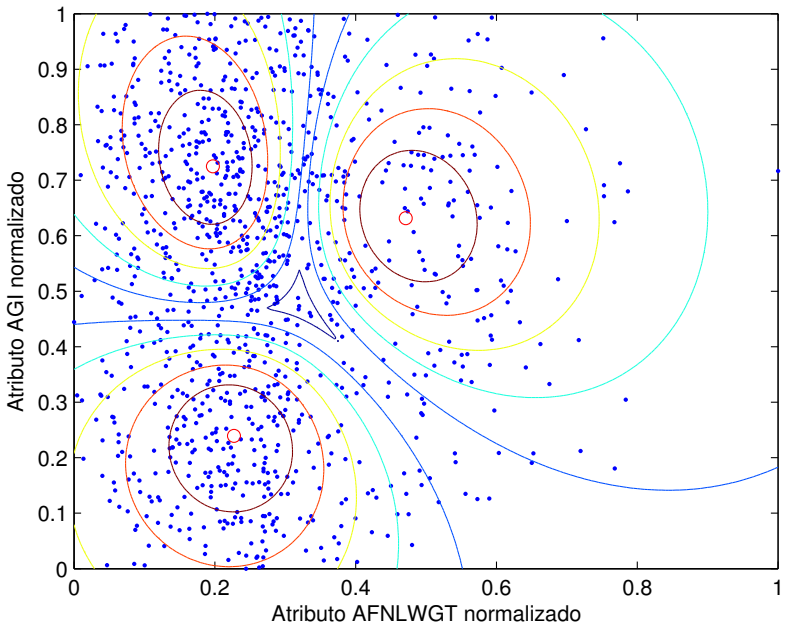


Figura 4.12: Partición de los atributos *AFNLWGT* y *AGI* por el método *Gustafson-Kessel* tratándolos en conjunto.

*Podemos ver de nuevo las curvas de pertenencia relativas a cada uno de los conjuntos creados, con los centro nuevamente denotados por círculos rojos.*

*Por último, obtenemos la partición del segundo método difuso, el de las* **c-medias difuso**. *Queda representado en la Figura 4.13, cuya estructura es la misma que el caso del método anterior, con los atributos normalizados en los ejes, los individuos por puntos azules y las curvas de pertenencia*

*alrededor de los centros de los conjuntos generados.*

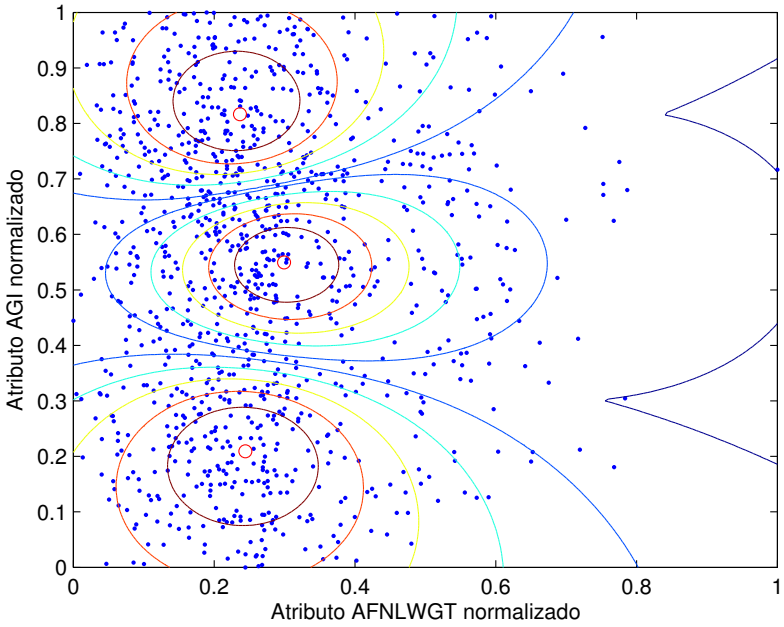


Figura 4.13: Partición de los atributos *AFNLWGT* y *AGI* por el método *c-medias difuso* tratándolos en conjunto.

*Como podemos ver, las formas de los conjuntos de ambos métodos difusos son distintas, debido a la utilización de distintas normas como habíamos estudiado en la sección anterior, al encontrarnos en un caso donde tenemos más de un atributo sobre el que aplicar el algoritmo.*

*Ahora, procedemos a comparar los tres métodos.*

*En este caso, los criterios que utilizaremos para comparar los resultados, serán los mismos que hemos utilizado en el caso en el que realizabamos la codificación independiente (k-anonymity, l-diversity, t-closeness y measure_risk), añadiendo el criterio Q-anonymity para comparar ambos métodos difusos.*

*Los valores para dichos criterios los encontramos en la Tabla 4.4 (todos*

*los resultados en el Anexo).*

| Criterio | | *k-medias* | *Gustafson-Kessel* | *c-medias difuso* |
|---|---|---|---|---|
| $k$-anonymity | | 165 | 221 | 334 |
| *measure_ risk* | | 165 | 162 | 0 |
| $Q$-anonymity | | - | 199 | 319 |
| *PTOTVAL* | $l$-diversity | 99 | 106 | 131 |
| | $t$-closeness | 0.0091 | 0.0035 | 0.0127 |
| *TAXINC* | $l$-diversity | 98 | 102 | 128 |
| | $t$-closeness | 0.0104 | 0.0032 | 0.0101 |

Tabla 4.4: Tabla comparativa para la codificación conjunta de los atributos *AFNLWGT* y *AGI*.

*Tenemos de nuevo por filas los criterios utilizados y por columnas los tres métodos disponibles. La Q-anonymity para el método nítido no tiene valor ya que carece de sentido para un caso de conjuntos no difusos.*

*Analicemos ahora los distintos criterios en función de los métodos:*

- **$k$-anonymity, measure_risk y $Q$-anonymity**:
  - **$k$-anonymity**: *al tener que mayor valor de k implica una mayor protección, podemos ver como es el método de las* **c-medias difuso** *el preferible por ser el que obtiene un mayor valor (334), seguido por las 221 del otro método difuso, el de* **Gustafson-Kessel**. *Finalmente, nuestro método nítido de las* **k-medias** *tiene valor 165 para dicho criterio.*
  - **measure_risk**: *en este caso, tenemos que el método de las* **c-medias difuso** *no obtiene individuos en riesgo para dicho criterio, frente a los 165 y 162 de los métodos de las* **k-medias** *y* **Gustafson-Kessel**.
  - **$Q$-anonymity**: *este criterio nos sirve para comparar los dos métodos difusos, y cuanto mayor sea su valor, mejor será la protección. En nuestro caso, es el método de las* **c-medias difuso** *el que obtiene un valor (319) mayor que el de* **Gustafson-Kessel** *(199).*

*Para estos tres criterios, siempre es el método de las* **c-medias difuso** *el que ofrece los mejores resultados, sin presentar individuos en riesgo, y teniendo mayor protección que los otros métodos respecto a la k-anonymity y la Q-anonymity.*

- **Atributo sensible** *PTOTVAL:*
  - *l*-**diversity***: de nuevo, el mayor valor que encontramos está en el método de las* **c-medias difuso** *(131), seguido por los 106 de* **Gustafson-Kessel** *y los 99 del método de las* **k-medias**.
  - *t*-**closeness***: en este criterio buscamos el valor mínimo, que en este caso lo encontramos en el otro método difuso, el de* **Gustafson-Kessel** *(0.0032), mientras que los otros dos métodos tienen un valor similar, 0.0091 para el de las* **k-medias** *y 0.0127 para el de las* **c-medias difuso**.

  *Vemos como para el atributo PTOTVAL como sensible, los métodos difusos son superiores, siendo preferible respecto a la l-diversity el de las* **c-medias difuso** *y respecto a la t-closeness el de* **Gustafson-Kessel**.
- **Atributo sensible** *TAXINC:*
  - *l*-**diversity***: nos encontramos en un caso muy similar al que trata el atributo PTOTVAL como sensible, llegando a la conclusión que el mejor método respecto a dicho criterio es el de las* **c-medias difuso**.
  - *t*-**closeness***: al igual que pasa con la l-diversity, llegamos a la misma conclusión que con el atributo PTOTVAL como sensible, es decir, que el que obtiene mejor valor es el de* **Gustafson-Kessel**.

  *Volvemos a obtener las mismas conclusiones, resultando más adecuado el método de las* **c-medias difuso** *en función del criterio l-diversity y el de* **Gustafson-Kessel** *si nos fijamos en la t-closeness.*

*Hemos observado como los métodos difusos son superiores al nítido en todos los criterios, siendo mejor el de las* **c-medias difuso** *en todos los*

*criterios salvo en la t-closeness, donde obtenemos mejores resultados para el caso del método de* **Gustafson-Kessel**.

En la Tabla 4.5 tenemos un resumen de los datos obtenidos para todos los cuasi-identificadores, donde para cada criterio de los anteriormente descritos (por filas) y cada uno de los tres métodos (por columnas) tenemos el número de casos en el que haya sido el mejor respecto al resto de métodos.

| Criterio | | *k-medias* | *Gustafson-Kessel* | *c-medias difuso* |
|---|---|---|---|---|
| $k$-anonymity | | 14 | 10 | 32 |
| *measure_risk* | | 27 | 19 | 34 |
| $Q$-anonymity | | - | 8 | 47 |
| *PTOTVAL* | $l$-diversity | 30 | 9 | 18 |
| | $t$-closeness | 0 | 43 | 12 |
| *TAXINC* | $l$-diversity | 29 | 10 | 18 |
| | $t$-closeness | 0 | 43 | 12 |

Tabla 4.5: Tabla resumen de los resultados obtenidos para los 55 cuasi-identificadores respecto a los criterios utilizados.

En la Tabla 4.6, obtendremos además: las **medias** para cada uno de los métodos para los criterios $k$-anonymity y *measure_risk*, que denotamos por $\overline{X}$, y las **medias de las distancias al valor óptimo** en los casos en el que cada método no haya sido el mejor valor, que denotaremos por $D$.

| | | *k-medias* | | *Gustafson-Kessel* | | *c-medias difuso* | |
|---|---|---|---|---|---|---|---|
| Criterio | | $\overline{X}$ | $D$ | $\overline{X}$ | $D$ | $\overline{X}$ | $D$ |
| $k$-anonymity | | 213.62 | 100.27 | 200.78 | 107.05 | 278.42 | 23.78 |
| *measure_risk* | | 164.42 | 189.36 | 236.78 | 257.83 | 109.22 | 107.9 |
| $Q$-anonymity | | - | - | 184.36 | 99.3 | 262.29 | 47.63 |
| *PTOTVAL* | $l$-diversity | 97.73 | 41.88 | 89.55 | 32.54 | 109.56 | 10.7 |
| | $t$-closeness | 0.0867 | 0.0827 | 0.0049 | 0.0042 | 0.0129 | 0.0115 |
| *TAXINC* | $l$-diversity | 93.76 | 37.42 | 86.13 | 30.96 | 104.6 | 10.19 |
| | $t$-closeness | 0.0879 | 0.0841 | 0.0047 | 0.004 | 0.0122 | 0.0107 |

Tabla 4.6: Medias y distancias a los óptimos de los criterios tratados para los tres métodos.

Analicemos los datos de las Tablas 4.5 y 4.6:

- $k$-**anonymity**: tenemos que en 32 de los 55 casos estudiados, es el método de las **c-medias difuso** el que obtiene mejores resultados, lejos de los 14 y 10 de las **k-medias** y **Gustafson-Kessel** respectivamente.

  También tenemos que el método de las **c-medias difuso** tiene una media superior al resto además de una mayor cercanía a los máximos en los casos en los que no es superior al resto.

- **measure_risk**: en cuanto a los individuos en riesgo, tenemos que el método de las **k-medias** obtiene 27 casos favorables de los 55 posibles, el método de **Gustafson-Kessel** 19 y el de las **c-medias difuso** 34.

  Si bien es el último el que tiene mejores resultados, para reforzar dicha hipótesis, observamos los resultados de la Tabla 4.6, donde vemos como la media más pequeña la encontramos en el método de las **c-medias difuso**, con cierta diferencia respecto a los otros dos métodos. Además, la diferencia en los casos desfavorables sigue siendo más pequeña que la del resto de métodos.

- $Q$-**anonymity**: en este caso, comparamos ambos métodos difusos, siendo concluyente lo obtenido en la Tabla 4.5, ya que en 47 casos el método de las **c-medias difuso** es favorable respecto a los 8 casos del de **Gustafson-Kessel**.

  De nuevo vemos como en las dos medidas que hemos tratado en la Tabla 4.6 ($\overline{X}$ y $D$) es el método de las **c-medias difuso** el que nos da los mejores resultados.

- $l$-**diversity**: en el caso de la $l$-diversity tratando el atributo $PTOTVAL$ como sensible podemos ver como el método nítido de las **k-medias** obtiene mejores resultados en 30 de los 55 casos posibles, frente a los 9 y 18 de los dos métodos difusos **Gustafson-Kessel** y **c-medias difuso**.

  Nos encontramos una diferencia similar en el caso en el que tratamos el atributo $TAXINC$ como sensible, donde tenemos los 29 casos favorables del caso nítido **k-medias** frente a los 10 y 18 de los dos métodos difusos **Gustafson-Kessel** y **c-medias difuso**.

Sin embargo, si obviamos el método de **Gustafson-Kessel** obtenemos los resultados de la Tabla 4.7, y vemos como los resultados son menos dispares.

| Atr. Sensible | *k-medias* | *c-medias difuso* |
|---|---|---|
| *PTOTVAL* | 33 | 24 |
| *TAXINC* | 34 | 25 |

Tabla 4.7: Comparativa de número de casos favorables entre *k-medias* y *c-medias difuso* para la *l*-diversity.

Si además nos fijamos en los resultados de la Tabla 4.6, vemos como obtenemos que tanto la media como las diferencias con los máximos son mejores en el caso del método de las **c-medias difuso**, tanto tratando el atributo *PTOTVAL* sensible como con el atributo *TAXINC*.

Es decir, si bien ante el número de casos favorables resulta preferible el método nítido de las **k-medias**, teniendo en cuenta las medias de los valores obtenidos en los distintos casos estudiados, vemos como, en media, son mejores los resultados obtenidos en el método de las **c-medias difuso**.

Esto nos refleja que el método nítido toma valores más extremos que el método difuso, lo cual puede no ser aconsejable ya que si bien puede presentar muy buena protección sobre algunos casos, puede resultar en una deficiencia en la protección de otros.

- *t*-**closeness**: en este caso, podemos ver como el método nítido no logra obtener en ningún caso ser el mejor respecto a la *t*-closeness, mientras que es el método difuso **Gustafson-Kessel** el que obtiene 43 casos en los que es la mejor opción frente a los 12 del método de las **c-medias difuso**, tanto tomando el atributo *PTOTVAL* como sensible como tomando el atributo *TAXINC*.

  Con las medidas de la Tabla 4.6 obtenemos las mismas conclusiones, donde vemos de nuevo que es el método de **Gustafson-Kessel** el que presenta mejores resultados.

En resumen, tenemos que el método de las **c-medias difuso** es superior al resto respecto a los criterios de la $k$-anonymity, *measure_risk* y $Q$-anonymity. Además, respecto a la $l$-diversity parece más aconsejable frente al método nítido de las **k-medias** por lo visto anteriormente. Respecto a la $t$-closeness tenemos que el método preferible es el de **Gustafson-Kessel**.

Por tanto, el método más aconsejable es el de las **c-medias difuso**, ya que la mejora en la $t$-closeness que presenta el otro método difuso no compensa la pérdida de privacidad reflejada por el resto de criterios estudiados.

### 4.3.3. Comentarios finales

Hemos utilizado una base de datos con 1080 individuos y 13 atributos, y tomado todos los posibles cuasi-identificadores formados por pares de atributos. Con esto, hemos dividido en dos grupos la experimentación:

- Tratando los dos atributos uno a uno y cruzando las dos particiones obtenidas para crear la partición final. En este caso, hemos utilizado tres métodos distintos: **k-medias**, **Global Recode** y **c-medias difuso**, donde los dos primeros son nítidos y el tercero difuso.

  Tras realizar la experimentación con los criterios que hemos definido, llegamos a la conclusión que el método **Global Recode** es el que mejor protege los datos frente a los otros dos métodos posibles.

- Tratando ambos atributos a la vez para obtener la partición. En este caso tenemos también tres métodos distintos: **k-medias**, **Gustafson-Kessel** y **c-medias difuso**, siendo el primero nítido y los otros dos difusos.

  Utilizando una experimentación análoga a la del caso anterior, llegamos a la conclusión que es el método de las **c-medias difuso** el que nos ofrece particiones con un mayor poder de protección, mejorando así los resultados obtenidos mediante el conocido método nítido de las **k-medias**.

Con los resultados obtenidos mediante la experimentación realizada, vemos como las particiones generadas, tomando los atributos en conjunto

y utilizando el método de las **c-medias difuso**, presentan una seria alternativa a los métodos ya existentes, aprovechando las propiedades que los **conjuntos difusos** nos ofrecen.

# Capítulo 5

# Conclusiones y puntos abiertos

En este trabajo se han desarrollado técnicas alternativas para proteger la privacidad que puede ser vulnerada al liberar cierta información en forma de microdatos.

Para ello, y partiendo de las usuales y ya existentes técnicas de $k$-anonymity, $l$-diversity y $t$-closeness, se ha recurrido a los conjuntos difusos, cuyas propiedades los hacen idóneos para el propósito buscado en el trabajo. De este modo, tras la descripción de los conceptos básicos de privacidad, los problemas existentes, y el desarrollo de las tres técnicas anteriormente nombradas, se han proporcionado las definiciones, conceptos básicos, y aspectos concretos relativos a la cardinalidad de los conjuntos difusos, materia fundamental en este trabajo.

En primer lugar se ha adaptado la $k$-anonymity cuando la herramienta de generalización se basa en conjuntos difusos. En concreto, se ha desarrollado una técnica similar al caso de conjuntos clásicos, pero con una potencia de protección mayor. Sin embargo, esta técnica no está exenta de ciertos problemas ya existentes para el caso de la $k$-anonymity clásica, como por ejemplo, su indefensión frente a ataques de homogeneidad, poniendo en riesgo la privacidad de los individuos participantes en los datos.

Para paliar dicho problema, se ha adaptado la $l$-diversity, tratándola con particiones difusas, definiendo una técnica similar al caso convencio-

nal, la cual plantea algunos problemas parecidos, relativos a los ataques relacionados a la similitud de los valores sensibles.

Además, hemos adaptado al caso borroso la última técnica de las nombradas anteriormente, la $t$-closeness. Para ello, hemos necesitado buscar una nueva distribución de probabilidad en función de cada individuo que conforman los datos, a partir de las distribuciones del atributo sensible de cada conjunto de la partición difusa, pudiendo así, adaptarla a la $t$-closeness para conjuntos clásicos con la EMD (Earth Mover's Distance).

Los estudios teóricos se han cumplimentado con el desarrollo de distintas experimentaciones sobre una base de datos real, con las que hemos obtenido conclusiones sobre la mejora de la privacidad con la utilización de particiones difusas respecto a distintos criterios estudiados, como han sido $k$-anonymity, $l$-diversity o $t$-closeness. Con dicha experimentación, hemos llegado a la conclusión que las particiones difusas suponen una alternativa eficiente a las técnicas nítidas conocidas, pudiendo dar lugar a una mayor protección de la privacidad, viendo como respecto a los criterios estudiados, presentan una mejora sustancial.

En conclusión, se han desarrollado técnicas de protección de datos, aprovechando las ventajas que proporcionan en este ámbito los conjuntos difusos, ya que la información publicada en microdatos que se obtiene con estas técnicas es más informativa a un nivel similar de generalización. Por otro lado, dado que el número de elementos de cada clase en que dividimos los microdatos es también difuso, la protección frente a ataques externos es mayor.

Este trabajo deja varias lineas abiertas. Entre otras, podemos destacar:

- búsqueda de la configuración óptima de todos los parámetros del sistema (tipo y número de conjuntos difusos utilizados, influencia de los operadores de agregación, etc.),
- estudio del comportamiento de las técnicas desarrolladas sobre estructuras más complejas.

# Anexo

En este Anexo exponemos todos los resultados obtenidos para los 55 cuasi-identificadores tratados en el apartado de experimentación.

La estructura de los datos se divide del siguiente modo:

- Codificación independiente:
  - Método de las k-medias.
  - Método Global Recode.
  - Método difuso.
- Codificación conjunta:
  - Método de las k-medias.
  - Método de Gustafson-Kessel.
  - Método de las c-medias difuso.

En cada uno de los casos, presentamos los datos en dos tablas, agrupando por un lado los valores de $k$-anonymity, *measure_ risk* y $Q$-anonymity (en caso de ser una partición difusa), y por otro lado, la $l$-diversity y $t$-closeness para ambos casos de atributos sensibles.

Los casos en los que no hay valores para los criterios $l$-diversity y $t$-closeness se deben a que o la $k$-anonymity o bien la $Q$-anonymity obtienen valor 1.

## 5.1. Codificación independiente

### 5.1.1. k-medias: $k$-anonymity y measure_risk

| Pares | | $k$-anonymity | *measure_risk* |
|---|---|---|---|
| 1 | 2 | 45 | 165 |
| 1 | 3 | 40 | 165 |
| 1 | 4 | 30 | 365 |
| 1 | 5 | 28 | 310 |
| 1 | 6 | 24 | 227 |
| 1 | 7 | 47 | 164 |
| 1 | 8 | 5 | 320 |
| 1 | 9 | 7 | 259 |
| 1 | 11 | 41 | 165 |
| 1 | 13 | 41 | 165 |
| 2 | 3 | 33 | 76 |
| 2 | 4 | 2 | 137 |
| 2 | 5 | 1 | 192 |
| 2 | 6 | 11 | 53 |
| 2 | 7 | 17 | 112 |
| 2 | 8 | 1 | 182 |
| 2 | 9 | 4 | 103 |
| 2 | 11 | 2 | 174 |
| 2 | 13 | 46 | 290 |
| 3 | 4 | 37 | 389 |
| 3 | 5 | 19 | 244 |
| 3 | 6 | 27 | 377 |
| 3 | 7 | 43 | 89 |
| 3 | 8 | 7 | 182 |
| 3 | 9 | 3 | 103 |
| 3 | 11 | 36 | 378 |
| 3 | 13 | 13 | 60 |
| 4 | 5 | 2 | 323 |
| 4 | 6 | 3 | 270 |

Sigue en la página siguiente.

| Pares | | $k$-anonymity | *measure_risk* |
|---|---|---|---|
| 4 | 7 | 1 | 218 |
| 4 | 8 | 1 | 358 |
| 4 | 9 | 6 | 285 |
| 4 | 11 | 16 | 127 |
| 4 | 13 | 8 | 91 |
| 5 | 6 | 14 | 279 |
| 5 | 7 | 19 | 499 |
| 5 | 8 | 12 | 337 |
| 5 | 9 | 1 | 277 |
| 5 | 11 | 3 | 211 |
| 5 | 13 | 11 | 356 |
| 6 | 7 | 8 | 338 |
| 6 | 8 | 4 | 318 |
| 6 | 9 | 1 | 250 |
| 6 | 11 | 15 | 151 |
| 6 | 13 | 20 | 104 |
| 7 | 8 | 12 | 182 |
| 7 | 9 | 4 | 103 |
| 7 | 11 | 5 | 278 |
| 7 | 13 | 43 | 155 |
| 8 | 9 | 6 | 232 |
| 8 | 11 | 5 | 412 |
| 8 | 13 | 1 | 370 |
| 9 | 11 | 3 | 103 |
| 9 | 13 | 3 | 334 |
| 11 | 13 | 1 | 217 |

Tabla 5.1: Resultados obtenidos para la codificación independiente en el caso del método de las *k-medias* para $k$-anonymity y *measure_risk*.

### 5.1.2. k-medias: $l$-diversity y $t$-closeness

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 1 | 2 | 34 | 0.0645 | 34 | 0.0629 |
| 1 | 3 | 35 | 0.0133 | 35 | 0.0127 |
| 1 | 4 | 23 | 0.0865 | 23 | 0.0945 |
| 1 | 5 | 21 | 0.1316 | 21 | 0.1361 |
| 1 | 6 | 18 | 0.0657 | 18 | 0.0679 |
| 1 | 7 | 36 | 0.0657 | 36 | 0.0662 |
| 1 | 8 | 5 | 0.0164 | 5 | 0.018 |
| 1 | 9 | 7 | 0.0106 | 6 | 0.0121 |
| 1 | 11 | 30 | 0.1721 | 30 | 0.1686 |
| 1 | 13 | 26 | 0.1507 | 26 | 0.1655 |
| 2 | 3 | 25 | 0.0347 | 25 | 0.0327 |
| 2 | 4 | 2 | 0.1067 | 2 | 0.1041 |
| 2 | 5 | - | - | - | - |
| 2 | 6 | 9 | 0.0769 | 9 | 0.0736 |
| 2 | 7 | 13 | 0.0731 | 13 | 0.0726 |
| 2 | 8 | - | - | - | - |
| 2 | 9 | 4 | 0.0759 | 4 | 0.0751 |
| 2 | 11 | 2 | 0.1627 | 2 | 0.1619 |
| 2 | 13 | 26 | 0.1571 | 26 | 0.1546 |
| 3 | 4 | 31 | 0.0498 | 31 | 0.0514 |
| 3 | 5 | 16 | 0.092 | 16 | 0.0896 |
| 3 | 6 | 21 | 0.053 | 21 | 0.0565 |
| 3 | 7 | 33 | 0.033 | 32 | 0.0352 |
| 3 | 8 | 7 | 0.0401 | 7 | 0.0426 |
| 3 | 9 | 3 | 0.0158 | 3 | 0.0171 |
| 3 | 11 | 22 | 0.1697 | 22 | 0.1634 |
| 3 | 13 | 10 | 0.1313 | 10 | 0.142 |
| 4 | 5 | 2 | 0.1302 | 2 | 0.1285 |
| 4 | 6 | 3 | 0.0745 | 3 | 0.0691 |
| 4 | 7 | - | - | - | - |
| 4 | 8 | - | - | - | - |

Sigue en la página siguiente.

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 4 | 9 | 6 | 0.1043 | 6 | 0.1086 |
| 4 | 11 | 14 | 0.1696 | 14 | 0.1649 |
| 4 | 13 | 7 | 0.1555 | 7 | 0.1528 |
| 5 | 6 | 11 | 0.1262 | 11 | 0.1292 |
| 5 | 7 | 14 | 0.1489 | 14 | 0.1381 |
| 5 | 8 | 11 | 0.0511 | 11 | 0.0493 |
| 5 | 9 | - | - | - | - |
| 5 | 11 | 3 | 0.1277 | 3 | 0.1183 |
| 5 | 13 | 10 | 0.0664 | 9 | 0.0626 |
| 6 | 7 | 8 | 0.0898 | 8 | 0.0902 |
| 6 | 8 | 3 | 0.055 | 3 | 0.0536 |
| 6 | 9 | - | - | - | - |
| 6 | 11 | 12 | 0.165 | 12 | 0.166 |
| 6 | 13 | 17 | 0.1536 | 16 | 0.162 |
| 7 | 8 | 11 | 0.0868 | 11 | 0.0859 |
| 7 | 9 | 4 | 0.0796 | 4 | 0.0805 |
| 7 | 11 | 5 | 0.1689 | 5 | 0.1131 |
| 7 | 13 | 22 | 0.0803 | 20 | 0.1078 |
| 8 | 9 | 6 | 0.0119 | 6 | 0.0106 |
| 8 | 11 | 4 | 0.1711 | 4 | 0.1541 |
| 8 | 13 | - | - | - | - |
| 9 | 11 | 3 | 0.1763 | 3 | 0.1657 |
| 9 | 13 | 3 | 0.156 | 3 | 0.1704 |
| 11 | 13 | - | - | - | - |

Tabla 5.2: Resultados obtenidos para la codificación independiente en el caso del método de las *k-medias* para $l$-diversity y $t$-closeness.

### 5.1.3. Global Recode: $k$-anonymity y measure_risk

| Pares | | $k$-anonymity | *measure_risk* |
|---|---|---|---|
| 1 | 2 | 102 | 313 |
| 1 | 3 | 110 | 0 |
| 1 | 4 | 100 | 0 |
| 1 | 5 | 102 | 0 |
| 1 | 6 | 92 | 92 |
| 1 | 7 | 101 | 310 |
| 1 | 8 | 101 | 0 |
| 1 | 9 | 101 | 204 |
| 1 | 11 | 105 | 0 |
| 1 | 13 | 106 | 0 |
| 2 | 3 | 47 | 94 |
| 2 | 4 | 31 | 162 |
| 2 | 5 | 1 | 90 |
| 2 | 6 | 13 | 297 |
| 2 | 7 | 20 | 108 |
| 2 | 8 | 95 | 0 |
| 2 | 9 | 66 | 132 |
| 2 | 11 | 2 | 42 |
| 2 | 13 | 41 | 115 |
| 3 | 4 | 57 | 122 |
| 3 | 5 | 49 | 101 |
| 3 | 6 | 66 | 135 |
| 3 | 7 | 53 | 110 |
| 3 | 8 | 107 | 0 |
| 3 | 9 | 102 | 0 |
| 3 | 11 | 32 | 88 |
| 3 | 13 | 38 | 87 |
| 4 | 5 | 3 | 240 |
| 4 | 6 | 12 | 274 |
| 4 | 7 | 12 | 74 |
| 4 | 8 | 82 | 82 |
| 4 | 9 | 56 | 120 |

Sigue en la página siguiente.

| Pares | | $k$-anonymity | *measure_risk* |
|---|---|---|---|
| 4 | 11 | 6 | 197 |
| 4 | 13 | 3 | 113 |
| 5 | 6 | 12 | 106 |
| 5 | 7 | 4 | 165 |
| 5 | 8 | 80 | 80 |
| 5 | 9 | 65 | 137 |
| 5 | 11 | 19 | 164 |
| 5 | 13 | 16 | 486 |
| 6 | 7 | 12 | 282 |
| 6 | 8 | 89 | 188 |
| 6 | 9 | 68 | 138 |
| 6 | 11 | 16 | 61 |
| 6 | 13 | 18 | 135 |
| 7 | 8 | 91 | 0 |
| 7 | 9 | 62 | 129 |
| 7 | 11 | 6 | 209 |
| 7 | 13 | 2 | 114 |
| 8 | 9 | 4 | 374 |
| 8 | 11 | 96 | 0 |
| 8 | 13 | 93 | 0 |
| 9 | 11 | 65 | 65 |
| 9 | 13 | 63 | 152 |
| 11 | 13 | 2 | 151 |

Tabla 5.3: Resultados obtenidos para la codificación independiente en el caso del método *Global Recode* para $k$-anonymity y *measure_risk*.

### 5.1.4. Global Recode: $l$-diversity y $t$-closeness

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 1 | 2 | 73 | 0.019 | 71 | 0.0255 |
| 1 | 3 | 67 | 0.0142 | 68 | 0.015 |
| 1 | 4 | 68 | 0.0198 | 66 | 0.0226 |
| 1 | 5 | 71 | 0.013 | 69 | 0.0161 |
| 1 | 6 | 67 | 0.016 | 66 | 0.0129 |
| 1 | 7 | 72 | 0.0118 | 70 | 0.0165 |
| 1 | 8 | 67 | 0.0183 | 66 | 0.0172 |
| 1 | 9 | 71 | 0.0152 | 72 | 0.0145 |
| 1 | 11 | 67 | 0.0292 | 68 | 0.0305 |
| 1 | 13 | 63 | 0.03 | 62 | 0.0303 |
| 2 | 3 | 39 | 0.0189 | 39 | 0.0203 |
| 2 | 4 | 22 | 0.021 | 22 | 0.0226 |
| 2 | 5 | - | - | - | - |
| 2 | 6 | 12 | 0.0269 | 12 | 0.0223 |
| 2 | 7 | 14 | 0.0136 | 14 | 0.0171 |
| 2 | 8 | 66 | 0.0114 | 64 | 0.0091 |
| 2 | 9 | 49 | 0.0142 | 49 | 0.0127 |
| 2 | 11 | 2 | 0.0096 | 2 | 0.0093 |
| 2 | 13 | 34 | 0.0186 | 34 | 0.0208 |
| 3 | 4 | 50 | 0.0212 | 48 | 0.0213 |
| 3 | 5 | 38 | 0.0249 | 37 | 0.0276 |
| 3 | 6 | 56 | 0.0178 | 53 | 0.0163 |
| 3 | 7 | 45 | 0.0239 | 43 | 0.0231 |
| 3 | 8 | 70 | 0.0114 | 71 | 0.0094 |
| 3 | 9 | 68 | 0.0191 | 68 | 0.0188 |
| 3 | 11 | 27 | 0.0512 | 26 | 0.0556 |
| 3 | 13 | 30 | 0.0614 | 30 | 0.0593 |
| 4 | 5 | 3 | 0.0208 | 3 | 0.0222 |
| 4 | 6 | 11 | 0.0191 | 11 | 0.0169 |
| 4 | 7 | 10 | 0.0117 | 10 | 0.0136 |
| 4 | 8 | 62 | 0.0173 | 61 | 0.0179 |

Sigue en la página siguiente.

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 4 | 9 | 43 | 0.0153 | 44 | 0.0162 |
| 4 | 11 | 6 | 0.0162 | 6 | 0.02 |
| 4 | 13 | 3 | 0.0162 | 3 | 0.0159 |
| 5 | 6 | 12 | 0.0118 | 12 | 0.0111 |
| 5 | 7 | 4 | 0.0213 | 4 | 0.0237 |
| 5 | 8 | 56 | 0.0201 | 55 | 0.0203 |
| 5 | 9 | 50 | 0.0132 | 51 | 0.0142 |
| 5 | 11 | 17 | 0.0494 | 17 | 0.0448 |
| 5 | 13 | 14 | 0.0222 | 14 | 0.0231 |
| 6 | 7 | 11 | 0.0152 | 11 | 0.0113 |
| 6 | 8 | 61 | 0.0172 | 59 | 0.0151 |
| 6 | 9 | 48 | 0.0185 | 49 | 0.0204 |
| 6 | 11 | 15 | 0.0129 | 15 | 0.0138 |
| 6 | 13 | 15 | 0.0157 | 15 | 0.0173 |
| 7 | 8 | 68 | 0.0092 | 66 | 0.0131 |
| 7 | 9 | 45 | 0.0156 | 46 | 0.0186 |
| 7 | 11 | 6 | 0.0141 | 6 | 0.0157 |
| 7 | 13 | 2 | 0.0094 | 2 | 0.0097 |
| 8 | 9 | 4 | 0.0127 | 4 | 0.0128 |
| 8 | 11 | 66 | 0.0199 | 64 | 0.016 |
| 8 | 13 | 63 | 0.0151 | 61 | 0.0171 |
| 9 | 11 | 54 | 0.0153 | 55 | 0.0147 |
| 9 | 13 | 51 | 0.0121 | 51 | 0.0163 |
| 11 | 13 | 2 | 0.0515 | 2 | 0.0523 |

Tabla 5.4: Resultados obtenidos para la codificación independiente en el caso del método *Global Recode* para $l$-diversity y $t$-closeness.

### 5.1.5. Método difuso: $Q$-anonymity. $k$-anonymity y measure_risk

| Pares | | $Q$-anonymity | $k$-anonymity | *measure_ risk* |
|---|---|---|---|---|
| 1 | 2 | 37 | 44 | 170 |
| 1 | 3 | 40 | 42 | 170 |
| 1 | 4 | 31 | 35 | 170 |
| 1 | 5 | 32 | 35 | 251 |
| 1 | 6 | 25 | 28 | 245 |
| 1 | 7 | 42 | 45 | 170 |
| 1 | 8 | 6 | 8 | 352 |
| 1 | 9 | 1 | 3 | 259 |
| 1 | 11 | 38 | 42 | 170 |
| 1 | 13 | 37 | 41 | 170 |
| 2 | 3 | 30 | 29 | 71 |
| 2 | 4 | 1 | 8 | 250 |
| 2 | 5 | 1 | 1 | 352 |
| 2 | 6 | 1 | 17 | 191 |
| 2 | 7 | 1 | 15 | 110 |
| 2 | 8 | 1 | 3 | 212 |
| 2 | 9 | 1 | 3 | 99 |
| 2 | 11 | 2 | 2 | 281 |
| 2 | 13 | 1 | 46 | 278 |
| 3 | 4 | 35 | 38 | 109 |
| 3 | 5 | 15 | 18 | 242 |
| 3 | 6 | 29 | 33 | 33 |
| 3 | 7 | 35 | 38 | 81 |
| 3 | 8 | 7 | 8 | 212 |
| 3 | 9 | 2 | 2 | 99 |
| 3 | 11 | 26 | 35 | 375 |
| 3 | 13 | 8 | 11 | 58 |
| 4 | 5 | 1 | 1 | 466 |
| 4 | 6 | 1 | 2 | 284 |
| 4 | 7 | 1 | 6 | 425 |
| 4 | 8 | 1 | 3 | 394 |

Sigue en la página siguiente.

| Pares | | $Q$-anonymity | $k$-anonymity | *measure_ risk* |
|---|---|---|---|---|
| 4 | 9 | 1 | 2 | 301 |
| 4 | 11 | 15 | 17 | 137 |
| 4 | 13 | 1 | 4 | 27 |
| 5 | 6 | 9 | 10 | 510 |
| 5 | 7 | 1 | 21 | 492 |
| 5 | 8 | 1 | 1 | 347 |
| 5 | 9 | 1 | 2 | 267 |
| 5 | 11 | 3 | 2 | 194 |
| 5 | 13 | 1 | 12 | 449 |
| 6 | 7 | 1 | 15 | 500 |
| 6 | 8 | 3 | 5 | 366 |
| 6 | 9 | 1 | 3 | 270 |
| 6 | 11 | 16 | 18 | 157 |
| 6 | 13 | 14 | 19 | 44 |
| 7 | 8 | 1 | 1 | 212 |
| 7 | 9 | 1 | 2 | 99 |
| 7 | 11 | 2 | 3 | 268 |
| 7 | 13 | 1 | 42 | 201 |
| 8 | 9 | 1 | 1 | 249 |
| 8 | 11 | 7 | 7 | 441 |
| 8 | 13 | 4 | 4 | 212 |
| 9 | 11 | 2 | 2 | 394 |
| 9 | 13 | 2 | 2 | 320 |
| 11 | 13 | 1 | 1 | 461 |

Tabla 5.5: Resultados obtenidos para la codificación independiente en el caso del método *difuso* para $Q$-anonymity. $k$-anonymity y *measure_ risk*.

### 5.1.6. Método difuso: $l$-diversity y $t$-closeness

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 1 | 2 | 29 | 0.0305 | 29 | 0.0279 |
| 1 | 3 | 33 | 0.0078 | 33 | 0.0077 |
| 1 | 4 | 24 | 0.0178 | 24 | 0.0163 |
| 1 | 5 | 24 | 0.0329 | 24 | 0.03 |
| 1 | 6 | 17 | 0.0056 | 17 | 0.0062 |
| 1 | 7 | 32 | 0.026 | 32 | 0.0227 |
| 1 | 8 | 4 | 0.0025 | 4 | 0.003 |
| 1 | 9 | - | - | - | - |
| 1 | 11 | 28 | 0.0924 | 28 | 0.0887 |
| 1 | 13 | 23 | 0.0462 | 23 | 0.0492 |
| 2 | 3 | 20 | 0.0221 | 20 | 0.0211 |
| 2 | 4 | - | - | - | - |
| 2 | 5 | - | - | - | - |
| 2 | 6 | - | - | - | - |
| 2 | 7 | - | - | - | - |
| 2 | 8 | - | - | - | - |
| 2 | 9 | - | - | - | - |
| 2 | 11 | 2 | 0.0805 | 2 | 0.0763 |
| 2 | 13 | - | - | - | - |
| 3 | 4 | 29 | 0.0278 | 29 | 0.027 |
| 3 | 5 | 13 | 0.0395 | 13 | 0.0372 |
| 3 | 6 | 20 | 0.0135 | 20 | 0.0135 |
| 3 | 7 | 26 | 0.0227 | 26 | 0.0215 |
| 3 | 8 | 5 | 0.0086 | 5 | 0.009 |
| 3 | 9 | 1 | 0.0119 | 1 | 0.0117 |
| 3 | 11 | 18 | 0.0864 | 18 | 0.0792 |
| 3 | 13 | 7 | 0.0521 | 7 | 0.0519 |
| 4 | 5 | - | - | - | - |
| 4 | 6 | - | - | - | - |
| 4 | 7 | - | - | - | - |
| 4 | 8 | - | - | - | - |

Sigue en la página siguiente.

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 4 | 9 | - | - | - | - |
| 4 | 11 | 10 | 0.0876 | 10 | 0.0856 |
| 4 | 13 | - | - | - | - |
| 5 | 6 | 8 | 0.0384 | 8 | 0.0338 |
| 5 | 7 | - | - | - | - |
| 5 | 8 | - | - | - | - |
| 5 | 9 | - | - | - | - |
| 5 | 11 | 2 | 0.0608 | 2 | 0.0547 |
| 5 | 13 | - | - | - | - |
| 6 | 7 | - | - | - | - |
| 6 | 8 | 2 | 0.0106 | 2 | 0.0103 |
| 6 | 9 | - | - | - | - |
| 6 | 11 | 13 | 0.0924 | 13 | 0.0875 |
| 6 | 13 | 10 | 0.0407 | 10 | 0.0509 |
| 7 | 8 | - | - | - | - |
| 7 | 9 | - | - | - | - |
| 7 | 11 | 1 | 0.0797 | 1 | 0.0735 |
| 7 | 13 | - | - | - | - |
| 8 | 9 | - | - | - | - |
| 8 | 11 | 5 | 0.091 | 5 | 0.0806 |
| 8 | 13 | 2 | 0.0487 | 2 | 0.051 |
| 9 | 11 | 2 | 0.0908 | 2 | 0.0821 |
| 9 | 13 | 2 | 0.0417 | 2 | 0.0427 |
| 11 | 13 | - | - | - | - |

Tabla 5.6: Resultados obtenidos para la codificación independiente en el caso del método *difuso* para $l$-diversity y $t$-closeness.

## 5.2. Codificación conjunta

### 5.2.1. k-medias: $k$-anonymity y measure_risk

| Pares | | $k$-anonymity | *measure_risk* |
|---|---|---|---|
| 1 | 2 | 165 | 165 |
| 1 | 3 | 165 | 165 |
| 1 | 4 | 165 | 165 |
| 1 | 5 | 164 | 164 |
| 1 | 6 | 164 | 164 |
| 1 | 7 | 165 | 165 |
| 1 | 8 | 165 | 165 |
| 1 | 9 | 164 | 164 |
| 1 | 11 | 165 | 165 |
| 1 | 13 | 165 | 165 |
| 2 | 3 | 335 | 0 |
| 2 | 4 | 340 | 0 |
| 2 | 5 | 320 | 0 |
| 2 | 6 | 339 | 0 |
| 2 | 7 | 318 | 0 |
| 2 | 8 | 336 | 0 |
| 2 | 9 | 339 | 0 |
| 2 | 11 | 339 | 0 |
| 2 | 13 | 301 | 0 |
| 3 | 4 | 230 | 230 |
| 3 | 5 | 220 | 220 |
| 3 | 6 | 275 | 275 |
| 3 | 7 | 325 | 0 |
| 3 | 8 | 32 | 181 |
| 3 | 9 | 18 | 429 |
| 3 | 11 | 324 | 0 |
| 3 | 13 | 209 | 209 |
| 4 | 5 | 182 | 182 |
| 4 | 6 | 246 | 246 |

Sigue en la página siguiente.

| Pares | | $k$-anonymity | *measure_risk* |
|---|---|---|---|
| 4 | 7 | 326 | 0 |
| 4 | 8 | 66 | 485 |
| 4 | 9 | 20 | 485 |
| 4 | 11 | 236 | 236 |
| 4 | 13 | 264 | 264 |
| 5 | 6 | 221 | 221 |
| 5 | 7 | 285 | 0 |
| 5 | 8 | 199 | 199 |
| 5 | 9 | 220 | 220 |
| 5 | 11 | 220 | 220 |
| 5 | 13 | 236 | 236 |
| 6 | 7 | 328 | 0 |
| 6 | 8 | 33 | 182 |
| 6 | 9 | 18 | 306 |
| 6 | 11 | 239 | 239 |
| 6 | 13 | 214 | 214 |
| 7 | 8 | 324 | 0 |
| 7 | 9 | 323 | 0 |
| 7 | 11 | 331 | 0 |
| 7 | 13 | 290 | 0 |
| 8 | 9 | 34 | 198 |
| 8 | 11 | 33 | 182 |
| 8 | 13 | 64 | 517 |
| 9 | 11 | 18 | 488 |
| 9 | 13 | 266 | 266 |
| 11 | 13 | 266 | 266 |

Tabla 5.7: Resultados obtenidos para la codificación independiente en el caso del método de las *k-medias* para $k$-anonymity y *measure_risk*.

### 5.2.2. k-medias: $l$-diversity y $t$-closeness

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 1 | 2 | 99 | 0.0091 | 98 | 0.0104 |
| 1 | 3 | 99 | 0.009 | 98 | 0.0096 |
| 1 | 4 | 99 | 0.009 | 98 | 0.0096 |
| 1 | 5 | 98 | 0.008 | 97 | 0.0082 |
| 1 | 6 | 98 | 0.0093 | 97 | 0.0091 |
| 1 | 7 | 99 | 0.009 | 98 | 0.0101 |
| 1 | 8 | 99 | 0.009 | 98 | 0.0096 |
| 1 | 9 | 98 | 0.0093 | 97 | 0.0091 |
| 1 | 11 | 99 | 0.009 | 98 | 0.0096 |
| 1 | 13 | 99 | 0.009 | 98 | 0.0096 |
| 2 | 3 | 151 | 0.0814 | 143 | 0.0811 |
| 2 | 4 | 149 | 0.0847 | 141 | 0.0846 |
| 2 | 5 | 126 | 0.0963 | 118 | 0.0976 |
| 2 | 6 | 149 | 0.0837 | 143 | 0.0835 |
| 2 | 7 | 139 | 0.0871 | 130 | 0.0872 |
| 2 | 8 | 151 | 0.0831 | 143 | 0.083 |
| 2 | 9 | 149 | 0.0837 | 143 | 0.0835 |
| 2 | 11 | 149 | 0.0837 | 143 | 0.0835 |
| 2 | 13 | 105 | 0.0781 | 100 | 0.0834 |
| 3 | 4 | 108 | 0.1046 | 101 | 0.1091 |
| 3 | 5 | 97 | 0.1419 | 93 | 0.1398 |
| 3 | 6 | 125 | 0.1103 | 120 | 0.1141 |
| 3 | 7 | 143 | 0.0825 | 134 | 0.0836 |
| 3 | 8 | 26 | 0.0289 | 26 | 0.0303 |
| 3 | 9 | 18 | 0.082 | 17 | 0.08 |
| 3 | 11 | 126 | 0.0956 | 125 | 0.0901 |
| 3 | 13 | 73 | 0.1972 | 70 | 0.2061 |
| 4 | 5 | 83 | 0.1522 | 80 | 0.151 |
| 4 | 6 | 115 | 0.0876 | 108 | 0.0902 |
| 4 | 7 | 143 | 0.0818 | 134 | 0.0829 |
| 4 | 8 | 50 | 0.1108 | 49 | 0.1057 |

Sigue en la página siguiente.

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 4 | 9 | 20 | 0.0614 | 19 | 0.0447 |
| 4 | 11 | 108 | 0.1055 | 101 | 0.1102 |
| 4 | 13 | 88 | 0.1652 | 85 | 0.1771 |
| 5 | 6 | 97 | 0.1412 | 93 | 0.1392 |
| 5 | 7 | 117 | 0.0953 | 109 | 0.0979 |
| 5 | 8 | 97 | 0.159 | 93 | 0.1588 |
| 5 | 9 | 97 | 0.1426 | 93 | 0.1406 |
| 5 | 11 | 97 | 0.1419 | 93 | 0.1398 |
| 5 | 13 | 88 | 0.1583 | 84 | 0.1585 |
| 6 | 7 | 143 | 0.0798 | 134 | 0.0809 |
| 6 | 8 | 27 | 0.0289 | 27 | 0.0303 |
| 6 | 9 | 18 | 0.0628 | 17 | 0.0602 |
| 6 | 11 | 111 | 0.1611 | 108 | 0.1589 |
| 6 | 13 | 76 | 0.1887 | 73 | 0.1932 |
| 7 | 8 | 144 | 0.0877 | 135 | 0.0895 |
| 7 | 9 | 142 | 0.0819 | 133 | 0.0833 |
| 7 | 11 | 143 | 0.0806 | 134 | 0.0811 |
| 7 | 13 | 102 | 0.0884 | 98 | 0.0967 |
| 8 | 9 | 28 | 0.024 | 28 | 0.0257 |
| 8 | 11 | 27 | 0.0289 | 27 | 0.0303 |
| 8 | 13 | 49 | 0.1326 | 48 | 0.1461 |
| 9 | 11 | 18 | 0.0935 | 17 | 0.0879 |
| 9 | 13 | 88 | 0.1659 | 85 | 0.1781 |
| 11 | 13 | 88 | 0.167 | 85 | 0.1793 |

Tabla 5.8: Resultados obtenidos para la codificación independiente en el caso del método de las *k-medias* para $l$-diversity y $t$-closeness.

### 5.2.3. Gustafson-Kessel: $Q$-anonymity. $k$-anonymity y measure_risk

| Pares | | $Q$-anonymity | $k$-anonymity | *measure_risk* |
|---|---|---|---|---|
| 1 | 2 | 199 | 221 | 162 |
| 1 | 3 | 185 | 211 | 211 |
| 1 | 4 | 188 | 212 | 163 |
| 1 | 5 | 189 | 203 | 203 |
| 1 | 6 | 181 | 203 | 0 |
| 1 | 7 | 201 | 222 | 222 |
| 1 | 8 | 65 | 66 | 411 |
| 1 | 9 | 41 | 41 | 265 |
| 1 | 11 | 307 | 325 | 0 |
| 1 | 13 | 170 | 200 | 200 |
| 2 | 3 | 297 | 323 | 0 |
| 2 | 4 | 258 | 273 | 273 |
| 2 | 5 | 251 | 283 | 283 |
| 2 | 6 | 246 | 262 | 262 |
| 2 | 7 | 303 | 335 | 335 |
| 2 | 8 | 136 | 147 | 424 |
| 2 | 9 | 57 | 62 | 527 |
| 2 | 11 | 234 | 255 | 255 |
| 2 | 13 | 264 | 288 | 0 |
| 3 | 4 | 313 | 350 | 0 |
| 3 | 5 | 320 | 351 | 0 |
| 3 | 6 | 282 | 319 | 459 |
| 3 | 7 | 299 | 322 | 0 |
| 3 | 8 | 140 | 150 | 318 |
| 3 | 9 | 97 | 105 | 243 |
| 3 | 11 | 314 | 343 | 0 |
| 3 | 13 | 316 | 344 | 344 |
| 4 | 5 | 205 | 229 | 138 |
| 4 | 6 | 217 | 232 | 0 |
| 4 | 7 | 191 | 192 | 270 |
| 4 | 8 | 147 | 151 | 409 |

Sigue en la página siguiente.

| Pares | | $Q$-anonymity | $k$-anonymity | *measure_risk* |
|---|---|---|---|---|
| 4 | 9 | 43 | 50 | 415 |
| 4 | 11 | 192 | 209 | 209 |
| 4 | 13 | 205 | 221 | 0 |
| 5 | 6 | 202 | 219 | 219 |
| 5 | 7 | 241 | 264 | 412 |
| 5 | 8 | 76 | 92 | 470 |
| 5 | 9 | 27 | 28 | 215 |
| 5 | 11 | 127 | 133 | 412 |
| 5 | 13 | 150 | 157 | 513 |
| 6 | 7 | 243 | 255 | 255 |
| 6 | 8 | 136 | 142 | 345 |
| 6 | 9 | 84 | 94 | 135 |
| 6 | 11 | 277 | 326 | 0 |
| 6 | 13 | 217 | 242 | 0 |
| 7 | 8 | 138 | 150 | 434 |
| 7 | 9 | 44 | 44 | 263 |
| 7 | 11 | 225 | 254 | 254 |
| 7 | 13 | 237 | 264 | 0 |
| 8 | 9 | 138 | 138 | 290 |
| 8 | 11 | 134 | 136 | 451 |
| 8 | 13 | 142 | 147 | 473 |
| 9 | 11 | 76 | 80 | 330 |
| 9 | 13 | 75 | 79 | 220 |
| 11 | 13 | 98 | 99 | 331 |

Tabla 5.9: Resultados obtenidos para la codificación independiente en el caso del método de *Gustafson-Kessel* para $Q$-anonymity. $k$-anonymity y *measure_risk*.

### 5.2.4. Gustafson-Kessel: $l$-diversity y $t$-closeness

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 1 | 2 | 106 | 0.0035 | 102 | 0.0032 |
| 1 | 3 | 107 | 0.0026 | 103 | 0.0031 |
| 1 | 4 | 102 | 0.0038 | 100 | 0.0028 |
| 1 | 5 | 104 | 0.0053 | 99 | 0.0048 |
| 1 | 6 | 92 | 0.0029 | 89 | 0.0027 |
| 1 | 7 | 112 | 0.0043 | 108 | 0.0025 |
| 1 | 8 | 50 | 0.0006 | 49 | 0.0008 |
| 1 | 9 | 33 | 0.0011 | 33 | 0.001 |
| 1 | 11 | 90 | 0.0283 | 86 | 0.0256 |
| 1 | 13 | 90 | 0.0049 | 88 | 0.0029 |
| 2 | 3 | 140 | 0.0055 | 134 | 0.0046 |
| 2 | 4 | 117 | 0.016 | 110 | 0.0137 |
| 2 | 5 | 110 | 0.0013 | 106 | 0.0015 |
| 2 | 6 | 113 | 0.0175 | 109 | 0.0163 |
| 2 | 7 | 137 | 0.0011 | 133 | 0.0013 |
| 2 | 8 | 84 | 0.0007 | 79 | 0.0029 |
| 2 | 9 | 41 | 0.0013 | 40 | 0.0015 |
| 2 | 11 | 103 | 0.0016 | 98 | 0.0016 |
| 2 | 13 | 103 | 0.001 | 97 | 0.0009 |
| 3 | 4 | 131 | 0.005 | 124 | 0.0054 |
| 3 | 5 | 119 | 0.0019 | 116 | 0.0019 |
| 3 | 6 | 118 | 0.0003 | 112 | 0.0017 |
| 3 | 7 | 135 | 0.0021 | 132 | 0.0015 |
| 3 | 8 | 86 | 0.0027 | 82 | 0.002 |
| 3 | 9 | 70 | 0.0026 | 69 | 0.0031 |
| 3 | 11 | 121 | 0.0029 | 119 | 0.0033 |
| 3 | 13 | 109 | 0.0009 | 105 | 0.0016 |
| 4 | 5 | 94 | 0.0056 | 90 | 0.0035 |
| 4 | 6 | 104 | 0.0097 | 99 | 0.01 |
| 4 | 7 | 96 | 0.0051 | 91 | 0.004 |
| 4 | 8 | 81 | 0.0041 | 76 | 0.0054 |

Sigue en la página siguiente.

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 4 | 9 | 35 | 0.001 | 34 | 0.0007 |
| 4 | 11 | 79 | 0.0007 | 77 | 0.0009 |
| 4 | 13 | 104 | 0.0032 | 94 | 0.0041 |
| 5 | 6 | 88 | 0.0078 | 86 | 0.0074 |
| 5 | 7 | 108 | 0.0034 | 103 | 0.0049 |
| 5 | 8 | 52 | 0.0013 | 52 | 0.0028 |
| 5 | 9 | 21 | 0.0005 | 21 | 0.0004 |
| 5 | 11 | 84 | 0.013 | 83 | 0.0126 |
| 5 | 13 | 74 | 0.0023 | 71 | 0.0014 |
| 6 | 7 | 115 | 0.0137 | 110 | 0.0116 |
| 6 | 8 | 76 | 0.0036 | 73 | 0.0033 |
| 6 | 9 | 55 | 0.0012 | 54 | 0.0021 |
| 6 | 11 | 95 | 0.0197 | 91 | 0.0187 |
| 6 | 13 | 93 | 0.0026 | 89 | 0.0025 |
| 7 | 8 | 84 | 0.0024 | 79 | 0.0016 |
| 7 | 9 | 34 | 0.0002 | 33 | 0.0003 |
| 7 | 11 | 95 | 0.0017 | 93 | 0.0022 |
| 7 | 13 | 113 | 0.0039 | 102 | 0.0048 |
| 8 | 9 | 82 | 0.0027 | 81 | 0.003 |
| 8 | 11 | 88 | 0.0072 | 88 | 0.0081 |
| 8 | 13 | 66 | 0.0064 | 63 | 0.0053 |
| 9 | 11 | 59 | 0.006 | 58 | 0.0066 |
| 9 | 13 | 57 | 0.0017 | 57 | 0.0028 |
| 11 | 13 | 70 | 0.017 | 67 | 0.012 |

Tabla 5.10: Resultados obtenidos para la codificación independiente en el caso del método de *Gustafson-Kessel* para $l$-diversity y $t$-closeness.

### 5.2.5. c-medias difuso: $Q$-anonymity. $k$-anonymity y measure_risk

| Pares | | $Q$-anonymity | $k$-anonymity | *measure_risk* |
|---|---|---|---|---|
| 1 | 2 | 319 | 334 | 0 |
| 1 | 3 | 187 | 225 | 225 |
| 1 | 4 | 280 | 323 | 0 |
| 1 | 5 | 173 | 201 | 201 |
| 1 | 6 | 190 | 214 | 0 |
| 1 | 7 | 314 | 330 | 0 |
| 1 | 8 | 163 | 172 | 172 |
| 1 | 9 | 167 | 172 | 172 |
| 1 | 11 | 174 | 197 | 197 |
| 1 | 13 | 193 | 230 | 230 |
| 2 | 3 | 311 | 331 | 0 |
| 2 | 4 | 298 | 307 | 0 |
| 2 | 5 | 305 | 312 | 0 |
| 2 | 6 | 303 | 319 | 0 |
| 2 | 7 | 318 | 327 | 0 |
| 2 | 8 | 326 | 334 | 0 |
| 2 | 9 | 328 | 333 | 0 |
| 2 | 11 | 315 | 326 | 0 |
| 2 | 13 | 298 | 306 | 0 |
| 3 | 4 | 289 | 334 | 334 |
| 3 | 5 | 296 | 325 | 0 |
| 3 | 6 | 248 | 292 | 0 |
| 3 | 7 | 305 | 331 | 0 |
| 3 | 8 | 305 | 322 | 0 |
| 3 | 9 | 310 | 321 | 0 |
| 3 | 11 | 306 | 324 | 0 |
| 3 | 13 | 297 | 324 | 0 |
| 4 | 5 | 235 | 242 | 242 |
| 4 | 6 | 247 | 258 | 258 |
| 4 | 7 | 291 | 295 | 0 |
| 4 | 8 | 241 | 261 | 261 |

Sigue en la página siguiente.

| Pares | | $Q$-anonymity | $k$-anonymity | *measure_risk* |
|---|---|---|---|---|
| 4 | 9 | 239 | 250 | 250 |
| 4 | 11 | 265 | 276 | 276 |
| 4 | 13 | 237 | 256 | 256 |
| 5 | 6 | 207 | 223 | 223 |
| 5 | 7 | 280 | 287 | 0 |
| 5 | 8 | 194 | 216 | 216 |
| 5 | 9 | 205 | 220 | 220 |
| 5 | 11 | 303 | 315 | 315 |
| 5 | 13 | 237 | 243 | 243 |
| 6 | 7 | 289 | 305 | 0 |
| 6 | 8 | 208 | 224 | 224 |
| 6 | 9 | 204 | 218 | 218 |
| 6 | 11 | 282 | 304 | 304 |
| 6 | 13 | 264 | 277 | 277 |
| 7 | 8 | 309 | 321 | 0 |
| 7 | 9 | 309 | 320 | 0 |
| 7 | 11 | 318 | 332 | 0 |
| 7 | 13 | 283 | 295 | 0 |
| 8 | 9 | 15 | 20 | 160 |
| 8 | 11 | 296 | 316 | 0 |
| 8 | 13 | 266 | 274 | 274 |
| 9 | 11 | 320 | 324 | 0 |
| 9 | 13 | 252 | 259 | 259 |
| 11 | 13 | 312 | 316 | 0 |

Tabla 5.11: Resultados obtenidos para la codificación independiente en el caso del método de las *c-medias difuso* para $Q$-anonymity. $k$-anonymity y *measure_risk*.

### 5.2.6. c-medias difuso: $l$-diversity y $t$-closeness

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 1 | 2 | 131 | 0.0127 | 128 | 0.0101 |
| 1 | 3 | 102 | 0.0021 | 98 | 0.0015 |
| 1 | 4 | 120 | 0.0076 | 116 | 0.0049 |
| 1 | 5 | 89 | 0.0031 | 85 | 0.0016 |
| 1 | 6 | 103 | 0.0017 | 101 | 0.0014 |
| 1 | 7 | 132 | 0.0115 | 127 | 0.0101 |
| 1 | 8 | 94 | 0.0003 | 93 | 0.0005 |
| 1 | 9 | 98 | 0.0006 | 97 | 0.0007 |
| 1 | 11 | 89 | 0.0023 | 86 | 0.001 |
| 1 | 13 | 84 | 0.0035 | 80 | 0.0041 |
| 2 | 3 | 128 | 0.0084 | 122 | 0.0085 |
| 2 | 4 | 132 | 0.0228 | 124 | 0.0201 |
| 2 | 5 | 123 | 0.0242 | 115 | 0.0231 |
| 2 | 6 | 129 | 0.0196 | 123 | 0.0175 |
| 2 | 7 | 139 | 0.0258 | 130 | 0.0226 |
| 2 | 8 | 141 | 0.0229 | 134 | 0.0237 |
| 2 | 9 | 148 | 0.0182 | 141 | 0.0201 |
| 2 | 11 | 130 | 0.0236 | 124 | 0.0222 |
| 2 | 13 | 104 | 0.0216 | 99 | 0.0205 |
| 3 | 4 | 123 | 0.0051 | 117 | 0.0043 |
| 3 | 5 | 115 | 0.005 | 112 | 0.0047 |
| 3 | 6 | 112 | 0.0012 | 108 | 0.0009 |
| 3 | 7 | 130 | 0.0091 | 124 | 0.0078 |
| 3 | 8 | 136 | 0.0006 | 130 | 0.0017 |
| 3 | 9 | 136 | 0.0008 | 130 | 0.0019 |
| 3 | 11 | 120 | 0.0152 | 114 | 0.0155 |
| 3 | 13 | 101 | 0.0084 | 97 | 0.0098 |
| 4 | 5 | 102 | 0.0083 | 96 | 0.0092 |
| 4 | 6 | 113 | 0.0144 | 108 | 0.0129 |
| 4 | 7 | 131 | 0.022 | 123 | 0.0184 |
| 4 | 8 | 107 | 0.0076 | 101 | 0.0073 |

Sigue en la página siguiente.

| Pares | | *PTOTVAL* | | *TAXINC* | |
|---|---|---|---|---|---|
| | | $l$-diversity | $t$-closeness | $l$-diversity | $t$-closeness |
| 4 | 9 | 109 | 0.0142 | 103 | 0.0102 |
| 4 | 11 | 106 | 0.0094 | 100 | 0.0103 |
| 4 | 13 | 95 | 0.012 | 91 | 0.0125 |
| 5 | 6 | 89 | 0.0125 | 86 | 0.0098 |
| 5 | 7 | 114 | 0.0264 | 106 | 0.0244 |
| 5 | 8 | 85 | 0.0174 | 81 | 0.0153 |
| 5 | 9 | 93 | 0.0232 | 89 | 0.0233 |
| 5 | 11 | 108 | 0.0237 | 104 | 0.0216 |
| 5 | 13 | 87 | 0.0204 | 83 | 0.02 |
| 6 | 7 | 126 | 0.0177 | 119 | 0.0152 |
| 6 | 8 | 96 | 0.0033 | 93 | 0.0045 |
| 6 | 9 | 99 | 0.0067 | 96 | 0.0057 |
| 6 | 11 | 106 | 0.0074 | 102 | 0.0077 |
| 6 | 13 | 90 | 0.0178 | 86 | 0.0184 |
| 7 | 8 | 134 | 0.0222 | 126 | 0.0177 |
| 7 | 9 | 139 | 0.0211 | 131 | 0.0198 |
| 7 | 11 | 120 | 0.0186 | 114 | 0.0159 |
| 7 | 13 | 101 | 0.0182 | 97 | 0.017 |
| 8 | 9 | 15 | 0.0007 | 14 | 0.0006 |
| 8 | 11 | 100 | 0.0182 | 94 | 0.0181 |
| 8 | 13 | 89 | 0.0126 | 86 | 0.0129 |
| 9 | 11 | 101 | 0.0194 | 94 | 0.0219 |
| 9 | 13 | 81 | 0.0113 | 78 | 0.0134 |
| 11 | 13 | 101 | 0.0274 | 97 | 0.0246 |

Tabla 5.12: Resultados obtenidos para la codificación independiente en el caso del método de las *c-medias difuso* para $l$-diversity y $t$-closeness.

# Índice de tablas

# Índice de figuras

# Bibliografía


[1] J. Abonyi, B. Balasko, and B. Feil. Fuzzy clustering and data analysis toolbox. `http://www.sunfinedata.com/wp-content/uploads/2009/10/FuzzyClusteringToolbox.pdf`, Apr. 2005.

[2] J. Bibby, K. Mardia, and J. Kent. *Multivariate Analysis*. Academic Press, 1982.

[3] A. Brodsky, C. Farkas, and S. Jajodia. Unauthorized inferences in semistructured databases. *Information Sciences*, 176(22):3269 – 3299, 2006.

[4] J. Casasnovas and J. Torrens. An axiomatic approach to fuzzy cardinalities of finite fuzzy sets. *Fuzzy Sets and Systems*, 133(2):193–209, 2003.

[5] M. D. Center. $k$-means clustering. `http://www.mathworks.es/es/help/stats/kmeans.html`.

[6] R. Chen, B. Desai, B. Fung, N. Mohammed, and K. Wang. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, in press.

[7] V. Ciriani, S. D. C. di Vimercati, S. Foresti, and P. Samarati. Microdata protection. In *Secure Data Management in Decentralized Systems*, pages 291–321. 2007.

[8] I. Díaz, J. Ranilla, L. Rodríguez-Muñiz, and L. Troiano. Identifying the risk of attribute disclosure by mining fuzzy rules. In *Communications*

*in Computer and Information Science 80: Information Processing and Management of Uncertainty in Knowledge-Based Systems Theory and Methods*, pages 455–464. Springer, 2010.

[9] I. Díaz, L. Rodríguez-Muñiz, and L. Troiano. On a fuzzy-based paradigm to study data privacy with continuous attributes. (to appear), 2012.

[10] I. Díaz, L. Rodríguez-Muñiz, and L. Troiano. Fuzzy sets in data protection: strategies and cardinalities. *Logic Journal of IGPL*, in press.

[11] J. Domingo-Ferrer and U. González-Nicolás. Rational behavior in peer-to-peer profile obfuscation for anonymous keyword search. *Information Sciences*, 185(1):191 – 204, 2012.

[12] J. Domingo-Ferrer and J. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189 –201, jan/feb 2002.

[13] D. Dubois and H. Prade. *Fuzzy Sets and Systems. Theory and Applications*. Academic Press, Inc., 1980.

[14] D. Dubois and H. Prade. Fuzzy cardinality and the modeling of imprecise quantification. *Fuzzy Sets and Systems*, 16(3):199 – 230, 1985.

[15] T. Folger and G. Klir. *Fuzzy Sets, Uncertainty and Information*. Prentice-Hall International (UK) Limited., 1988.

[16] J. Gehrke, D. Kifer, A. Machanavajjhala, and M. Venkitasubramaniam. *l*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1), 2007.

[17] F. Hoppner, R. Klawonn, and T. Runkler. *Fuzzy Cluster Analysis*. Wiley, Chichester, 1999.

[18] A. Hundepool, R. Ramaswamy, and A. van de Wetering. $\mu$ argus. version 4.1. `http://neon.vb.cbs.nl/casc/Software/MuManual4.1.pdf`, Feb. 2007.

[19] L. Jang and D. Ralescu. Cardinality concepts for type-two fuzzy sets. *Fuzzy Sets Syst.*, 118(3):479–487, 2001.

[20] A. Kowarik, B. Meindl, and M. Templ. Statistical disclosure control methods for the generation of public- and scientific-use files. `http://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf`, Sept. 2012.

[21] M. Li, X. Sun, and H. Wang. A family of enhanced (l, alpha)-diversity models for privacy preserving data publishing. *Future Generation Comp. Syst.*, 27(3):348–356, 2011.

[22] N. Li and T. Li. Towards optimal k-anonymization. *Data Knowl. Eng.*, 65(1):22–39, 2008.

[23] N. Li, T. Li, and S. Venkatasubramanian. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In *ICDE*, pages 106–115. IEEE, 2007.

[24] O. Maimon, N. Matatov, and L. Rokach. Privacy-preserving data mining: A feature set partitioning approach. *Information Sciences*, 180(14):2696 – 2720, 2010.

[25] B. Malin. k-unlinkability: A privacy protection model for distributed data. *Data Knowl. Eng.*, 64(1):294–311, 2008.

[26] A. Ralescu. A note on rule representation in expert systems. *Inf. Sci.*, 38(2):193–203, 1986.

[27] D. Ralescu. Cardinality, quantifiers, and the aggregation of fuzzy criteria. *Fuzzy Sets Syst.*, 69(3):355–365, 1995.

[28] L. Rokach, B. Shapira, E. Shmueli, T. Tassa, and R. Wasserstein. Limiting disclosure of sensitive data in sequential releases of databases. *Information Sciences*, 191:98–127, 2012.

[29] E. Ruspini. On the semantics of fuzzy logic. *Int. J. Approx. Reasoning*, 5:45–88, January 1991.

[30] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. on Knowl. and Data Eng.*, 13(6):1010–1027, 2001.

[31] D. Shah and S. Zhong. Two methods for privacy preserving data mining with malicious participants. *Information Sciences*, 177(23):5468 – 5483, 2007.

[32] L. Sweeney. Achieving $k$-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[33] M. Wygralak. Questions of cardinality of finite fuzzy sets. *Fuzzy Sets Syst.*, 102(2):185–210, 1999.

[34] M. Wygralak. An axiomatic approach to scalar cardinalities of fuzzy sets. *Fuzzy Sets Syst.*, 110(2):175–179, 2000.

[35] S. Zhong. Privacy-preserving algorithms for distributed mining of frequent itemsets. *Information Sciences*, 177(2):490 – 503, 2007.