

Gestión de información difusa en bases de datos usando medidas de similitud

Management of Diffused Information in Databases Using Similarity Measures

José Manuel Rodríguez-Jiménez

Policía Local de Mijas

jmrodriguez@mijas.es | <https://orcid.org/0000-0003-3776-9887>

DOI: <https://doi.org/1014201/cp.32374>

Recibido: 03-07-2025 | Aceptado: 25-08-2025

Resumen

El concepto subjetivo diferente que tienen las personas cuando tratan ciertos datos que son similares hace que la transcripción de la información sea un problema existente en muchas bases de datos policiales. Conceptos como el color o la altura son en realidad valores difusos, ya que existen diferentes grados dentro de cada uno de estos valores. La información recopilada se suele almacenar con valores discretos, cuando en realidad son una representación de información que se podría denominar, siguiendo conceptos usados en ciencias que tratan la gestión de la información difusa. Al transformar la información difusa en estos valores discretos se puede perder información, o incluso tergiversarla, si no se valora adecuadamente la misma ni se comprueba la certidumbre de la fuente o la formación del encargado de gestionarla. Dentro de las matemáticas, el estudio de la lógica difusa intenta establecer relaciones entre estos valores para poder dar una solución al problema planteado. Las relaciones de similitud son quizás la mejor medida para poder valorar adecuadamente los datos existentes y ayudar a los investigadores. En este artículo se plantean 3 casos de estudio: información difusa, información parcial e información errónea. Para ello se van a usar ejemplos de tratamiento de colores en bases de datos, la evaluación de matrículas con datos parciales o la detección de anomalías en registros de personas.

Palabras clave: Medidas de similitud; Lógica difusa; Bases de datos; Investigación policial; Relaciones.

Abstract

The different subjective concept that people have when dealing with similar data makes the transcription of information a problem in many police databases. Concepts such as colour or height are actually fuzzy values, as there are different degrees within each of these values. The collected information is often stored with discrete values, when in reality they are a representation of information that could be called, following concepts used in sciences dealing with information management fuzzy. When transforming fuzzy information into these discrete values, information can be lost or even misrepresented if it is not properly valued and the certainty of the source or the training of the person in charge of managing it is not checked. Within mathematics, the study of fuzzy logic attempts to establish relationships between these values in order to provide a solution to the problem at hand. Similarity relations are perhaps the best measure to be able to properly assess the existing data and help researchers. In this article, three case studies are presented: fuzzy information, partial information and misinformation. Examples of colour processing in databases, the evaluation of number plates with partial data or the detection of anomalies in people's records will be used for this purpose.

Keywords: Similarity measures; Fuzzy logic; Data bases; Police research; Relationships.

1 Introducción

Dentro de los diferentes hechos delictivos registrados en las bases de datos policiales, los investigadores encargados se encuentran con el problema de conseguir establecer relaciones para poder ampliar la información obtenida de forma separada. Tratan de unirla recabando los datos suficientes y necesarios que conlleven al esclarecimiento de dichos hechos para llevar a los responsables ante la justicia.

La investigación no necesariamente busca conseguir un patrón que relacione todos los eventos, pero sí poder enlazar al menos la mayoría, a través de un arduo proceso que consiga

completar la información recibida de diferentes fuentes. La valoración de cada una de las fuentes es una situación complicada ya que los testimonios de testigos no son todo lo fiables que sería deseable, existiendo incluso contradicciones en las versiones de varias personas ante un mismo hecho que a veces no debería presentar dudas.

Cuando se le pregunta a una persona de qué color es el mar, la mayoría responde que es de color azul. Si se pregunta a la misma persona sobre el color del cielo la respuesta suele ser la misma, de color azul. Sin embargo, cuando se le pregunta si los considera del mismo color, la respuesta suele ser negativa. Si se anota en una base de datos, lo que queda registrado es simplemente la palabra azul, lo que en una comparativa indicaría que el mar y el cielo son del mismo color, hecho que normalmente no es así debido a multitud de factores como la ubicación, la hora del día, la contaminación, etc. No solamente se asigna el mismo valor a 2 colores que en la realidad son diferentes, sino que cualquier tonalidad de color similar a la transición entre el azul del mar y el azul del cielo, siendo preguntado el mismo encuestado, también se designaría posiblemente con el término azul.

Este hecho se produce porque se *discretiza*, es decir, se da un valor discreto o exacto a un valor que se puede considerar difuso, ya que no existe una definición clara de cómo debe ser un color.

Esta adaptación a valores discretos no solo se da en colores, sino también en otros datos anotados que se suelen interpretar como discretos cuando no lo son. En el caso de la altura, se puede anotar una altura en centímetros, pero siempre puede variar con la real ya que siempre se pueden añadir más decimales y se trataría el valor como un truncamiento de un valor difuso. Este dato se puede obtener de un examen físico de la persona a la hora de realizar la ficha policial, pero las víctimas o testigos nunca serán capaces de aportar la medición que coincida con esa ficha. En las descripciones de una persona para el concepto “altura” se suelen usar términos como “alto”, “bajo” o “altura media”, que no pueden determinarse exactamente con un número y dependen de la concepción subjetiva de la persona del concepto “altura” y de los términos lingüísticos en los que se expresa.

Para la edad, existe una complicación adicional al tener que considerarse la fecha en la que se produjeron los hechos denunciados. El valor de la edad se debe calcular a partir de la fecha de nacimiento, que es la que aparece en la ficha policial, la cual se debe tener en cuenta para calcular los años que tendría en el momento de los hechos a investigar. El valor estimado por los denunciantes puede incrementar o disminuir de forma considerable la similitud entre la edad real y la que se aporta en la denuncia. Como se ha descrito anteriormente, ante el concepto “edad”, es común que el valor aportado sea “mediana edad”, el cual es un valor bastante difícil de interpretar para el receptor de la información.

Existe además la conceptualización limitada por los términos que la base de datos de la herramienta policial incluye. A veces, el término “constitución-atlético” y “constitución-grueso o muy grueso” se emplea mal por la víctima o el testigo ya que no saben diferenciar entre ambos términos cuando se trata de una persona fornida. El término “pelo-teñido” podría ser interesante para la investigación si el color no fuera común, ya que una persona mayor puede tener el pelo teñido y el compareciente indicarlo por dicho motivo.

El poder establecer relaciones entre estos datos transformados en valores discretos hace que los investigadores puedan trabajar con probabilidades, o porcentajes, de que una persona o vehículo a los que buscan entre dentro del rango de la persona o vehículo descritos en la denuncia. Cuantos más datos se aporten, la exploración hará que los candidatos o sospechosos aumenten o disminuyan su probabilidad de ser los buscados.

El estudio de la lógica difusa fue introducido por Zadeh en el año 1965, contando este autor con multitud de artículos relacionados sobre cómo estudiar las relaciones de similitud existente entre datos difusos (Zadeh, 1965, 1971, 1992, 1996). En ellos, basándose en relaciones de similitud, establece los principios por los cuales se pueden interpretar datos difusos para extraer conocimiento y poder trabajar con ellos de forma adecuada.

El modelo relacional de Codd para bases de datos (Codd, 1990) se ha convertido en un modelo estándar en el área de las bases de datos. En estos años se han propuesto algunas extensiones al

modelo original para poder adaptarlo a determinados entornos y aplicaciones a la vida real, como es el caso de la lógica difusa, donde el modelo debería ser más expresivo. Uno de los modelos propuestos que más se aproximan al motivo de este estudio es el de la Tabla de Atributos Difusos (Cordero *et al.*, 2011), si bien es necesario modificar la base de datos policial para poder usar dicho modelo, motivo por el cual no se puede emplear en la actualidad.

El uso de toda esta teoría que adapta el uso de los valores difusos a las tablas de datos *clásicas*, como pueden ser las bases de datos policiales, permite establecer nuevos puntos de vista hacia los resultados que se pueden obtener mediante la exploración de conceptos como las medidas de similitud.

La consideración de que se pueda contener información errónea de partida en una investigación, casi siempre proporcionada de forma involuntaria, hace considerar que se debe acotar cualquier conclusión aportada por este tipo de información. Este problema es complejo ya que, a priori, se desconoce que dicha información es errónea, siendo un hándicap para los investigadores.

Aunque se posea una base de datos con valores discretos, como la matrícula o la marca de un vehículo, la información aportada por los testigos o colaboradores en una investigación no tiene por qué consistir en la aportación de datos que puedan considerarse discretos. Ante una incidencia con un vehículo, además del color del mismo, se suelen aportar datos sobre la matrícula, pero no siempre de forma correcta o completa, siendo los datos conseguidos una parte de la numeración o la serie, hecho que podemos denominar aportación de *información parcial*.

El objetivo de este artículo es considerar diferentes casuísticas con las que se pueden encontrar los investigadores en los que los datos que se evalúan, tanto en las bases de datos oficiales como en las bases de datos policiales, tienen la consideración de datos difusos, planteando propuestas para el tratamiento de esta información. El empleo del término probabilidad, al no poder cuantificarse adecuadamente la información difusa y consistir los valores aportados en aproximaciones de los mismos, no se considera adecuado, siendo sustituido por medidas de similitud. Se van a usar como sinónimos los términos grado de similitud y

medida de similitud, siendo el primero más específico para definir lo que se quiere representar.

Los casos teóricos planteados, en apariencia sencillos, introducen la base teórica para un estudio más profundo de la metodología usada y cómo puede adaptarse a casos más complejos. Esta teoría es la base de los motores de búsqueda y análisis de algunos programas de inteligencia artificial, los cuales pueden ser de interés para la labor policial.

2 Método

Antes de iniciar el desarrollo de la metodología, se van a establecer varias definiciones para concretar los términos que se usarán durante la misma. Partiendo del hecho de que la base de datos policial usa datos transformados en discretos para representar datos que en realidad son difusos, cabe hacer una valoración de los diferentes campos de una base de datos y los atributos que pueden figurar en los mismos. La forma de tratar los datos es diferente en cada base de datos ya que, por ejemplo, los campos que aparecen en la base de datos de personas no son los mismos que los campos que pueden aparecer en una base de datos de vehículos. Desde este punto de vista se planteará el uso de una base de datos específica para el caso de estudio en cuestión. Definamos inicialmente los conceptos de “objetos” y “atributos”.

Definición: Se define atributo como cada uno de los valores que pueden tomar los datos en un conjunto dado.

Ejemplo: Para el campo color de pelo, son atributos posibles rubio, castaño, moreno, etc.

Definición: Se define objeto como una colección de atributos determinada previamente.

Ejemplo: Los atributos de un objeto, en este caso una persona, pueden ser nombre = María, color de pelo = rubio, edad = 43.

El hecho de usar la denominación “objeto” es por la generalización del uso del término, sobre todo en bases de datos y su uso en áreas como la programación o el Análisis de Conceptos Formales (Ganter *et al.*, 2005).

Se representa que un conjunto de atributos $a_1, \dots, a_n$, pertenece a un objeto obj de la forma $(obj)R(a_1, \dots, a_n)$. Nótese que, para un mismo campo, el objeto no puede tener 2 atributos, ya que sería contradictorio.

Definición: Se define una medida de similitud r entre 2 miembros x , y de un conjunto X , como aquella que toma valores entre 0 y 1, y cumple las propiedades reflexiva y simétrica.

1) Reflexiva: Para todo x perteneciente a X , $r(x, x) = 1$.

2) Simétrica: Para todo x , y pertenecientes a X , $r(x, y) = r(y, x)$.

Existe una tercera propiedad denominada transitiva que caracteriza a las medidas de similitud como relaciones de similitud.

3) Transitiva: Para todo x , y , z pertenecientes a X , $r(x, z) \geq \min\{r(x, y), r(y, z)\}$

Que se cumpla la propiedad transitiva, y por tanto se acredite que exista una relación de similitud, dota al análisis de los datos de una serie de herramientas y metodologías que hacen que el tratamiento de los mismos se pueda gestionar de una forma más eficiente. Sin embargo, no va a ser considerada para este artículo, aunque puede cumplirse para ciertos casos.

Como ejemplo de que habitualmente no se cumple en algunas bases de datos la propiedad transitiva, vamos a usar las medidas de similitud existentes entre las imágenes de los números del 0 al 9 en una zona de lectura mecánica de un pasaporte, que serán desarrolladas posteriormente.

Ejemplo: $r(0, 2) = 0.58$, $r(0, 4) = 0.49$, $r(2, 4) = 0.43$. No se cumple la propiedad transitiva ya que $r(2, 4) = 0.43 < \min\{r(2, 0), r(0, 4)\} = 0.49$.

La forma de representar los datos en una tabla puede ser variada, dependiendo de la metodología que se estime como la más adecuada para la investigación. En un contexto formal, se indican todos los atributos posibles, marcándose con valores booleanos, sí o no, si ese atributo pertenece al objeto. Se marcará el valor afirmativo del objeto obj con el atributo a en el caso de que $(obj)R(a)$. Estos contextos formales permiten el estudio a través del análisis de conceptos formales, siendo parte fundamental de las dependencias funcionales los axiomas de Armstrong (Armstrong, 1974), los cuales sirven de base al razonamiento empleado en diversos estudios. Debido a que puede haber ausencia de datos, no cabe aplicar la metodología que emplea atributos negativos (Rodríguez-Jiménez *et al.*, 2014) para obtener, o en este caso descartar, información.

Debido a que el número de atributos posibles puede ser considerablemente extenso y la información que aportan puede ser inexistente, en la metodología usada en este estudio se va a plantear el establecimiento de una base de datos donde aparezca la información proporcionada por los denunciantes o testigos con los atributos adaptados para un mejor ajuste a las medidas de similitud propuestas.

Algunas medidas de similitud provienen de distancias matemáticas definidas sobre varias coordenadas. Si se define un objeto como conjunto de coordenadas de cada punto en un espacio de varias dimensiones, todos sus valores serán numéricos. Para referirse a este tipo especial de objetos, se usará la denominación “tupla”. Al ser todos los valores numéricos, es sencillo operar con ellos y sobre estas tuplas se calculan las diferentes distancias. Estas distancias no tienen necesariamente que aportar el mismo valor y pueden adaptarse a un caso específico por sus características. Entre las distancias más comunes se encuentran:

- Distancia euclídea: Dados 2 puntos en el espacio n -dimensional cuyas tuplas son $(x_1, \dots, x_n)$ e $(y_1, \dots, y_n)$, $d_e = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
- Distancia absoluta o Manhattan: Dados 2 puntos en el espacio n -dimensional cuyas tuplas son $(x_1, \dots, x_n)$ e $(y_1, \dots, y_n)$, $d_a = \sum_{i=1}^n |x_i - y_i|$

- Distancia de Chebyshov: Dados 2 puntos en el espacio n-dimensional cuyas tuplas son $(x_1, \dots, x_n)$ e $(y_1, \dots, y_n)$, $d_c = \max_{i=1, \dots, n} |x_i - y_i|$

Estas distancias están pensadas para un caso n-dimensional sobre puntos en el espacio y sus valores son siempre positivos, pero no necesariamente acotados. Si se desea obtener un número representativo para poder comparar los resultados es necesario normalizarlas para que sus valores estén en el intervalo $[0,1]$.

Para mejor comprensión de los resultados, a los valores de las diferentes medidas de similitud propuestas se les denominará grado de similitud y se denotarán con una letra griega diferente para cada uno de los casos.

77

Estos grados que se van a definir, al estar acotados y normalizados en el intervalo $[0,1]$, permiten también establecer medidas compuestas, otorgando pesos a cada una de las medidas individuales que lo componen. No necesariamente estos pesos han de ser iguales, pudiendo adaptarse las medidas a las necesidades de los investigadores. De esta forma, si una medida α se quiere complementar con una medida β , puede definirse una nueva medida $\lambda = a*\alpha + b*\beta$, siendo $a + b = 1$, con a y $b \geq 0$.

2.1 Información difusa: similitud en colores

Como se ha mencionado anteriormente, los objetos que contengan un atributo del tipo color han de gestionar un valor para ese atributo que se ha de transformar en discreto, siendo el valor que da origen a ese atributo un valor difuso.

La adaptación es importante en este tipo de campo color, no siendo fácil dar respuesta a qué valor asignar al atributo. La propuesta que se va a realizar, para aplicar una metodología basada en ella, es una conversión de datos según la decodificación de colores RGB. Esta decodificación convierte cada color en un conjunto de 3 números que representan una tupla de 3 valores entre 0 y 255. En cada atributo de dicha tupla se refleja la cantidad aportada de los colores rojo (R-ed), verde (G-reen) y azul (B-lue) respectivamente. Esta nomenclatura aparece en

los sistemas de codificación de imágenes usados por diferentes herramientas informáticas y se considera un estándar a nivel internacional. Existen valores predeterminados para ciertos colores que se consideran como referencia, como por ejemplo el rojo, que se codifica (255, 0, 0); el blanco, que se codifica (255, 255, 255), o el negro, que se codifica (0, 0, 0).

Esto implica que la definición de un objeto color puede usar 2 definiciones, la discreta que usa el nombre del color y la numérica, que usa una tupla. De esta forma se podría denotar (color1) R (blanco) o (color1) R (255,255,255).

Los valores RGB también pueden encontrarse decodificados de forma hexadecimal, variando entre 00 y FF los valores que aparecen para su definición. La codificación hexadecimal usa los valores 0-9 y A-F, siendo A=10, ... , F=15. Un valor hexadecimal XY se puede transformar a un valor numérico entre 0 y 255 usando la formula $X*16+Y$.

Con la conversión a valores numéricos de los colores usando la decodificación RGB, la creación de una medida de similitud se puede definir de una manera más objetiva usando la distancia euclídea para 3 dimensiones. Por definición, la distancia euclídea no tiene necesariamente que estar acotada, sin embargo, en este caso particular, al referirse a un conjunto finito de valores, sí lo está. Esta acotación nos permite saber el valor máximo que toma dicha distancia y poder aplicar un factor corrector que nos permite establecer que los valores resultantes pertenezcan al intervalo [0,1], donde 0 indicaría que los colores no se parecen en absoluto y 1 que son iguales. Esta escala, que se gradúa a través de una distancia preestablecida, permite considerar una aproximación de una relación ya que, cuanto mayor sea el valor, más se parecerán los colores.

Definición: Siendo el color 1 representado con la tupla (r1, g1, b1) y el color 2 con la tupla (r2, g2, b2), puede definirse el grado de similitud entre ambos colores como

$$\mu(\text{color1}, \text{color2}) = 1 - \frac{\sqrt{(r1-r2)^2 + (g1-g2)^2 + (b1-b2)^2}}{(255*\sqrt{3})}$$

Ejemplo: Dados los colores predefinidos blanco= (255, 255, 255), negro= (0, 0, 0), gris= (128, 128, 128), azul marino= (0, 0, 128), verde oliva= (128, 128, 0) y morado= (128, 0, 128), se representa en la Tabla 1 los diferentes valores de la medida de similitud μ entre cada par de colores.

Tabla 1: Ejemplo de medidas de similitud μ entre pares de colores

μ	blanco	negro	gris	azul marino	verde oliva	morado
blanco	1	0	0,5020	0,1344	0,2938	0,2938
negro	0	1	0,4980	0,7102	0,5902	0,5902
gris	0,5020	0,4980	1	0,5902	0,7102	0,7102
azul marino	0,1344	0,7102	0,5902	1	0,4980	0,7102
verde oliva	0,2938	0,5902	0,7102	0,4980	1	0,5902
morado	0,2938	0,5902	0,7102	0,7102	0,5902	1

Que el grado de similitud entre estos colores predefinidos sea bajo significa que los colores no se parecen lo suficiente y que se pueden distinguir fácilmente, por lo que no llevarían a confusión al gestor de los datos y puedan descartarse en los primeros pasos de una investigación.

Quizás el gestor tenga problemas cuando deba determinar cuál es el grado mínimo que debe considerar para comenzar a descartar posibles valores. En el ejemplo se puede ver que el color morado se parece al gris o al azul marino con un grado de similitud de 0,7102, que es bastante alto, pero puede no considerarse como suficiente ya que los colores indicados se pueden distinguir bien de forma objetiva. Dicho valor sí puede usarse como grado comparativo para indicar que un color se parece más a otro según sus grados de similitud.

El uso de la distancia euclídea para este caso determinado se considera adecuado una vez adaptado mediante factores de normalización, si bien se debe considerar aplicarla en otros casos por los inconvenientes que presenta, como no estar acotada, ser sensible a cambio de escalas y que las variables sean independientes estocásticamente (Cuadras *et al.*, 1996). Nótese que existen valores, como el color gris (128, 128, 128), cuyo grado de similitud con respecto a cualquier otro color siempre será superior a 0,5.

Para referenciar el color se debe tener una tabla prefijada, como la MIT'S XConsortum (Walsh, 2006), que transforma el valor discreto que aporta el testigo en una tupla de valores que describen el color en la escala RGB. Si dicha descripción del color se basa en una identificación visual de la tabla, la declaración tendrá un valor mejor definido y los resultados de las medidas de similitud serán más precisos.

2.2 Información parcial: matrículas parciales

80

En los hechos delictivos donde participan vehículos, posiblemente el detalle más útil es que la víctima o los testigos aporten la matrícula que llevaba el vehículo o vehículos implicados. Salvo el caso en que se usen matrículas falsas, puede considerarse el atributo de referencia principal.

Lamentablemente, estas matrículas se aportan de forma parcial o con caracteres cambiados en algunos casos, llegando a confundir esos caracteres indicando alguno que ni siquiera pertenece a la matrícula. Podemos definir el objeto matrícula como aquel que posee como atributos los caracteres que la conforman en la posición en la que aparecen. Así, la matrícula 1234ABC quedaría definida como (matrícula)R(1,2,3,4,A,B,C). De esta forma podemos establecer qué atributo va en cada posición y si un valor pertenece a un objeto matrícula. Abusando de la notación, matrícula puede considerarse tanto objeto como atributo (vehículo)R(matrícula)R(1,2,3,4,A,B,C).

Si se considera el caso de las matrículas españolas actuales con formato NNNNLLL, existen un total de 80 millones de combinaciones posibles. Los números de cada serie van del 0000 al 9999 y las letras usadas son 20, prescindiendo de las vocales y la letra Q. Teniendo en cuenta los casos posibles hasta el final de la serie M, en la actualidad tendríamos 40 millones de combinaciones potencialmente activas.

De esos 40 millones de casos posibles, las combinaciones se reducen en función de la información conocida de la siguiente forma:

- 1 dígito conocido, pero no la posición: 13.756.000 casos posibles.

- 1 dígito conocido y su posición: 4.000.000 casos posibles.
- 2 dígitos conocidos distintos, pero no la posición: 3.896.000 casos posibles.
- 2 dígitos conocidos y su posición: 400.000 casos posibles.
- 3 dígitos conocidos distintos, pero no la posición: 816.000 casos posibles.
- 3 dígitos conocidos y su posición: 40.000 casos posibles.
- 4 dígitos conocidos distintos, pero no la posición: 96.000 casos posibles.
- 4 dígitos conocidos y su posición: 4.000 casos posibles.
- 1 letra conocida, pero no la posición (M o anterior): 7.510.000 casos posibles.
- 1 letra conocida, pero no la posición (N o posterior): 3.900.000 casos posibles.
- 1 letra conocida y su posición 5: 4.000.000 casos posibles.
- 1 letra conocida y su posición 6 o 7: 2.000.000 casos posibles.

El resto de casos posibles de las letras depende de la combinatoria de las mismas y la limitación al estudio sobre la codificación activa el día de los hechos a investigar, siendo los casos posibles dependientes de si las letras son anteriores o posteriores a la letra M y el conocimiento o no de la posición.

De este cálculo de combinaciones posibles se desprende que el investigador obtiene mayor beneficio si la víctima o testigo son capaces de aportar al menos una letra antes que un dígito. De la misma forma, la determinación exacta de la posición del dígito o letra reduce considerablemente el número de casos posibles, pero hace que sea improductiva si la información proporcionada es errónea.

Se va a definir un grado de similitud que permita proponer una matrícula candidata en una investigación a partir de la información aportada por los testigos o por otros medios de prueba. Se ha considerado que la información aportada puede no ser correcta, por lo que en la definición de dicho grado se tiene que tener en cuenta que esa información inválida no condicione de forma grave el resultado.

Para comparar cadenas de caracteres, como es el caso de las matrículas, se podría partir de la distancia de Levenshtein (Levenshtein, 1965), que establece el número de cambios necesarios para que una palabra se transforme en otra mediante modificaciones simples de cambio de un carácter, inserción o eliminación del mismo. Para el caso de estudio planteado, al poder ser la información parcial, no podría aplicarse la misma, debiendo considerarse un algoritmo que sea capaz de medir esta particularidad adecuadamente.

Se define $v(\text{condición, matrícula}) = 1$ si la matrícula cumple la condición propuesta y 0 en caso contrario. Cuando la condición es compuesta, v toma el valor de la media aritmética de considerar todas las condiciones como individuales. De esta forma, el valor de v pertenece siempre al intervalo $[0,1]$.

La condición se puede expresar de 2 formas. Si se sabe la posición, se representa con el número o la letra en esa posición, y el resto de caracteres desconocidos se representan con un asterisco *. Cuando no se sabe la posición, se denota como X_e para que indicar que el carácter X pertenece a la matrícula. Por ejemplo, para indicar la condición de que la letra M ocupa la primera posición de las letras de una matrícula de formato NNNNLLL (N = número, L = letra), se denotaría $****M**$.

Con el fin de que la información errónea no afecte a los resultados considerablemente, cuando existe más de una condición en la propuesta, se realiza la media aritmética de las diferentes condiciones individuales. Esto implica que, si la condición es verdadera en su totalidad, el valor de v no se ve afectado, pero si las condiciones propuestas contienen alguna condición individual errónea, se mantienen las condiciones individuales verdaderas y el valor de v global, aunque se vea disminuido, no

resulte perjudicado de la misma forma que si se valorara la condición completa.

Debe considerarse también que las condiciones pueden ser compatibles o no. Si 2 condiciones no pueden darse a la vez, se da por supuesto que una de las 2 no es correcta, pero no se puede determinar cuál de ellas lo es. Estas condiciones no compatibles afectan al valor de v significativamente ya que impiden que pueda conseguirse alcanzar un valor de certidumbre máximo ($v = 1$).

Ejemplo: Para el caso en que la persona en su declaración diga “la matrícula portaba la letra M o N en la primera posición de las letras”, porque no es capaz de afirmar ninguna de las 2 opciones, pero crea que una de ellas es posible, $v(\text{condición}, \text{matrícula}) = \frac{v(****M**, \text{matrícula}) + v(****N**, \text{matrícula})}{2}$. El valor será 0 cuando la matrícula no porte la M o la N en la primera posición o 0.5 cuando la matrícula cumpla una de las condiciones indicadas. Al ser un conjunto de condiciones no compatibles, ya que esta es excluyente y nunca puede darse el caso de que en la posición de la primera letra puedan aparecer 2 valores diferentes a la vez, v nunca podrá ser 1.

Dada un conjunto de n declaraciones d_i con $i=1, \dots, n$ (la información aportada que se considera un conjunto de condiciones), se define el grado de similitud $\frac{\sum_{i=1}^n v(d_i, \text{matrícula})}{n}$.

Ejemplo: Para el caso de un vehículo que porta la matrícula 1234BCD, se han recibido 4 declaraciones de testigos (d_1 - d_4):

d_1 : “La matrícula empezaba por 1 y acababa en D”.

d_2 : “La matrícula contenía un 3 y un 4, pero no recuerdo las posiciones”.

d_3 : “Las 2 primeras letras eran BC o BG”.

d_4 : “La matrícula empezaba por 12”.

Se observa inicialmente que al menos 2 testigos afirman que la primera posición es un 1, por lo cual las matrículas que cumplan esa condición tendrán mayor valoración como candidatas.

$$v(d_1, \text{matrícula}) = \frac{v(1*****,\text{matrícula}) + v(*****D,\text{matrícula})}{2}$$

$$v(d_2, \text{matrícula}) = \frac{v(3E,\text{matrícula}) + v(4E,\text{matrícula})}{2}$$

$$v(d_3, \text{matrícula}) = \frac{v(****BC*,\text{matrícula}) + v(****BG*,\text{matrícula})}{2} \\ = \frac{2*v(****B**,\text{matrícula}) + v(*****C*,\text{matrícula}) + v(*****G*,\text{matrícula})}{4}$$

$$v(d_4, \text{matrícula}) = v(12*****,\text{matrícula}) \\ = \frac{v(1*****,\text{matrícula}) + v(*2*****,\text{matrícula})}{2}$$

Esto implica que $\theta(\text{matrícula}) = \frac{1}{4}v(1*****,\text{matrícula}) + \frac{1}{8}v(*2*****,\text{matrícula}) + \frac{1}{8}v(3E,\text{matrícula}) + \frac{1}{8}v(4E,\text{matrícula}) + \frac{1}{8}v(****B**,\text{matrícula}) + \frac{1}{16}v(*****C*,\text{matrícula}) + \frac{1}{16}v(*****G*,\text{matrícula}) + \frac{1}{8}v(*****D,\text{matrícula})$.

Al realizar los cálculos, los resultados indican 4 matrículas cuyo valor de θ alcanza el máximo posible, 0.9375. Estas matrículas son 1234BCD, 1234BGD, 1243BCD y 1243BGD, lo cual reduce ampliamente la lista de candidatos. Con un valor de $\theta = 0.875$ hay una serie de candidatos que coinciden en la premisa de la incertidumbre del valor en la posición 6 (C o G) aportada por el testigo en d_3 . Esa duda planteada por el testigo sobre el valor refuerza la idea de que no puede afirmarse nada sobre su certidumbre, y el algoritmo plantea como una opción posible, con alta probabilidad, los casos en los que la letra en la posición 6 no coincida con alguna de las señaladas por el testigo en d_3 .

La aportación de un dato erróneo, por regla general, debe afectar considerablemente los cálculos y, por ello, el orden de los candidatos propuestos en función de su grado de similitud θ . El algoritmo planteado intenta minimizar esos errores en la medida de lo posible, por lo que la validez del resultado dependerá de cuantos errores tengan las declaraciones y de la relevancia que tengan estas en su conjunto.

Ejemplo: Para el caso del ejemplo anterior se va a cambiar la declaración del testigo en d_4 , el cual afirma no solo que las 2 primeras posiciones son 12, sino que añade que el dígito en la tercera posición es 8. Esta declaración lleva a contradicción con la aportación del testigo en d_2 , ya que no es compatible que exista una matrícula que empiece por 128 y que contenga un 3 y un 4.

La aportación del testigo en d_4 quedará de la siguiente forma:

$$v(d_4, \text{matrícula}) = v(128****, \text{matrícula}) =$$

$$\frac{v(1*****, \text{matrícula}) + v(*2*****, \text{matrícula}) + v(**8****, \text{matrícula})}{3}$$

Este cambio afecta a la fórmula general planteada para el grado de similitud θ , la cual quedaría de la siguiente forma:

$$\theta(\text{matrícula}) = \frac{5}{24}v(1*****, \text{matrícula}) + \frac{1}{12}v(*2*****, \text{matrícula}) + \frac{1}{12}v(**8****, \text{matrícula}) + \frac{1}{8}v(3e, \text{matrícula}) + \frac{1}{8}v(4e, \text{matrícula}) + \frac{1}{8}v(****B**, \text{matrícula}) + \frac{1}{16}v(*****C*, \text{matrícula}) + \frac{1}{16}v(*****G*, \text{matrícula}) + \frac{1}{8}v(*****D, \text{matrícula}).$$

Tras realizar los nuevos cálculos, las opciones planteadas varían. A los 4 candidatos propuestos en el caso anterior, se añaden las matrículas 1384BCD, 1384BGD, 1483BCD y 1483BGD. Para estas propuestas, el grado de similitud θ ha bajado de 0.9375 a 0,8533, que sigue siendo alto y aceptable.

Este descenso en el valor de θ refleja que la fiabilidad de las declaraciones del primer ejemplo puede considerarse objetivamente mejor que la de las del segundo ejemplo, reflejando el hecho de que, no por tener un mayor número de datos proporcionados por los testigos, los resultados sean necesariamente de mayor utilidad para los investigadores.

Existe una variante que se puede considerar en el caso de que, tal como se ha reflejado en el segundo ejemplo, las declaraciones recibidas puedan no ser totalmente fiables. Cuando un testigo indica en su declaración términos como “creo”, “puede que”, o términos similares, refleja que existen dudas sobre si lo

comunicado es cierto. El investigador en este caso puede añadir un grado entre 0 y 1 a cada valoración que podría denominarse grado de veracidad. Este grado sería subjetivo y, de esa valoración subjetiva del gestor, los datos obtenidos llevarían a una investigación orientada hacia resultados posiblemente diferentes.

Para los casos de matrículas extranjeras, al ser la combinatoria diferente, no se puede proporcionar un cálculo de casos posibles adecuado sin saber el país de origen. Existen casos, como las matrículas italianas y francesas, que comparten la misma codificación LLNNNLL, en los que no basta simplemente con conocer los candidatos y sus grados de similitud, sino que el análisis debería duplicarse y ampliarse a estos 2 países en caso de no conocerse exactamente el país de procedencia.

Cabe recalcar que el uso de placas de matrícula falsas también hace inviable el análisis que se puede iniciar a partir de los datos obtenidos, aunque sí se puede llegar a la conclusión de confirmar la falsedad de dicha codificación.

2.3 Información errónea: referencias

Cuando se realiza una reseña de una persona, se anota un valor de referencia del documento que porta o que, usando una mala praxis, manifiesta verbalmente el reseñado. Para el caso de españoles o residentes en España, existe una serie de referencia clara que es el “número de DNI” o el “número de NIE”, abusando de la notación “número” ya que contiene caracteres que no son dígitos. Esta referencia cuenta con una letra de control que permite detectar, en la mayoría de los casos, cuándo se ha introducido incorrectamente dicha referencia, permitiendo corregirla si se detecta el error antes de validarla.

El objeto “número de DNI” o “número de NIE” está formado por los atributos ordenados por su posición, así el “número de DNI” (en adelante, DNI para acortar notación) 12345678Z vendría denotado como (DNI)R(1,2,3,4,5,6,7,8,Z).

En algunas partes de servicio, al ser realizados en la vía pública, no existe la verificación de dichas letras de control de

forma automática y se pueden dar errores de anotación. Estos errores se producen habitualmente por el intercambio de posiciones entre un carácter y otro adyacente, por el cambio de un carácter por otro que no es el correcto o por la omisión de uno de los caracteres.

Una vez detectado que existe un error en la anotación del número de registro, la comprobación más sencilla de hacer es intercambiar las posiciones de los diferentes pares de dígitos consecutivos. Esto conlleva hacer simplemente 7 comprobaciones en los casos de los DNI y 6 para los NIE.

La omisión de un carácter, que no sea la letra de control, conlleva mayor número de comprobaciones ya que se ha de probar cada uno de los caracteres posibles en cada posición (0 a 9 más X, Y, Z en la primera posición por si se trata de un NIE). Si el primer carácter es un dígito habría que comprobar 83 casos y, si fuera una letra y por lo tanto un NIE, se entendería que hay un dígito omitido en las posiciones 2 a 8 y por tanto las comprobaciones serían solamente 70. Encontrar el dato correcto es una tarea ardua ya que, al combinar los posibles caracteres omitidos y comprobar que cumplen el algoritmo de la letra de control, proporcionará varios candidatos que deberán ser verificados uno a uno.

Descartados estos casos, para poder comprobar el error detectado sin saber cuál es el origen del fallo en la anotación del registro de identidad, es necesario establecer una medida de similitud entre los diferentes caracteres usados. Dado que estos caracteres en los DNI y NIE son en su mayoría dígitos, se va a enfocar la primera parte del estudio en establecer los grados de similitud entre estos.

Para ello, usando como base de datos imágenes reales obtenidas de documentos, se han extraído diferentes muestras de diferentes dígitos, localizadas en las zonas de lectura mecánica de documentos de viaje. Estas muestras se han normalizado para un tamaño estándar de 10 x 10 píxeles en una imagen en blanco y negro. Se pueden obtener muestras diferentes cambiando el origen de las imágenes y el tamaño de las mismas en píxeles. Al variar las muestras y la morfología de los caracteres, los grados de similitud que se van a definir pueden variar de los propuestos.

Definición: El valor numérico de un píxel en las coordenadas i, j en una imagen c en formato blanco y negro se define como:

$$v_c(i,j) = \begin{cases} 1 & \text{si el píxel } (i,j) \text{ es negro} \\ 0 & \text{en otro caso} \end{cases}$$

Definición: Dado C un conjunto de imágenes con la misma dimensión en píxeles, el valor común del conjunto C para las coordenadas i, j se define como:

$$v_c(i,j) = \begin{cases} 1 & \text{si } \sum_{c \in C} v_c(i,j) \geq \frac{|C|}{2} \\ 0 & \text{en otro caso} \end{cases}$$

Definición: Dadas c_1, c_2 imágenes de $a * b$ píxeles, siendo a la anchura y b la altura, pertenecientes al conjunto C , se define el grado de similitud entre c_1 y c_2 como:

$$\phi(c_1, c_2) = \frac{\sum_{i=1}^a \sum_{j=1}^b \{v_{c_1}(i,j) = v_{c_2}(i,j)\}}{a * b}$$

Esta función mide la cantidad de coincidencias de píxeles blancos y negros entre dos caracteres a partir de sus imágenes tomadas, es decir, cómo se parecen. Para el caso de imágenes con diferente anchura o altura en píxeles, no es posible calcular el grado de similitud, motivo por el cual el conjunto de imágenes debe ser homogéneo en cuanto a estos parámetros se refiere.

A pesar de no ser una relación de similitud por no cumplir la transitividad, esta función permite establecer, fijando un carácter de inicio, un orden numérico entre los caracteres, usando como referencia el grado ϕ .

Para cada conjunto de imágenes del mismo carácter, se puede obtener el valor común de dicho conjunto, que será el que se use como referencia para hacer las comparaciones usando el grado ϕ .

A modo de ejemplo, se muestran en la Tabla 2 los valores de ϕ para los dígitos del 0 al 9 usando esta metodología con las imágenes de 10x10 píxeles tomadas de una muestra.

Tabla 2: Valores de ϕ para los dígitos del 0 al 9

ϕ	0	1	2	3	4	5	6	7	8	9
0	1	0.59	0.58	0.54	0.49	0.64	0.65	0.43	0.70	0.64
1	0.59	1	0.45	0.61	0.54	0.65	0.50	0.46	0.59	0.59
2	0.58	0.45	1	0.60	0.43	0.52	0.45	0.77	0.60	0.62
3	0.54	0.61	0.60	1	0.47	0.76	0.59	0.63	0.64	0.42
4	0.49	0.54	0.43	0.47	1	0.51	0.64	0.40	0.59	0.55
5	0.64	0.65	0.52	0.76	0.51	1	0.63	0.55	0.64	0.56
6	0.65	0.50	0.45	0.59	0.64	0.63	1	0.36	0.75	0.41
7	0.43	0.46	0.77	0.63	0.40	0.55	0.36	1	0.51	0.53
8	0.70	0.59	0.60	0.64	0.59	0.64	0.75	0.51	1	0.54
9	0.64	0.59	0.62	0.42	0.55	0.56	0.41	0.53	0.54	1

Usando los grados de similitud entre dígitos indicados en la tabla y dada una referencia de documento con un error de transcripción, podemos hacer una valoración orientada de los posibles candidatos a ser el dato correcto.

El orden de comprobación sería la sustitución en función del valor del grado ϕ , de mayor a menor valor, siempre y cuando la referencia contenga uno de los dígitos. Así, se partiría del cambio del 2 por el 7, o viceversa, por poseer el mayor grado de similitud 0.77. Seguiría el 3 por el 5 con un grado de similitud de 0.76, el 6 por el 8 con grado 0.75, el 0 por el 8 con grado 0.70, etc.

De forma análoga, si se considera que el carácter erróneo cambiado es la letra de control, se puede comprobar también el grado de similitud entre la letra anotada y la correcta según el algoritmo de verificación de la letra de control.

Ejemplo: Se ha encontrado en un parte de servicio anotado el DNI 32052158T, el cual no verifica el algoritmo de la letra de control. El formato es correcto, por lo que no es necesario hacer las diferentes pruebas incluyendo los diferentes candidatos que se darían si faltara un carácter.

La comprobación de las 7 permutaciones posibles no confirma la letra de control, por lo que, por descarte, el caso posible

que queda es el de sustitución de la letra de control o uno de los dígitos por uno erróneo.

Según el algoritmo de comprobación de la letra de control, a la serie de dígitos le corresponde el valor W, el cual tiene un grado de similitud de 0.39 con respecto a la letra T, bastante bajo, por lo que no parece la opción correcta.

Los candidatos se van a obtener cambiando uno a uno los dígitos existentes por aquellos que tengan mayor grado de similitud. El primero de los cambios es el 2 por el 7, no cumpliendo ninguno de los 2 candidatos el algoritmo de verificación. Los siguientes 2 candidatos se producen al cambiar el 3 por el 5 y viceversa, resultando que tampoco cumplen el algoritmo de verificación.

El quinto candidato se genera al cambiar el 8 por el 6, siendo su grado 0.75. Este candidato sí es verificado por el algoritmo, por lo que solamente ha sido necesario comprobar 5 de los 72 candidatos posibles para hallar una posible solución.

La medida de similitud θ del apartado anterior se puede generalizar, ya que en ella no se va a usar la terminología de carácter correcto o incorrecto, sino los términos carácter propuesto y carácter candidato. Una matrícula es en realidad una referencia a la serie de un vehículo, por lo que se trataría de un caso de estudio análogo.

Haciendo un abuso de la notación, se denotarán con j las posiciones de los caracteres contenidos en una declaración, obviando aquellos no presentes y la parte de las declaraciones donde se indica que un carácter pertenece a una referencia, quedando de-

notado por $\theta^*(\text{referencia}) = \frac{\sum_{i=1}^n v(d_{ij}, \text{referencia}) * \varphi(j, r_j)}{n}$, donde r_j es el carácter en la posición j de la referencia.

Esta medida de similitud permite evaluar los datos de la referencia sin necesidad de conocer si estos son correctos o incorrectos. Se debe usar con la precaución de saber en todo momento que pueden existir errores y que no siempre se puede verificar la fiabilidad de la fuente.

3 Resultados

Con la metodología descrita en el apartado anterior, se van a combinar diferentes medidas en casos teóricos para ver su comportamiento. Dado que los datos de color son valores difusos y no se desea tergiversar los resultados usando datos subjetivos, se han tomado las muestras de color a partir de fotografías. Las matrículas que aparecen en los resultados, aunque pertenezcan a vehículos reales, han sido generadas aleatoriamente para no desvirtuar los experimentos realizados en los ejemplos.

91

3.1 Información difusa: similitud en colores

Se ha tomado una serie de imágenes de vehículos, recortando trozos donde aparece en mayor medida la carrocería, para determinar el color de los mismos. Dichas imágenes están compuestas por un conjunto de píxeles, con tamaños no fijados previamente, en la escala RGB. A través de una aplicación creada al efecto, se extraen los diferentes valores de los colores de cada píxel.

Debido a que estas imágenes contienen *ruido*, píxeles con un color notablemente diferente al resto de su entorno, se ha de proceder a una *limpieza* de los valores anómalos. Para ello, se ha usado el algoritmo del método DBSCAN (Ester *et al.*, 1996) con valor de distancia máxima de 2, grupos mínimos formados por 1 objeto y midiendo las distancias entre tuplas con la distancia euclídea normalizada para 3 valores. Se ha tomado el conjunto o *clúster* con mayor número de componentes, descartando los píxeles fuera de dicho grupo. Para definir el color tomado de la fotografía, se usa la media de cada uno de los componentes RGB de los píxeles pertenecientes al grupo.

Ejemplo: Se ha tomado una fotografía de un coche blanco, a la sombra para que dicho color aparezca atenuado. De dicha fotografía se ha recortado un trozo lateral de una puerta, obteniendo una imagen de 378 x 359 píxeles. Los valores iniciales se indican en la Tabla 3.

Tabla 3: Valores iniciales en la toma de una fotografía de un vehículo blanco

	Rojo	Verde	Azul
Mínimo	64	71	75
Media	183	182	186
Máximo	212	221	230

Se observa claramente la presencia de ruido, por lo que, tras ser analizados los datos con DBSCAN y tomar el conjunto principal, este consta de 537 valores, siendo estos los que figuran en la Tabla 4.

Tabla 4: Valores de la fotografía de un vehículo blanco tras aplicar DBSCAN

	Rojo	Verde	Azul
Mínimo	161	159	161
Media	188	190	195
Máximo	210	218	227

Teniendo en cuenta que el color blanco *puro* en la escala RGB se corresponde con la tupla (255, 255, 255), se aprecia una gran diferencia con los valores resultantes. Esto es debido a que el color imprimado en el coche ya presenta una diferencia con estos valores, a lo que se añade cómo estos colores se ven afectados por las circunstancias físicas externas, como la luz o la curvatura de la puerta, cuando se les toma una fotografía. Usando la medida de similitud entre colores, $\mu((255, 255, 255), (188, 190, 195)) = 0.7488$, que podemos considerar un valor alto.

Se pueden considerar otros colores predefinidos, aunque estos no tienen por qué tener valores fijos. De esta forma se puede tomar como referencia, pero no de forma exacta, el color gris claro de la tabla MIT'S XConsortum. Según su definición, los valores en la escala RGB se corresponden con la tupla (211, 211, 211), aunque puede haber diferencias si se toma una definición diferente de este color, usando la medida de similitud entre colores, $\mu((211, 211, 211), (188, 190, 195)) = 0.9207$, que es superior a la medida de similitud con el color blanco.

Esto es un ejemplo claro de cómo un color puede ser definido por un testigo, o por la captura de una imagen, de forma errónea

o tergiversada sin intención, por las condiciones en las cuales se apreciaron los hechos y se asimiló un valor de forma subjetiva.

Ejemplo: Tomando una segunda imagen de un coche blanco, al sol, se ha recortado la parte del capó, incluyendo algunas tonalidades de ruido dentro de la misma. El tamaño resultante es de 118 x 104 píxeles. Los valores iniciales figuran en la Tabla 5.

Tabla 5: Valores iniciales de la fotografía de un vehículo blanco

	Rojo	Verde	Azul
Mínimo	155	155	157
Media	235	236	240
Máximo	254	255	255

Igualmente, se observa la presencia de ruido, por lo que, tras ser analizados los datos con DBSCAN y tomar el conjunto principal, este consta de 175 valores, siendo estos los datos que aparecen en la Tabla 6.

Tabla 6: Valores de la fotografía de un vehículo blanco tras aplicar DBSCAN

	Rojo	Verde	Azul
Mínimo	207	208	212
Media	234	234	236
Máximo	254	255	255

Tomando las medidas de similitud con respecto a las definiciones mediante tuplas en escala RGB de los colores blanco y gris claro del ejemplo anterior, tenemos que $\mu((255, 255, 255), (234, 234, 236)) = 0.9202$ y $\mu((211, 211, 211), (234, 234, 236)) = 0.9071$, siendo para este ejemplo el color observado más similar al blanco que al color gris claro.

Teniendo en consideración el resultado de los ejemplos indicados con un color claro de referencia como es el color blanco, puede considerarse que los valores obtenidos y anotados en las bases de datos pueden diferir de forma considerable cuando se trata con colores que presentan una gama variada de valores, como las familias de los colores que pueden ser interpretados como azul, verde, gris o amarillo.

3.2 Información parcial: matrículas parciales

La metodología para obtener una medida de similitud entre matrículas parciales θ quedó definida anteriormente, reflejando cómo se obtendrían candidatos usando una base de datos general muy amplia que recoge cualquier posible combinación de matrículas. En algunas investigaciones, se parte de una base de datos más específica, con datos de vehículos que contienen, entre otros campos, la matrícula y el color.

94

Para trabajar con esta base de datos específica, ampliando la metodología usada para obtener las medidas de similitud θ entre matrículas, se va a definir una nueva medida de similitud θ' que añada el color del vehículo.

Dada un conjunto de n declaraciones d_i con $i = 1, \dots, n$ (la información aportada que se considera un conjunto de condiciones), se define el grado de similitud θ' (matrícula, color) = $\frac{\sum_{i=1}^n v(d_i, \text{matrícula})}{2n} + \frac{\sum_{i=1}^n \mu(d_i, \text{color})}{2n}$. Cada declaración d_i debe tener 2 partes, la dedicada a la matrícula y la dedicada al color.

Esta nueva definición equipara la valoración de los datos matrícula y color en la declaración, otorgándoles el mismo peso de $\frac{1}{2}$ en la ecuación, si bien pueden adaptarse según la valoración de las declaraciones. En caso de que las declaraciones tengan mayor relevancia describiendo la matrícula que en la descripción, puede otorgársele más peso a este y viceversa.

Ejemplo: Tras el análisis de cámaras de seguridad, se comprueba cómo 5 vehículos han sido captados transitando por una vía cercana al lugar donde se ha producido un hecho delictivo. Estos vehículos se corresponden con las matrículas y colores siguientes: <3748FDV, (190, 83, 120)>, <8263HGR, (59, 131, 189)>, <2372DDF, (238, 180, 34)>, <4764GTF, (35, 107, 142)>, <2822HCN, (36, 24, 130)>.

Un testigo afirma, en la declaración d_1 , que la matrícula del vehículo contenía un 6 y una letra G, siendo de color “azul grisáceo”,

identificándolo en la tabla MIT'S XConsortum con los valores (54, 100, 139) correspondiente a “azul acero 4”.

Según la declaración d_1 , $\theta'(\text{matrícula}, \text{color}) = \frac{v(6e, \text{matrícula}) + v(Ge, \text{matrícula})}{4} + \frac{\mu((54, 100, 139), \text{color})}{2}$. Haciendo la comparación con los 5 vehículos, resultaría la Tabla 7.

Tabla 7: Valores de θ' de la declaración d_1 para los 5 vehículos

	Matrícula	Color	θ'
3748FDV	0	0.3433	0.3433
8263HGR	0.5	0.4332	0.9332
2372DDF	0	0.2436	0.2436
4764GTF	0.5	0.4768	0.9768
2822HCN	0	0.4110	0.4110

Tomando en cuenta esta declaración, los vehículos 8263HGR y 4764GTF tienen una medida θ' superior a 0.9 ya que sus medidas parciales son bastante relevantes. Estos serían los principales candidatos si solo se tuviera la declaración d_1 .

Un segundo testigo aporta la declaración d_2 , en la cual indica que los dígitos en las posiciones segunda y tercera son 3 y 6, siendo de color “azul oscuro”, identificándolo en la tabla MIT'S XConsortum con los valores (16, 78, 139) correspondiente a “azul evasor 4”.

Según la declaración d_2 , $\theta'(\text{matrícula}, \text{color}) = \frac{v(*3****, \text{matrícula}) + v(**6****, \text{matrícula})}{4} + \frac{\mu((16, 78, 139), \text{color})}{2}$.

Haciendo la comparación con los 5 vehículos, resultaría la Tabla 8.

Tomando en cuenta esta declaración d_2 , los vehículos 8263HGR y 4764GTF tienen una medida θ' alta, pero no de la misma consideración que la declaración d_1 , ya que es seguro que hay un error en los datos de la matrícula en esta declaración por no ser cumplida por ninguno de los vehículos de la base de datos restringida. El investigador, ante estas declaraciones, puede

Tabla 8: Valores de θ' de la declaración d_2 para los 5 vehículos

	Matrícula	Color	θ'
3748FDV	0	0.3017	0.3017
8263HGR	0.25	0.4042	0.6542
2372DDF	0.25	0.1990	0.4490
4764GTF	0.25	0.4606	0.7106
2822HCN	0	0.4340	0.4340

optar por cambiar los pesos y dar más valor al grado que mide la similitud del color.

Haciendo la valoración con estas 2 declaraciones sobre los vehículos 8263HGR y 4764GTF, se tiene que $\theta'(8263HGR, (59, 131, 189)) = \frac{0.9332 + 0.6542}{2} = 0.7937$ y $\theta'(4764GTF, (35, 107, 142)) = \frac{0.9768 + 0.7106}{2} = 0.8437$.

Con estas 2 únicas declaraciones, conteniendo una de ellas información errónea, no puede afirmarse claramente cuál de los 2 vehículos es el principal candidato, pero sí que se han conseguido descartar los otros 3 vehículos, reduciendo la muestra inicial de candidatos. Una tercera declaración con información fidedigna de la matrícula puede clarificar la elección entre los 2 candidatos. Sin embargo, al ser los colores de los vehículos candidatos muy similares, con una medida de similitud $\mu((59, 131, 189), (35, 107, 142)) = 0.8687$, la declaración sobre el color no sería determinante.

3.3 Información errónea: referencias

Se ha realizado un estudio durante 6 meses, entre enero y junio de 2025, sobre diversa documentación donde figuraban números de DNI o NIE, encontrándose 13 casos donde existían anomalías producidas por errores de anotación de los mismos por parte de los agentes.

De estos casos, se ha podido llegar a la conclusión de que 1 se corresponde con la falta de un carácter, 2 a errores de permutación y 10 a cambio de un carácter por otro incorrecto.

Para resolver estas incidencias, se programó un algoritmo asistente siguiendo las pautas indicadas anteriormente. Este programa asistente actualmente se encuentra en una aplicación programada para sistemas Android, para que pueda ser consultado por determinados agentes en la vía pública ante documentos con mala calidad y por el personal encargado de registrar los datos en la base de datos policial.

Para resolver el caso de la falta de 1 carácter, el algoritmo desarrollado comprobó la verificación de la letra de control de los 83 candidatos posibles. De estos 83 candidatos, tan solo 3 cumplían la verificación, debiendo contrastarse cada uno de ellos con la base de datos y localizándose el correcto por coincidencia de otros datos.

Los 2 casos de errores de permutación ofrecían 7 candidatos cada uno, verificando tan solo 1 de ellos, en cada caso, el algoritmo de verificación de la letra de control.

Para los restantes 10 casos, con la sustitución de un carácter por otro, se procedió a seguir el orden indicado por el algoritmo para obtener candidatos, con base en los grados de similitud y los caracteres presentes en cada referencia. Uno de los casos contemplaba el cambio de una Y por una X en un NIE, con un grado alto de similitud de 0.74, comprobándose su validez usando el tercer candidato por orden de grado de similitud. De los 10 casos, el que más se demoró para encontrar el error usando el algoritmo lo hizo durante la comprobación del noveno candidato.

Retornando al caso de la declaración d_2 en el apartado 3.2, donde la declaración con respecto a la matrícula era errónea al no ser cumplida totalmente por los candidatos de la muestra, la medida usada puede ser modificada para considerar que existe un error en la valoración. Dicha modificación se haría cambiando el valor v del carácter indicado de la matrícula por el valor v del carácter candidato, multiplicado por la medida de similitud entre ambos caracteres, el indicado y el candidato, como se definió anteriormente.

Los datos del color no van a ser modificados, manteniendo los valores obtenidos en el apartado anterior, por lo que de forma análoga se define $\theta^*(\text{matrícula}, \text{color})$.

Para los vehículos candidatos, los resultados de θ^* ofrecen valores diferentes a los de θ' , siempre mayores o iguales por las propiedades matemáticas de la fórmula para calcular el grado, quedando definida para d_2 , $\theta^*(matrícula, color) = \frac{v(*m_2****,matrícula)*\varphi(3,m_2)+v(**m_3****,matrícula)*\varphi(6,m_3)}{2} + \frac{\mu((16,78,139),color)}{4}$. Los valores⁴ calculados figuran en la Tabla 9.

Tabla 9: Valores de θ^* para la declaración d_2 para los 5 vehículos

	Matrícula	Color	θ^*
3748FDV	0.3175	0.3017	0.6182
8263HGR	0.4	0.4042	0.8042
2372DDF	0.34	0.1990	0.5390
4764GTF	0.4075	0.4606	0.8681
2822HCN	0.2725	0.4340	0.7065

Esta nueva valoración usando la medida θ^* en vez de θ para considerar que existe información errónea eleva el grado de similitud a unos valores que permiten considerar la aceptación de la declaración d_2 para la investigación. El inconveniente es que también eleva los valores de los candidatos que otras declaraciones pueden descartar sin tener en cuenta los errores, disminuyendo la diferencia relativa entre los candidatos.

4 Discusión

Se han presentado diversas medidas de similitud para considerar el uso de valores difusos en las bases de datos policiales y en posibles investigaciones. Estas medidas pretenden tratar con datos que no son exactos necesariamente, valorando 3 tipos de datos: los que contienen información difusa, los que contienen información parcial y los que contienen información errónea.

A priori, las bases de datos manejadas por humanos son sensibles a los errores de transcripción de información y a la valoración subjetiva de los datos difusos, por lo cual se pueden encontrar los 3 tipos de datos mencionados en las investigaciones policiales.

Como dato difuso de referencia se ha valorado el color, poniendo el foco en vehículos por su mayor facilidad en la determinación de dicho color. La base de datos de la Dirección General de Tráfico no sirve de referencia en cuanto al color ya que contiene información errónea, siendo el blanco el color que aparece por defecto cuando no hay una referencia anotada. A pesar de proponer la tabla de colores MIT'S XConsortum, esta solo sirve de referencia para acercar un poco más la realidad a lo contenido en las declaraciones de testigos o datos anotados.

En el caso de una persona, anotar el color de piel es una tarea más compleja pues el rango del color de la piel es más limitado, siendo la diferencia entre datos más pequeña dentro de los mismos grupos. Así, una piel blanquecina varía hasta el color negro pasando solo por una escala de tonos con diferencias en tonalidades, obviando la mayoría de valores de la tabla MIT's XConsortum ya que no existen personas con piel de color verde o azul.

El color del pelo es un dato problemático ya que evoluciona a lo largo del tiempo y de la decisión de la persona. Los colores artificiales usados hacen que se deban anotar los diferentes tonos de color de pelo de una persona con una referencia de tiempo asociada. No solo para aquellas personas mayores que se tiñen las canas, sino para aquellos delincuentes que ocultan su color natural por diferentes motivos. Queda excluido el caso en que se usan pelucas para cometer actos delictivos, ya que no es posible controlar este tipo de factor al ser tan impredecible y, en todo caso, entraría en el caso de estudio de la información errónea.

El uso de datos con información parcial puede ser admitido si la base de datos permite la comprobación de sus referencias con esta información. Generalmente se usan asteriscos para denotar información desconocida cuando se realizan consultas a estas bases de datos, pero la respuesta que dan es si coincide o no con lo planteado, pudiendo descartar algunos de los candidatos. Con las medidas propuestas, se proporcionaría un valor a cada dato devuelto, permitiendo discernir entre todas las respuestas cuál puede ser el que más interesa a los investigadores.

Se han estimado, como dato de referencia que usa información parcial, las matrículas de los vehículos. Al ser usadas en bases de

datos de matrículas españolas, la información de respuesta está delimitada por las combinaciones posibles. En caso de que las matrículas a buscar sean extranjeras, se añade la complicación de conocer la estructura de los caracteres que forman dichas matrículas y, fundamentalmente, conocer el país de origen. El acceso a bases de datos extranjeras es más limitado y la casuística es variada en cuanto a combinaciones de caracteres dependiendo del país.

Conociendo la estructura de los caracteres en las matrículas extranjeras, las medidas se pueden adaptar fácilmente. Incluso sin conocer esas estructuras, solo basándose en la cantidad de caracteres, se puede obtener una medida de similitud representativa, aunque sería necesaria la base de datos para aplicar esas medidas con las declaraciones propuestas.

El uso de información errónea es uno de los mayores problemas en las bases de datos policiales, y más concretamente desconocer que el valor anotado es un dato erróneo. Existen errores documentados en los cuales se anotaron en los registros identificaciones de personas con documentación falsa, los cuales han llevado incluso a proporcionar identidades legales a personas que han conseguido una nueva vida *limpiando* su historial delictivo. Este error es difícil de detectar cuando proviene de documentación y solo es detectable con conocimientos suficientes en documentoscopia o mediante la implementación de algoritmos que verifiquen posibles caracteres de control. Aunque existe una alternativa para controlar la anotación de una persona en la base de datos mediante la identificación dactilar, la cual representa de forma única a esta persona, si esa huella viene relacionada con una identidad falsa, la base de datos seguirá conteniendo información errónea. Esta casuística se da principalmente en extranjeros, ya que las personas registradas con DNI sí tendrían una entrada inicial verificada. El proyecto BUDA (Mélida Lledó, 2023) contempla el uso de las huellas dactilares para la identificación de personas basados en estos registros.

Los errores que sí entran dentro de los valores detectables son aquellos que se producen por una introducción incorrecta de datos o por un error en la fuente de origen. Se distingue entre dato incorrecto y dato falso basándose en la no voluntariedad de introducir o aportar el dato de forma errónea.

Se ha hecho referencia a cómo las declaraciones de testigos pueden contener información incorrecta y a cómo se pueden adaptar las medidas de similitud para tratar esa información. Dichas medidas adaptadas conservan la información parcial correcta y controlan que la información parcial incorrecta no tergiverse los resultados de las conclusiones al analizar los datos.

5 Conclusiones y trabajos futuros

101

Las medidas de similitud descritas permiten la combinación entre ellas para el análisis de los 3 tipos de datos con información difusa indicados, sin necesidad de conocer las diferentes teorías matemáticas en las que se basa. Esta plasticidad hace que las investigaciones sean más moldeables para un investigador experimentado, adaptando las medidas en función de los datos que haya que analizar.

Estas propuestas son una base para un desarrollo más amplio. Pueden contemplarse como la raíz para la aplicación de otras metodologías, como sistemas de inteligencia artificial, algoritmos evolutivos o minería de datos, que trabajen con datos difusos.

Entre las aplicaciones prácticas desarrolladas con la metodología presentada hay proyectos bastante avanzados para la detección de errores en las zonas de lectura mecánica de documentos de viaje, superando ampliamente el índice de detección de aquellas aplicaciones que solamente detectan los errores en los dígitos de control establecidos por la Organización de Aviación Civil Internacional. Lamentablemente, los algoritmos para calcular esos dígitos de control son conocidos tanto por los agentes que controlan dichos documentos como por los falsificadores.

Las limitaciones de la aplicación de esta metodología se fundamentan en el diseño de las propias bases de datos y en el manejo por parte de personal que puede carecer de la información adecuada para ello. La información que pueda ser considerada como difusa por tener una valoración subjetiva siempre está

expuesta a errores. El diseño de los campos en las bases de datos y la forma de introducirlos es bastante complejo, siendo la formación la mejor manera de limitar los posibles errores. Se han dado varios casos de intervenciones con vehículos en los que la base de datos de la Dirección General de Tráfico aportaba el color blanco, el que muestra por defecto, que no coincidía con el color del vehículo presente. Reformar una base de datos que contiene datos erróneos es una tarea compleja y a veces imposible de realizar.

Como proyecto futuro en actual desarrollo, se están habilitando herramientas que permitan comprobar datos de hospederías para detectar anomalías en las anotaciones realizadas por las personas encargadas de la recopilación de datos. La inconsistencia de los valores introducidos puede ser detectada mediante medidas de similitud, indicando al encargado del análisis de los datos dicha anomalía y dando un grado de similitud que puede incluso llegar a discernir si se trata de un error por anotación incorrecta, por error en el tipo de dato anotado o incluso si se trata de un dato proporcionado por un documento falso.

Un proyecto futuro en estudio analizaría las descripciones proporcionadas por testigos de personas y vehículos, para proporcionar candidatos que se ajusten de forma más precisa a lo indicado por las fuentes. En este artículo se han combinado a lo sumo 2 tipos de valores, como matrícula y color, cuando las medidas de similitud pueden ampliarse para medir la consistencia de numerosos valores. La dificultad de este estudio radica en la otorgación de pesos para ponderar las medidas de similitud, ya que cada distribución de pesos en la ponderación puede otorgar resultados completamente diferentes si no se aplica adecuadamente.

6 Agradecimientos

Los autores agradecen a los miembros de ADOFOR su colaboración para llevar a cabo este estudio mediante sus valoraciones y sugerencias.

7 Referencias bibliográficas

- Armstrong, W. W. (1974). Dependency structures of data base relationships. En J. W. Klimbie y K. L. Koffeman (Eds.), *International Federation for Information Processing* (pp. 580-583). North-Holland Publishing Company.
- Codd, E. F. (1990). *The relational model for database management: Versión 2*. Addison-Wesley Longman Publishing Co. Inc.
- Cordero, P., Enciso, M., Mora, A., De Guzmán, I. P. y Rodríguez-Jiménez, J. M. (2011). Specification and inference of fuzzy attributes. En *Lectures Notes in Computer Science*, 6692, 492-499). Springer. https://doi.org/10.1007/978-3-642-21498-1_52
- Cuadras, C. M., Arenas, C. y Fortiana, J. (1996). Some computational aspects of a distance-based model for prediction. *Communications in Statistics – Simulation and Computation*, 25(2), 395-410. <https://doi.org/10.1080/03610919608813336>
- Ester, M., Kriegel, H. P., Sander, J., y Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. En *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD-96*, 226-231. AAAI
- Ganter, B., Stumme, G. y Wille, R. (2005). Formal Concept Analysis. Foundations and Applications. *Berlin Heidelberg. Springer*. <https://doi.org/10.1007/11528784>
- Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8), 707-710.
- Mélida Lledó, P. L. (2023). Retos de la investigación científico-forense. *Ciencia Policial*, 179, 11-43.
- Rodríguez-Jiménez, J. M., Cordero, P., Enciso, M. y Mora, A. (2014). Negative attributes and implications in formal concept analysis. *Procedia Computer Science*, 31, 1054-1063. <https://doi.org/10.1016/j.procs.2014.05.373>

Walsh, J. K. (2006). *RGB to color name mapping (triplet and hex)*. Recuperado el 25 de junio de 2025 de <https://web.njit.edu/~walsh/rgb.html>

Zadeh, L. A (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)

Zadeh, L. A. (1971). Similarity relations and fuzzy orderings. *Information Sciences*, 3(2), 177-200. [https://doi.org/10.1016/0020-0255\(71\)90012-0](https://doi.org/10.1016/0020-0255(71)90012-0)

Zadeh, L. A. (1992). Knowledge representation in fuzzy logic. En *An introduction to fuzzy logic applications in intelligent systems*. Springer.

Zadeh, L. A. (1996). Fuzzy logic=computing with words. *IEEE transactions on fuzzy systems*, 4(2), 103-111. <https://doi.org/10.1109/91.485416>