Universidades de Burgos, León y Valladolid

Máster Universitario en Inteligencia de Negocio y Big Data en Entornos seguros

Trabajo fin de Máster

UNIVERSIDAD DE BURGOS

universidad de león

Universidad de Valladolid

Análisis de la consistencia de bases de datos de diferencias de color

Curso 2022-2023

Daniel Arranz Ortega

Tutores:

Pedro Latorre Carmona
Samuel Morillas Gómez
Rafael Huertas Roa

Índice

1. RESUMEN

La medición y especificación precisa del color, así como la medida de las diferencias de color entre dos pares de muestras son facetas de suma importancia en diversos ámbitos como el sector industrial, textil, agrícola o en el ámbito de la salud. Esta relevancia se acentúa especialmente en aquellos campos donde el color no solo es un atributo, sino que agrega un valor significativo al producto final. En medicina, la medición del color puede ser esencial para detectar cambios sutiles en la piel o los tejidos, lo que podría indicar problemas de salud. Por ejemplo, en dermatología, la evaluación del color de lunares o lesiones cutáneas puede ayudar a identificar posibles signos de cáncer de piel.

En este trabajo, se ha realizado un análisis exhaustivo de varias bases de datos de diferencias de color, utilizando técnicas de análisis de datos mediante lógica difusa. El objetivo final de este análisis ha sido identificar parejas de colores inconsistentes en comparación con otras parejas, con el fin de mejorar la calidad y la consistencia de las bases de datos iniciales. Para lograrlo, se ha implementado una metodología que se basa en la detección de discrepancias entre las diferencias de color visualmente percibidas y las diferencias de color calculadas en comparación con el resto de datos. Se ha explorado una variedad de umbrales, evaluando cómo los diferentes niveles de tolerancia afectan a la detección de inconsistencias.

El resultado de este trabajo es la identificación, análisis y eliminación de parejas de colores que se consideran inconsistentes en distintas bases de datos utilizando algoritmos desarrollados en Python. Esta depuración de datos contribuye significativamente a mejorar la calidad de las bases de datos de diferencias de color, lo que a su vez tiene un impacto positivo en diversas aplicaciones en las que la precisión del color es fundamental.

Estas bases de datos desempeñan un papel fundamental en el desarrollo y evaluación de nuevas fórmulas de diferencia de color. Su objetivo es lograr que los resultados de estas fórmulas se ajusten de manera más precisa a la percepción visual de la diferencia de color por parte del sistema visual humano.

2. ABSTRACT

The precise measurement and specification of color, as well as the measurement of color differences between two pairs of samples, are extremely important in various fields such as industry, textiles, agriculture and healthcare. This relevance is especially accentuated in those fields where color is not only an attribute, but adds significant value to the final product. In medicine, color measurement can be essential for detecting subtle changes in skin or tissue, which could indicate health problems. For example, in dermatology, assessing the color of moles or skin lesions can help identify possible signs of skin cancer.

In this work, a comprehensive analysis of several color difference databases has been performed using fuzzy logic data analysis techniques. The ultimate goal of this analysis has been to identify inconsistent color pairs compared to other pairs, in order to improve the quality and reliability of the initial databases. To achieve this, a methodology has been implemented that is based on the detection of discrepancies between visually perceived color differences and color differences calculated from color coordinates. A variety of thresholds have been explored, evaluating how different tolerance levels affect the detection of inconsistencies.

The result of this work is the identification, analysis and elimination of color pairs that are considered inconsistent in different databases using algorithms developed in Python. This data cleaning contributes significantly to improving the quality of color difference databases, which in turn has a positive impact on various applications where color accuracy is critical.

These databases play a key role in the development and evaluation of new color difference formulas. Their goal is to match the results of these formulas more accurately to the visual perception of color difference by the human visual system.

3. INTRODUCCIÓN

El color es algo que, en cierta medida, percibimos y entendemos de manera subjetiva. Diferentes personas pueden describir el color de un mismo objeto de manera diferente. Con el fin de conseguir expresar el color de manera numérica y objetiva, se puede definir según su tonalidad (el tipo de color), luminosidad (cuán brillante es) y saturación (cuán puro es). Estas mediciones nos permiten hablar de colores de una manera precisa y sin ambigüedades, siendo uno de los más utilizados, pero no el único.

La Comisión Internacional de la Iluminación (CIE por sus siglas en francés, Commission Internationale de l'Éclairage) es una organización científica y técnica reconocida a nivel internacional que se dedica al estudio y la estandarización de la luz, el color y la iluminación. La CIE tiene como objetivo principal proporcionar normas y recomendaciones relacionadas con la percepción visual, la medición de la luz y el color, así como la iluminación adecuada en una amplia gama de aplicaciones.

Uno de los sistemas más destacados desarrollados por la CIE es el sistema de color CIELAB, también conocido como el espacio de color Lab o simplemente Lab [1, 2]. Este sistema fue creado para que la distancia euclídea entre dos puntos (colores) tuviera una buena correspondencia con la diferencia de color percibida entre esos dos colores, lo que lo convierte en una herramienta ampliamente utilizada en diversas disciplinas.

Espacio de color CIELAB:

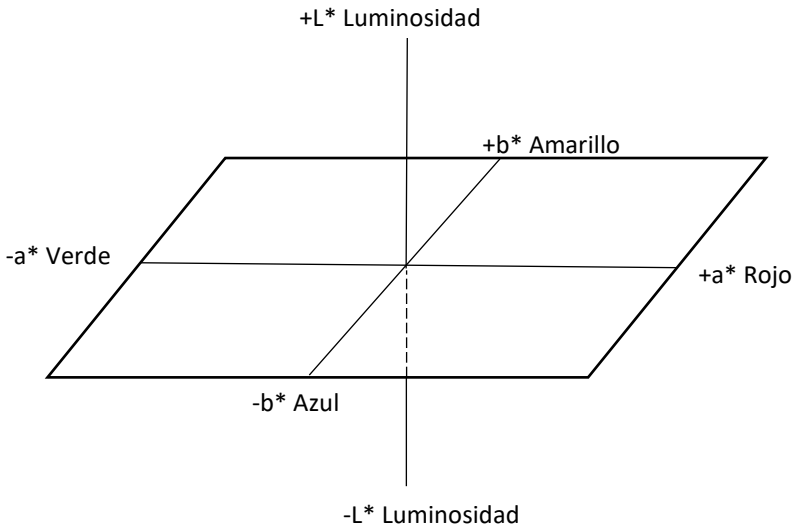

Figura 1- Espacio de color CIELAB

La Figura 1 muestra la forma como se representa el espacio de color CIELAB. Es un espacio tridimensional, en el que cada uno de los ejes representa un parámetro de color utilizado en colorimetría. Esta representación en tres dimensiones ha sido creada con la finalidad de que las variaciones de color en este espacio sean lo más uniformes posible en términos de percepción visual, manteniendo una igualdad perceptual en todo el espacio [3].

En el espacio CIELAB, las coordenadas cartesianas corresponden a:

- **Luminosidad (L*):** Este componente varía desde 0 (negro absoluto) hasta 100 (blanco absoluto) y representa la intensidad luminosa o el brillo del color. Valores más altos de L indican colores más claros, mientras que valores más bajos corresponden a colores más oscuros.
- **Componente a*:** La coordenada "a" se extiende desde valores negativos que representan tonos de verde hasta valores positivos que indican tonos de rojo. Valores cercanos a cero se asocian con colores neutros o grises.
- **Componente b*:** La coordenada "b" abarca desde valores negativos que sugieren tonos azules hasta valores positivos que indican tonos amarillos. Valores cercanos a cero se relacionan con colores neutros o grises.

Cada punto en el espacio tridimensional representa un color específico. Las coordenadas L*, a* y b* se utilizan para definir la ubicación precisa de un color. Esto permite una descripción objetiva y cuantitativa del color, lo que es esencial en una variedad de aplicaciones, desde el diseño de productos y la calidad de la impresión hasta la caracterización de alimentos y la industria textil.

En nuestro trabajo, hemos empleado un conjunto de bases de datos, obtenidos de una serie de experimentos. En concreto, dichas bases están formadas por datos que son pares de colore, cada uno de los cuales incluye información detallada sobre sus coordenadas en el espacio Lab (CIELAB) y otros parámetros significativos.

Las bases de datos de uso privado son:

- Pointer: Desarrollada por el investigador M. Pointer en la Universidad de Brandfor.
- MMB: Morley-Munn-Billmeyer [4].
- Qiao: Desarrollada por Yue Qiao en el Rochester Institute of Technology.
- BIGC: Beijing Institute of Graphic Communication [5].
- lcamDIN99WDC: Desarrollada por el investigador Michal Vik en el Departamento de ingeniería de Materiales en Technical University of Liberec.
- RIT_DuPont_Individual: Desarrollada en el Rochester Institute of Technology.

- RIT-COM_V5: Desarrollada en el Rochester Institute of Technology.
- WangHan: Desarrollada por el investigador Wang Han en la Colour, Imaging and Design Research Centre, University of Leeds.
- NCSU: Desarrollada por el investigador R. Shamey en la North Carolina State University.

En conjunto, estas bases de datos proporcionan una base sólida y diversa para nuestro análisis de las diferencias de color percibidas, permitiéndonos explorar y comprender mejor cómo percibe el ser humano el color.

En las bases de datos utilizadas en este estudio, encontramos una valiosa colección de datos experimentales. Estos datos experimentales están compuestos por pares de colores, cada uno de los cuales incluye información detallada sobre sus coordenadas en el espacio Lab (CIELAB) y otros parámetros significativos.

A continuación, se presentan en detalle los elementos que componen la matriz de datos iniciales, a la cual llamaremos a partir de ahora DataIn:

Columna	Descripción
1	Identificador del dato experimental (pareja de colores).
2-4	Coordenadas L*a*b* de primer color de la pareja.
5-7	Coordenadas L*a*b* de segundo color de la pareja.
8	**DV (ΔV):** Medida experimental de la diferencia de color percibida por un conjunto de observadores.
9	**DV (ΔV):** Normalizado: ΔV Normalizado a la escala de la fórmula de diferencia de color CIEDE2000
10	**DE (ΔE):** Medida calculada mediante la distancia euclídea en CIELAB: $$\boldsymbol{\Delta E} = \sqrt{\left(\boldsymbol{L_1^*} - \boldsymbol{L_2^{*2}}\right) + (\boldsymbol{a_1^*} - \boldsymbol{a_2^*})^2 + (\boldsymbol{b_1^*} - \boldsymbol{b_2^*})^2} \quad \textbf{(1)}$$

Tabla 1 - Descripción del contenido de la matriz de datos de entrada, DataIn

Muestra de dos datos experimentales de la base de datos Pointer:

Pair	L1	a1	b1	L2	a2	b2	DV	DV norm	DE00
11.1	64.46	-30.1	40.245	65.685	-29.885	40.815	1.063008	0.826932	1.046462
12.2	64.46	-30.1	40.245	66.47	-30.35	40.805	1.116652	0.868663	1.655305

Tabla 2 - Muestra de datos experimentales de la base de datos Pointer

4. MARCO TEÓRICO

Parte de este trabajo se basa en el marco teórico y experimental desarrollado en [6]-[8]. El método expuesto en [6] se centra en analizar conjuntos de datos que contienen diferencias de color percibidas por observadores en experimentos psicofísicos. Sin embargo, es una información ampliamente aceptada que la percepción humana no sigue una relación lineal cuando se trata de magnitudes de diferencias de color diversas [1,2,9-11].

En el contexto de una base de datos experimentales formada por colores, la "inconsistencia de un color" se refiere a la presencia de datos que no concuerdan adecuadamente. Estas inconsistencias pueden surgir por diversas razones y pueden afectar a la calidad de la base de datos en relación con la representación de colores. El enfoque propuesto consta de dos pasos fundamentales para evaluar la calidad y consistencia de estos datos.

En un primer paso, se identifican los pares de colores que muestran inconsistencias en el conjunto de datos. Esto es crucial, ya que las diferencias entre los mismos colores deben ser coherentes y no variar significativamente en función de diferentes condiciones experimentales, a menos que las condiciones experimentales no sean iguales. La detección de inconsistencias se basa en criterios de proximidad en el espacio de color CIELAB, lo que significa que los pares de colores que están muy cerca entre sí en este espacio y tienen diferencias de color visualmente similares deben mostrar diferencias de percepción consistentes. Si esto no ocurre, se consideran inconsistentes y se marcan para su posterior revisión o eliminación.

El segundo paso es asignar grados de consistencia a cada par de colores en comparación con otros pares similares en el conjunto de datos. Este proceso se basa en la idea de que los colores que son visualmente similares deberían mostrar diferencias de percepción también similares. Para lograrlo, se considera el marco matemático de la denominada lógica difusa [13] para modelizar conceptos como "consistente" y "similar".

A diferencia de la lógica clásica, que opera con valores de verdad estrictos de 0 o 1, la lógica difusa permite que las afirmaciones tengan grados de pertenencia a un conjunto determinado en un continuo entre 0 y 1. Esta característica es esencial para capturar y representar adecuadamente la imprecisión y la incertidumbre inherentes a numerosos problemas del mundo real.

La lógica difusa se ha convertido en un pilar en la teoría de conjuntos difusos, donde su capacidad para modelar y razonar en presencia de información vaga la hace una herramienta esencial para abordar problemas complejos y multifacéticos. La proximidad en el espacio de color y las diferencias de percepción se utilizan para calcular grados de consistencia para cada par de colores en función de su similitud con otros pares.

El grado de vecindad es un concepto esencial en este trabajo y se refiere a cuántos vecinos (datos experimentales) tiene un dato experimental dentro de la base de datos. En esencia, se trata de una medida que indica cuán relacionado o cercano está un dato en particular con otros datos en el conjunto de datos.

La lógica difusa se emplea para calcular este valor, lo que significa que el grado de vecindad de un dato experimental se expresa como un número en el rango de 0 a 1. Cuando este valor es cercano a 1, indica que el dato tiene una alta proximidad o similitud con otros datos en la base de datos, lo que representa es que tiene una gran cantidad de vecinos. Por otro lado, si el valor está más cerca de 0, sugiere que el dato tiene una relación débil o poca similitud con otros datos, lo que implica que tiene pocos "vecinos" o ninguno en absoluto. Proporciona información valiosa sobre la estructura y la distribución de los datos en el conjunto de datos.

Las mejoras introducidas en este método en comparación con enfoques anteriores incluyen un criterio de proximidad basado en la distancia entre las muestras de color en lugar de la distancia entre los centros de los colores, una comparación más precisa de pares de colores con diferencias visuales similares y la sustitución de medidas sensibles al ruido por un valor absoluto de la diferencia entre ΔV y ΔE_{00} [12].

Estas mejoras aumentan significativamente la precisión y la consistencia de los datos de diferencia de color perceptual en diversas condiciones experimentales. En conjunto, este método permite una gestión más efectiva de datos de diferencia de color perceptual y contribuye a la mejora de la calidad en la investigación en este campo.

En los conjuntos de datos que se van a analizar, cada dato experimental (par de colores) se representa como $\boldsymbol{S}i=\{\boldsymbol{A}_i,\boldsymbol{B}_i,\Delta V_i,\Delta E_{00,i}\}$. Para entenderlo mejor, desglosemos estos componentes:

- $\boldsymbol{A_i}$ y $\boldsymbol{B_i}$ son las coordenadas de color CIELAB de las dos muestras de color en el par. Estas coordenadas representan la información sobre la luminosidad (L^*), las componentes de cromaticidad a* y b* en el espacio de color CIELAB.

- $\boldsymbol{\Delta V_i}$ es la diferencia de color visual promedio que ha sido reportada por observadores. Esta medida refleja cómo perciben las personas la diferencia entre los dos colores en el par. Considerando que todos los observadores tienen visión normal del color y puede variar ligeramente de una persona a otra.

- $\boldsymbol{\Delta E_{00,i}}$ es la diferencia de color calculada a partir de las coordenadas de color CIELAB de las dos muestras utilizando la fórmula de diferencia de color CIEDE2000. Es una medida objetiva de la diferencia de color que se calcula utilizando una fórmula estándar. Es una forma de cuantificar cuán diferentes son los colores en función de sus coordenadas en el espacio de color CIELAB.

Calculamos dos medidas absolutas de diferencia:

$$\Delta\Delta V_{ij} = |\Delta V_i - \Delta V_j| \tag{2}$$

$$\Delta\Delta E_{ij} = |\Delta E_{00,i} - \Delta E_{00,j}| \tag{3}$$

Estos valores representan cuánto difieren las diferencias de color visual (ΔV) (Ec. 2) y las diferencias de color calculadas (ΔE_{00}) (Ec. 3) entre los pares $\boldsymbol{S}_i$ y $\boldsymbol{S}_j$, respectivamente.

Detectamos las inconsistencias existentes al comparar cómo deberían ser las diferencias de color visual (ΔV) en función de las diferencias de color calculadas (ΔE_{00}) entre dos conjuntos de colores que están cerca en el espacio de color y tienen valores de $\Delta E00$ similares.

Para expresar la similitud en el espacio de color de este par de datos S_i y S_j, definimos un par de distancias ($\Delta M_{i\,j,1}$, $\Delta M_{i\,j,2}$) de la siguiente manera:

Si $$||A_i - A_j|| + ||B_i - B_j|| \leq ||A_i - B_j|| + ||B_i - A_j|| \quad (4)$$

Entonces: $$(\Delta M_{ij,1}, \Delta M_{ij,2}) = \left(\left|\left|A_i - A_j\right|\right|, \left|\left|B_i - B_j\right|\right|\right) \quad (5)$$

Si no: $$(\Delta M_{ij,1}, \Delta M_{ij,2}) = \left(\left|\left|A_i - B_j\right|\right|, \left|\left|B_i - A_j\right|\right|\right) \quad (6)$$

El símbolo || representa la distancia euclídea.

Estas fórmulas cuantifican la diferencia en la proximidad de dos pares de colores en el espacio CIELAB, considerando tanto las coordenadas Lab del primer color como las del segundo color en cada par.

En ese trabajo, se proponen dos reglas difusas fundamentales para evaluar la inconsistencia entre dos pares de colores en función de medidas relacionadas con la percepción del color. Estas reglas son esenciales para determinar cuándo dos pares de colores se consideran inconsistentes entre sí.

4.1.Reglas difusas:

En el contexto del análisis de datos de color, se han desarrollado dos reglas difusas fundamentales, conocidas como regla difusa 1 y regla difusa 2, que desempeñan un papel crucial en la detección de inconsistencias en conjuntos de datos cromáticos. A continuación, exploraremos en detalle estas dos reglas:

4.1.1. Regla difusa 1

Esta regla se utiliza para evaluar la inconsistencia entre dos pares de datos. Para que se considere que el par de datos $\boldsymbol{S}_i$ y $\boldsymbol{S}_j$ es inconsistente según esta regla, deben cumplirse las siguientes condiciones:

- La distancia entre los dos pares de colores en el espacio de color CIELAB, representada por $\Delta M_{ij,1}$ y $\Delta M_{ij,2}$, debe ser pequeña, lo que indica que los colores están cercanos entre sí en este espacio.

- La diferencia entre las percepciones de color, $\Delta\Delta E_{ij}$,(Ec. 3) debe ser muy pequeña, lo que significa que los observadores apenas perciben diferencias entre estos colores.

- La diferencia en la percepción visual del color, $\Delta\Delta V_{ij}$, (Ec. 2) no debe ser pequeña, lo que indica que los observadores sí perciben una diferencia en la apariencia de estos colores.

Si $\Delta M_{ij,1}$ y $\Delta M_{ij,2}$ son pequeños, $\Delta\Delta E_{ij}$,(Ec. 3) es muy pequeño y $\Delta\Delta V_{ij}$ (Ec. 2) no es pequeño, entonces se considera que los pares de colores $\boldsymbol{S}_i$ y $\boldsymbol{S}_j$ son inconsistentes:

$$I_{ij} = \Delta M_{ij,1}^{pequeño} \cdot \Delta M_{ij,2}^{pequeño} \cdot \Delta\Delta E_{ij}^{muy\ pequeño} \Delta\Delta V_{ij}^{no\ pequeño} \qquad (7)$$

Si estas condiciones se cumplen, se asigna un valor en el rango de [0,1] para representar el grado de inconsistencia entre los pares de colores, donde un valor más cercano a 1 indica una mayor inconsistencia. Para calcular el grado de certeza de que estas condiciones se cumplen, se utiliza una t-norma continua, en este caso, la t-norma de producto clásica, que multiplica las certezas de cada uno de los términos vagos involucrados en la regla.

4.1.2. Regla difusa 2

Esta regla es similar a la regla difusa 1, pero intercambia las medidas relacionadas con la percepción del color. Para considerar que el par de datos $\boldsymbol{S}_i$ y $\boldsymbol{S}_j$ es inconsistente según esta regla, deben cumplirse las siguientes condiciones:

- La distancia entre los dos pares de colores en el espacio de color CIELAB, $\Delta M_{ij,1}$ y $\Delta M_{ij,2}$, debe ser pequeña.

- La diferencia entre las percepciones visuales del color, $\Delta\Delta V_{ij}$, (Ec. 2) debe ser muy pequeña.

- La diferencia en las percepciones de color, $\Delta\Delta E_{ij}$,(Ec. 3) no debe ser pequeña.

Si $\Delta M_{ij,1}$ es pequeño, $\Delta M_{ij,2}$ es pequeño, $\Delta\Delta V_{ij}$ es muy pequeño y $\Delta\Delta E_{ij}$ no es pequeño, entonces se considera que los pares de colores $\boldsymbol{S}_i$ y $\boldsymbol{S}_j$ son inconsistentes.

$$I_{ij}^{*} = \Delta M_{ij,1}^{pequeño} \cdot \Delta M_{ij,2}^{pequeño} \cdot \Delta\Delta V_{ij}^{muy\ pequeño} \Delta\Delta E_{ij}^{no\ pequeño} \qquad (8)$$

Si estas condiciones se cumplen, se asigna un grado de inconsistencia difusa entre los pares de colores, denotado como I_{ij}^{*}. Al igual que en la Regla Difusa 1, la

decisión de si los pares de colores son inconsistentes se basa en si los valores de I_{ij} [Ec. (6)] o I_{ij}^{*} [Ec. (7)]superan un umbral fijo específico. Es importante destacar que estas reglas identifican pares de colores inconsistentes, pero no indican cuál de los dos pares de colores podría considerarse incorrecto.

El umbral se define como un valor numérico. Por ejemplo, al establecer un umbral de 0.75, implica que cuando los valores de los parámetros que correspondan según lo determinado por la regla difusa exceden dicho valor, el par de datos es inconsistente.

4.2. Standardized Residual Sum of Squares (STRESS)

La Figura de mérito conocida como STRESS [14] (Standardized Residual Sum of Squares) se usa en un número no pequeño de diferentes contextos, para evaluar la calidad de las representaciones visuales de bases de datos antes y después de eliminar datos experimentales inconsistentes. Es una métrica utilizada en el contexto de la evaluación de la precisión de un modelo o sistema en relación con las observaciones visuales realizadas por un grupo de individuos.

Antes de eliminar los datos inconsistentes, calculamos la métrica STRESS para evaluar cuán bien se ajustaban las representaciones visuales a las similitudes originales entre los objetos en la base de datos. Un valor alto de STRESS indicaría que la representación visual no captura adecuadamente las relaciones reales entre los datos, lo que podría deberse a datos inconsistentes o ruidosos.

Una vez eliminados los datos experimentales inconsistentes, recalculamos el STRESS para la base de datos depurada. Si el STRESS disminuye significativamente después de la depuración de datos, esto indicaría que la calidad de la representación visual ha mejorado, ya que las similitudes entre los objetos se representan de manera más precisa en el espacio visual.

STRESS se define como:

$$STRESS = 100 \cdot \sqrt{\frac{\sum (F_2 \Delta E_i - \Delta V_i)^2}{\sum \Delta V_i^2}} \tag{9}$$

$$\text{donde } F_2 = \frac{\sum \Delta E_i \Delta V_i}{\Delta E_i^2} \tag{10}$$

F_2 es un factor de escala que se utiliza para ajustar los valores de ΔE_i y ΔV_i a la misma escala.

5. OBJETIVOS

El objetivo de este trabajo es evaluar la consistencia en conjuntos de datos que contienen diferencias de color medidas mediante el espacio de color CIELAB, así como las correspondientes medidas de color percibidas. Este método se utiliza para identificar parejas de colores inconsistentes en los datos y, en última instancia, mejorar la calidad de los conjuntos de datos utilizados en el estudio de las diferencias de color perceptual.

Para comenzar, hemos desarrollado un script en MATLAB que automatiza el análisis de consistencia de datos utilizando dos reglas difusas, denominadas FuzzyRule1 y FuzzyRule2. El script opera en una base de datos particular y genera resultados que se almacenan en archivos CSV para su posterior análisis. Se elige la regla difusa que se utilizará y se especifica la base de datos a analizar.

El script realiza un proceso donde se establecen varios umbrales y, después, ejecuta las funciones de regla difusa 1 o 2, según corresponda, en un ciclo. Esto se repite para cada uno de los umbrales de forma secuencial. Los resultados obtenidos en cada ejecución se guardan en archivos CSV, incluyendo los valores de umbral y los datos resultantes.

Este enfoque automatizado permite analizar la consistencia de los datos de manera sistemática y eficiente en diferentes condiciones y con diferentes reglas difusas. De este modo se obtienen unos resultados que permiten detectar los datos que son inconsistentes.

5.1. Detección de inconsistencias

A continuación, vamos a detallar un par de funciones creadas en MATLAB que tienen la capacidad de detectar pares inconsistentes en función de un umbral y una base de datos proporcionados. Estas funciones son herramientas útiles para analizar la consistencia de datos en conjuntos de datos de diferencia de color y pueden ayudar a identificar y eliminar datos que no se ajustan adecuadamente a ciertas reglas de consistencia.

5.1.1. Implementación de la regla difusa 1

La función de MATLAB, denominada `InconsistenciesFuzzyRule1`, se utiliza para analizar la consistencia de datos en un conjunto de datos particular. Funciona tomando un conjunto de datos de entrada `DatosIn` y un umbral como parámetro. El objetivo principal de esta función es calcular y evaluar la inconsistencia entre pares de datos en función de varios criterios relacionados con la percepción del color.

Primero, la función inicializa algunas variables y matrices necesarias para almacenar los resultados. Posteriormente, se recorre cada par de datos del conjunto, calculando la inconsistencia en relación con otros pares similares.

Utiliza medidas como la diferencia de color visual promedio (ΔE_{00}) y la distancia entre las coordenadas de color CIELAB para evaluar la discrepancia entre los pares.

La función utiliza conceptos de lógica difusa para calcular grados de inconsistencia, como se ha explicado en el apartado 4. Esto significa que no considera la inconsistencia como un valor binario, sino que asigna un grado de inconsistencia en un rango de 0 a 1, donde 0 indica completa consistencia y 1 indica completa inconsistencia.

El cálculo del grado de inconsistencia involucra factores como la proximidad en el espacio de color, la diferencia en las coordenadas de color y la diferencia en la percepción visual del color. Si el grado de inconsistencia calculado para un par de datos supera el umbral especificado, se considera que estos datos son inconsistentes y se registran en la matriz ParesInconsistentes, que posteriormente pasa a ser un fichero CSV para su procesamiento en Python.

5.1.2. Implementación de la regla difusa 2

La función “InconsistenciesFuzzyRule2” en MATLAB se utiliza para analizar la consistencia de datos en un conjunto de datos específicos. Al igual que la función “FuzzyRule1”, toma un conjunto de datos de entrada DatosIn y un umbral como parámetro. Su objetivo principal es calcular y evaluar la inconsistencia entre pares de datos en función de varios criterios relacionados con la percepción del color.

La función comienza inicializando algunas variables y matrices necesarias para almacenar los resultados. Luego, inicia un bucle que recorre cada par de datos en el conjunto. Para cada par, calcula la inconsistencia en relación con otros pares similares utilizando medidas como la diferencia de color percibida (ΔV) y la diferencia de color calculada (ΔE).

Al igual que en FuzzyRule1, se utiliza la lógica difusa para calcular grados de inconsistencia en un rango de 0 a 1, donde 0 representa completa consistencia y 1 representa completa inconsistencia. El cálculo de la inconsistencia involucra factores como la proximidad en el espacio de color y la diferencia en los valores de ΔV y ΔE.

Si el grado de inconsistencia calculado para un par de datos supera el umbral especificado, se considera que estos datos son inconsistentes y se registran en la matriz ParesInconsistentes. Además, se calcula la inconsistencia máxima para el par de datos actual.

El proceso de cálculo y evaluación de la inconsistencia se repite para todos los pares de datos en el conjunto. Los resultados se almacenan en la matriz DatosOut para su posterior análisis. Esta matriz contiene información sobre el grado de inconsistencia, el número de vecinos considerados y otros valores relacionados con la inconsistencia.

Las principales diferencias entre las funciones de MATLAB FuzzyRule1 y FuzzyRule2 radican en cómo evalúan y calculan la inconsistencia entre pares de datos en función de criterios relacionados con la percepción del color:

Criterios de Inconsistencia:

- FuzzyRule1 se enfoca en evaluar la inconsistencia en función de las diferencias en las coordenadas de color CIELAB y en la diferencia de color percibido (ΔV) entre los pares de datos.

- FuzzyRule2, por otro lado, se centra en evaluar la inconsistencia en función de diferencias en las coordenadas de color CIELAB y en las diferencias de percepción visual del color (ΔE).

Normas Difusas y Umbral:

- Ambas funciones utilizan lógica difusa para calcular grados de inconsistencia, pero las normas difusas pueden variar.

- Los umbrales de inconsistencia también pueden variar entre ambas funciones, lo que significa que un par de datos puede considerarse inconsistente en una función, pero no en la otra, dependiendo del umbral específico que se establezca.

Las funciones FuzzyRule1 y FuzzyRule2 se diferencian en los criterios específicos utilizados para evaluar la inconsistencia entre pares de datos, así como en las variables principales y las normas difusas empleadas en sus cálculos. Estas diferencias permiten abordar diferentes aspectos de la percepción del color en el análisis de consistencia de datos.

Estas funciones generan un fichero compuesto por una matriz con los datos experimentales originales, junto con nuevos parámetros, en adelante, DataOut. Estos parámetros adicionales brindan información valiosa sobre cada par de color en estudio y complementan las coordenadas L*a*b* y las diferencias de color que conforman los aspectos fundamentales de nuestros datos.

Las propiedades adicionales, que se registran en la nueva matriz de datos experimentales de salida, se incluyen las anteriormente mencionadas y, además:

Columna	Descripción
11	Número de vecinos difusos con los que se compara cada par. Esta columna representa la suma de los grados de vecindad difusa de los pares similares en el conjunto de datos. Es esencial que este valor alcance al menos el valor 1 para considerar que el par en estudio tiene algún otro par con el que compararse.
12	Número de vecinos con grados de vecindad no nulos. Estos son aquellos cuyos grados de vecindad difusa se suman para calcular el valor de la columna 11.
13	Inconsistencia promedio encontrada para el par al compararse con todos sus vecinos difusos. Esta columna representa la discrepancia promedio entre el par de colores y sus vecinos en términos de diferencia de color perceptual.
14	Inconsistencia máxima encontrada para el par al compararse con todos sus vecinos difusos. Esta columna refleja la mayor discrepancia entre el par de colores y sus vecinos en términos de percepción de diferencia de color.

Tabla 3 - Descripción del contenido de la matriz de datos de salida, DataOut.

Es importante destacar que los datos en las columnas 12, 13 y 14 son únicamente descriptivos y no se han utilizado en este trabajo, pero proporcionan información adicional sobre la relación de cada par de colores con sus vecinos en el espacio de color Lab y su consistencia perceptual.

Adicionalmente, se crea un archivo de pares inconsistentes formado por una matriz de pares de diferencias de color para los cuales se ha encontrado una inconsistencia superior al umbral establecido.

Si se encuentran N pares inconsistentes, el fichero de pares inconsistentes tendrá $2N$ filas. Es decir, habrá 2 filas para cada inconsistencia, y cada una contendrá un par de colores. Por lo tanto, en esta matriz, los pares de colores en las filas 1 y 2 son inconsistentes entre sí, al igual que los pares en las filas 3 y 4, 5 y 6, y así sucesivamente, siguiendo un patrón de $2_n - 1\ y\ 2_n$.

El tamaño de la matriz, si se encuentran N inconsistencias, será de $2Nx11$. Estas 11 columnas incluyen las mismas 10 columnas que se encuentran en DataIn, además de una última columna que indica el grado de inconsistencia encontrado para ese par de colores y el siguiente/anterior. Es importante destacar que el valor en las 11 columnas es idéntico para las filas $2_n - 1\ y\ 2_n$, ya que representa el grado de inconsistencia encontrado entre los pares en esas dos filas consecutivas.

Una vez que se han generado los pares inconsistentes y calculado los valores de salida para una amplia gama de umbrales, que van desde 0.05 hasta 1 con un paso de 0.05, el objetivo final es identificar y eliminar los datos experimentales menos consistentes de la base de datos. Esta fase es esencial para mejorar la calidad y consistencia de los datos que se están analizando.

Durante este proceso de detección y eliminación, se busca identificar aquellos datos que muestran una baja consistencia en comparación con el resto de la base de datos. Estos datos experimentales menos consistentes pueden introducir ruido o sesgos en el análisis general y, por lo tanto, es fundamental reducir su influencia en los resultados finales.

Al eliminar estos datos menos consistentes, se logra un conjunto de datos más depurado y preciso, lo que a su vez mejora la calidad de los análisis y evaluaciones que se realicen posteriormente. Esto es especialmente valioso en investigaciones y aplicaciones donde la calidad de los datos es crítica para la toma de decisiones informadas y la obtención de resultados confiables.

6. METODOLOGÍA

En este trabajo, es fundamental comprender la metodología que se ha empleado para analizar la consistencia de los datos perceptuales de diferencia de color en grandes conjuntos de datos. Nuestra metodología implica una serie de pasos clave que se describen a continuación.

Los scripts desarrollados en este análisis, están disponibles en el siguiente repositorio https://github.com/DanielUVA/colorConsistency Estos scripts están diseñados para ser fácilmente ejecutables, lo que permite que cualquier persona interesada pueda acceder a ellos y probar el código fuente.

6.1. Preparación de los datos

El primer paso de nuestra metodología se enfocó en la preparación de los datos. Para ello, reunimos varias bases de datos que albergan una gran cantidad de observaciones experimentales. Cada una de estas observaciones estás formada por la diferencia de color entre dos muestras con colores diferentes, aunque parecidos. Para facilitar la gestión y el análisis de estos datos, asignamos un identificador único a cada conjunto de datos. Es importante destacar que cada color en estas bases de datos se encuentra definido por sus respectivos parámetros en el espacio de color Lab. Esta preparación inicial de datos es esencial para abordar de manera efectiva el análisis de la consistencia en la percepción de la diferencia de color.

6.2. Procesamiento de los datos experimentales

En el segundo paso de nuestra metodología, nos centramos en el procesamiento de los datos. Para ello, ejecutamos un script en MATLAB que invoca las funciones InconsistenciesFuzzyRule1 e InconsistenciesFuzzyRule2, dependiendo de la regla seleccionada (FuzzyRule1 o FuzzyRule2).

Los resultados de esta evaluación, almacenados en varios ficheros CSV, contienen información sobre el umbral, los datos de salida y los pares inconsistentes encontrados, para cada valor de umbral utilizado.

6.3. Análisis de los datos y cálculo de métricas

En la tercera etapa de nuestro proceso, nos adentramos en el análisis exhaustivo de los datos y la aplicación de métricas. Elegimos una estrategia ventajosa al incorporar un cuaderno de Python en nuestro proceso. Esta elección se fundamenta en la necesidad de contar con herramientas versátiles y visualmente intuitivas para comprender y procesar con eficacia los resultados previamente obtenidos en las fases anteriores del estudio.

Este conjunto de scripts se ha diseñado con el propósito de automatizar tareas repetitivas y complejas, lo que agiliza significativamente el proceso de análisis de datos. Además, proporciona una mayor flexibilidad al permitir la personalización de las métricas y análisis específicos que se deseen aplicar a los datos.

Comenzamos evaluando nuestros datos experimentales y los resultados de las reglas difusas 1 y 2 para determinar su consistencia. Llevamos a cabo este proceso al analizar la interacción de todos los datos experimentales entre sí., considerando cuántos pares de colores similares existen en nuestro conjunto de datos en relación con cada par específico que estamos estudiando.

Para medir esta similitud, utilizamos el concepto de grado de vecindad, que refleja cuántos pares de colores similares se pueden comparar con el par en cuestión. En casos donde este valor sea inferior a 1, sugerimos que el análisis carece de relevancia para ese par en particular en nuestro trabajo.

En esta tarea, hacemos uso de una serie de bibliotecas de *Python*, entre las que destacan *Pandas, Matplotlib, Seaborn y Numpy*. Estas bibliotecas proporcionan una combinación poderosa que nos permite abordar distintos aspectos del análisis de datos.

Inicialmente, *Pandas* emerge como una libreria esencial, ya que nos brinda la capacidad de cargar y estructurar los datos exportados en formato CSV desde MATLAB. Esta tarea resulta crucial para organizar los datos de manera eficiente, lo que simplifica significativamente el proceso de análisis subsiguiente.

En la siguiente fase, empleamos las capacidades de visualización de *Matplotlib* para generar gráficos informativos. Estos gráficos nos permiten explorar la distribución de los datos, identificar patrones y tendencias ocultas, y detectar cualquier posible anomalía en la coherencia de los datos perceptuales de diferencia de color.

Por ejemplo, creamos gráficos que evalúan la consistencia de los datos en función del número de vecinos difusos para el par de colores (una medida que refleja la suma de grados de vecindad difusos de pares de colores similares en el conjunto de datos). También generamos gráficos que muestran cómo varía el número de pares inconsistentes para diferentes valores de umbral.

Antes de eliminar los datos inconsistentes, calculamos la métrica STRESS para evaluar cuán bien se ajustaban las representaciones visuales a las similitudes originales entre los objetos en la base de datos. Un valor alto de STRESS indicaría

que la representación visual no captura adecuadamente las relaciones entre los datos, lo que podría deberse a datos inconsistentes o ruidosos.

Posteriormente, se eliminan los datos experimentales inconsistentes y recalculamos el STRESS para la base de datos depurada. Si el STRESS disminuye significativamente después de la depuración de datos, esto indicaría que la calidad de la base de datos ha mejorado, ya que los datos son más consistentes.

Este enfoque exhaustivo nos permite identificar patrones y tendencias en los datos que podrían no ser evidentes si se analizaran de manera aislada. Además, al abordar las dos reglas difusas, podemos comparar y contrastar los resultados obtenidos, lo que puede proporcionar información de gran utilidad sobre cómo se comportan las medidas de color en diferentes contextos y bajo diferentes criterios de consistencia.

7. RESULTADOS

Tras realizar el procedimiento propuesto en la sección 6, presentaremos los resultados de nuestro análisis y comparación de los datos obtenidos de las distintas bases de datos, así como los resultados de las dos reglas Fuzzy1 y Fuzzy2. Utilizaremos gráficos y visualizaciones para ilustrar de manera efectiva cómo se comportan estos datos en diferentes contextos.

Además, llevaremos a cabo comparaciones entre las bases de datos, lo que nos permitirá evaluar la consistencia de nuestros resultados en un contexto más amplio. Esto nos ayudará a determinar si las tendencias y patrones identificados son consistentes a través de diferentes conjuntos de datos y, por lo tanto, más robustos en términos de aplicabilidad.

7.1. Distribución de los datos experimentales en función del número de vecinos difusos

En la siguiente sección de este estudio, se presentarán una serie de gráficas que muestran la distribución de los datos experimentales en función del número de vecinos difusos para diversas bases de datos. Estas gráficas proporcionarán una representación visual clara de cómo los datos experimentales se distribuyen y agrupan en relación con diferentes configuraciones de vecinos difusos.

Las gráficas proporcionarán una valiosa perspectiva visual que complementará el análisis cuantitativo de los datos, ofreciendo una visión completa de la estructura de los datos experimentales en este contexto.

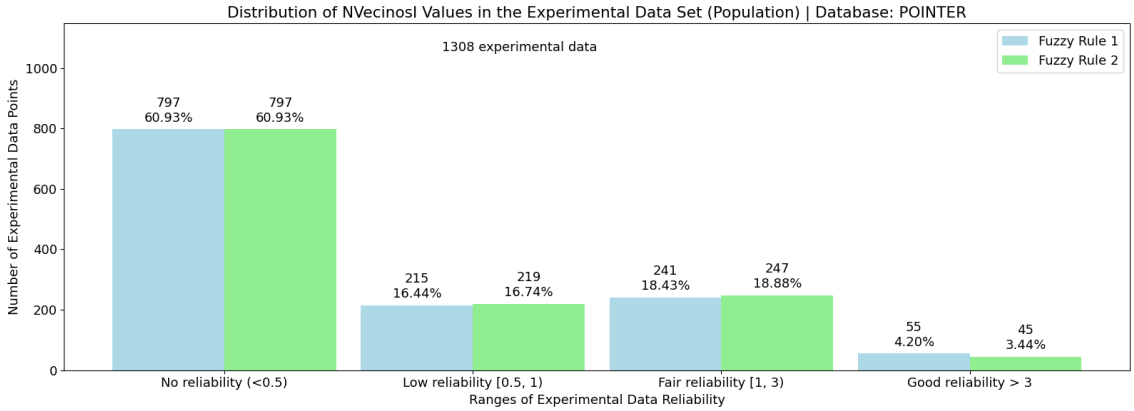

Figura 2 - Distribución del número de vecinos difusos de la base de datos Pointer

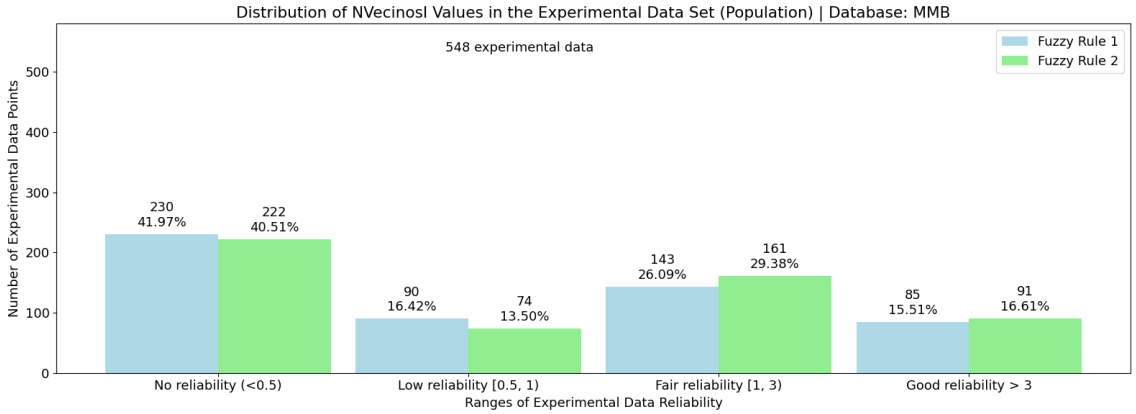

Figura 3 - Distribución del número de vecinos difusos de la base de datos MMB

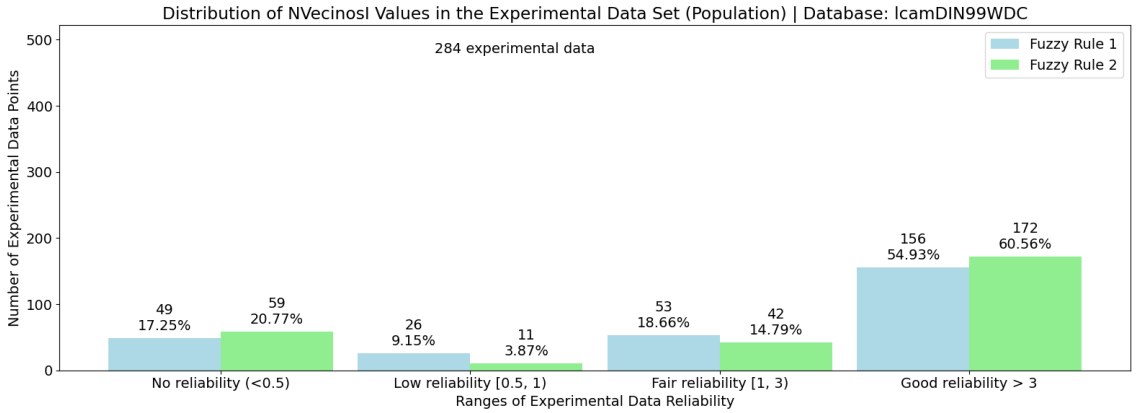

Figura 4 - Distribución del número de vecinos difusos de la base de datos IcamDIN99WDC

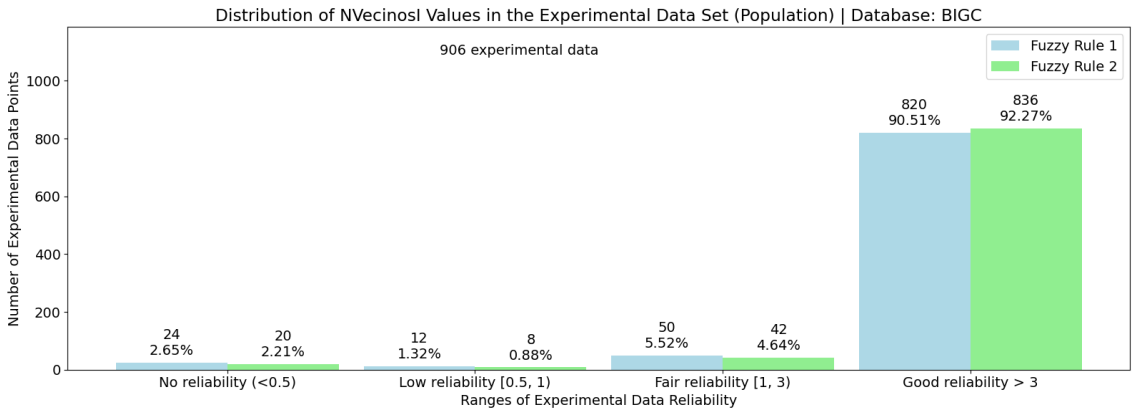

Figura 5 - Distribución del número de vecinos difusos de la base de datos BIGC

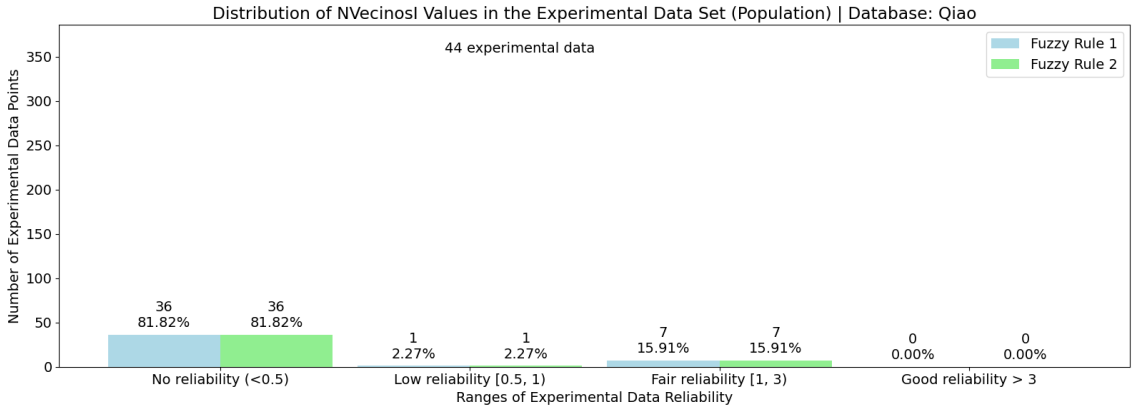

Figura 6 - Distribución del número de vecinos difusos de la base de datos Qiao

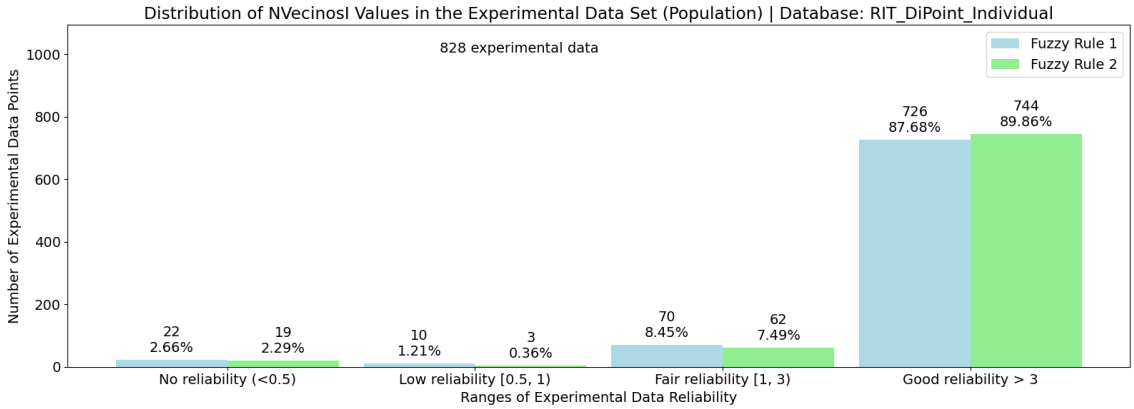

Figura 7 - Distribución del número de vecinos difusos de la base de datos RIT_DiPoint_Individual

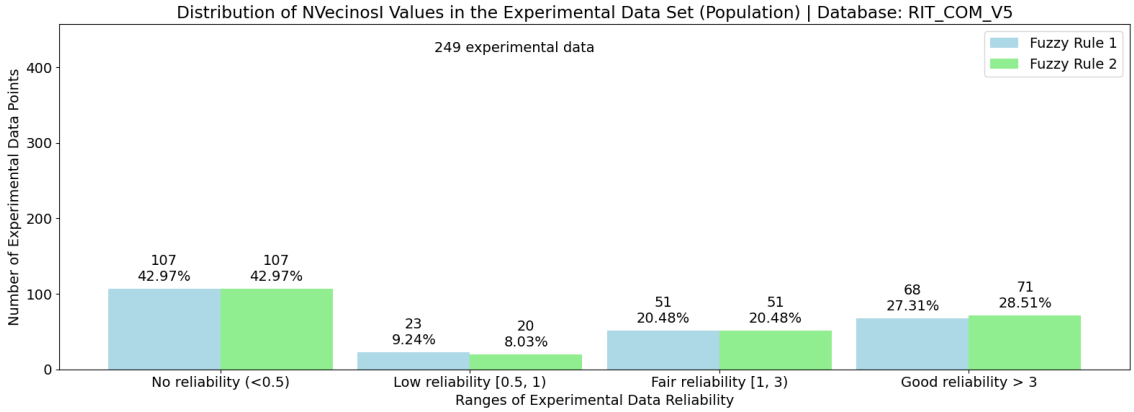

Figura 8 - Distribución del número de vecinos difusos de la base de datos RIT_Com_V5

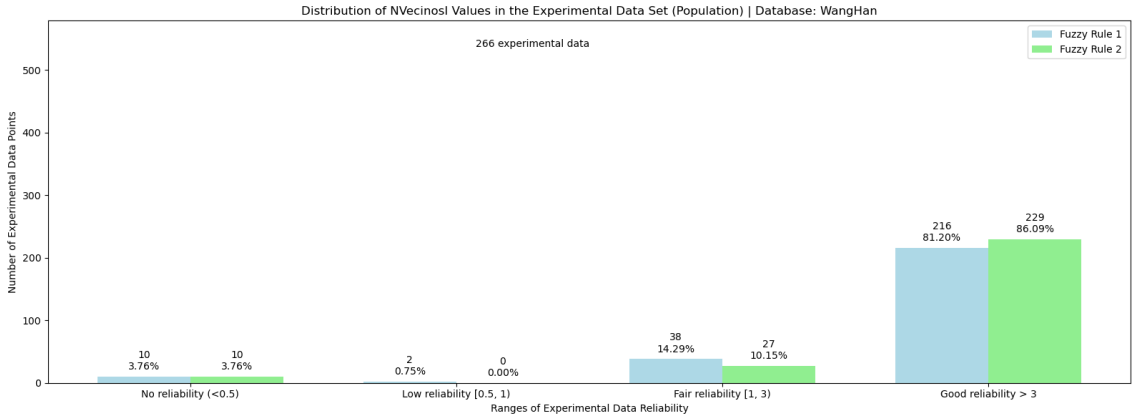

Figura 9 - Distribución del número de vecinos difusos de la base de datos WangHan

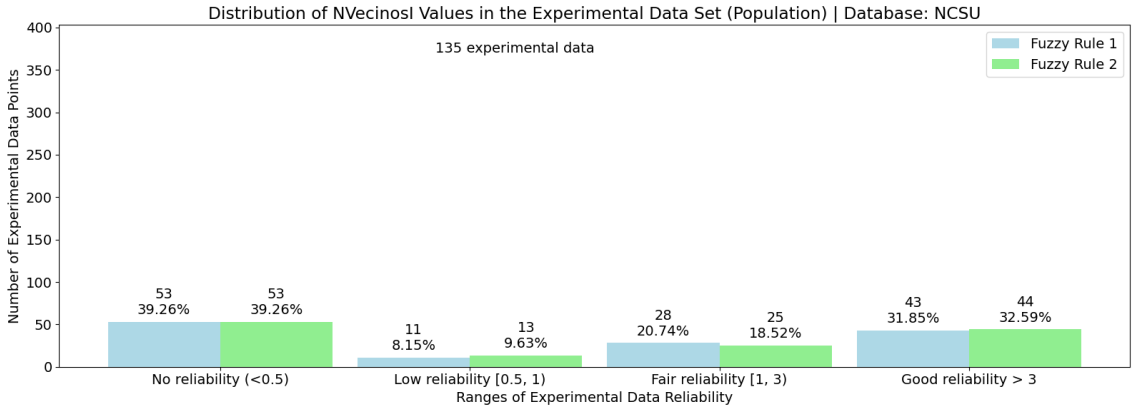

Figura 10 - Distribución del número de vecinos difusos de la base de datos NCSU

Las Figuras 2-10 representan la agrupación de los datos experimentales de cada una de las bases de datos en función de la suma de grados de vecindad total o número de vecinos difusos.

Establecer varios niveles de vecindad nos permite comprender mejor cómo influye este valor en la evaluación de la calidad de los datos experimentales:

- Grado de vecindad muy bajo: Niveles menores que 0.5. Cuando el valor del número de vecinos difusos es menor que 0.5, indica una situación en la que hay una falta significativa de información circundante para determinar la calidad de un dato experimental. En este escenario, la evaluación puede ser altamente incierta debido a la escasez de puntos de referencia cercanos o de similitud.

- Grado de vecindad bajo: Niveles en el rango [0.5, 1). Un valor del número de vecinos en este intervalo sugiere que existe cierta información disponible para evaluar la calidad de los datos experimentales, pero aún no se dispone de una cantidad suficiente para hacer una evaluación concluyente. Es un nivel de confianza moderada.

- Grado de vecindad moderado: Niveles en el rango [1, 3). En esta franja, que comprende desde 1 hasta 3, el grado de vecindad indica que existe una cantidad razonable de información disponible para realizar una evaluación sólida de la calidad de los datos experimentales. Los datos tienen un nivel adecuado de contexto y vecindad para tomar decisiones más informadas sobre su validez.

- Grado de vecindad bueno: Niveles mayores que 3. Cuando el grado de vecindad es mayor que 3, denota una situación en la que abunda la información circundante y se dispone de un sólido conjunto de referencias para determinar con confianza si un dato experimental es de buena calidad o no. Este nivel de información permite tomar decisiones sólidas y precisas en la evaluación de los datos.

Al analizar estas categorías en diferentes bases de datos, se obtienen observaciones interesantes:

- La Figura 2, representa la base de datos "Pointer", donde más de la mitad de los datos experimentales, son clasificados como datos que no tienen un grado de vecindad muy bajo respecto a otros datos de la base de datos.
- En las Figuras 3, 8 y 10 se representan las bases de datos MMB y RIT_COM_V5, NCSU cuya la distribución de datos es más equilibrada, con un énfasis ligeramente mayor en la categoría de bajo nivel de vecindad. Sin embargo, también son notables los datos con un grado de vecindad moderado.
- En la base de datos IcamDIN99WDC, Figura 4, cuyos datos estan distribuidos de manera bastante equitativa en todos los niveles de veciendad definidos.
- En la bases de datos BIGC, RIT_DuPoint_Individual WangHan, Figuras 5, 7 y 9, aproximadamente el 90% de los datos tienen un alto grado de vecindad. Esto destaca notablemente sobre las demas bases de datos.
- Finalmente, en la base de datos Qiao, en la Figura 6, la gran mayoria de sus datos tienen un nivel de vecindad muy bajo.

Estos gráficos, resaltan cómo los grados de vecindad de los datos experimentales varían según la base de datos, lo que puede guiar la elección de la base de datos más adecuada para un estudio específico y la necesidad de aplicar correcciones.

7.2. Número de pares inconsistentes en función del umbral

A continuación, se presentaran una serie de Figuras que exploran una relación fundamental: el número de pares de datos experimentales inconsistentes en función del valor del umbral. Estas gráficas proporcionarán una valiosa visión sobre cómo la elección del umbral afecta la identificación de datos experimentales inconsistentes.

Estas gráficas permiten tomar decisiones informadas sobre qué umbral utilizar en una determinada aplicación o análisis, ya que nos mostrarán cuán sensible es la identificación de datos inconsistentes cuando se producen cambios en este parámetro.

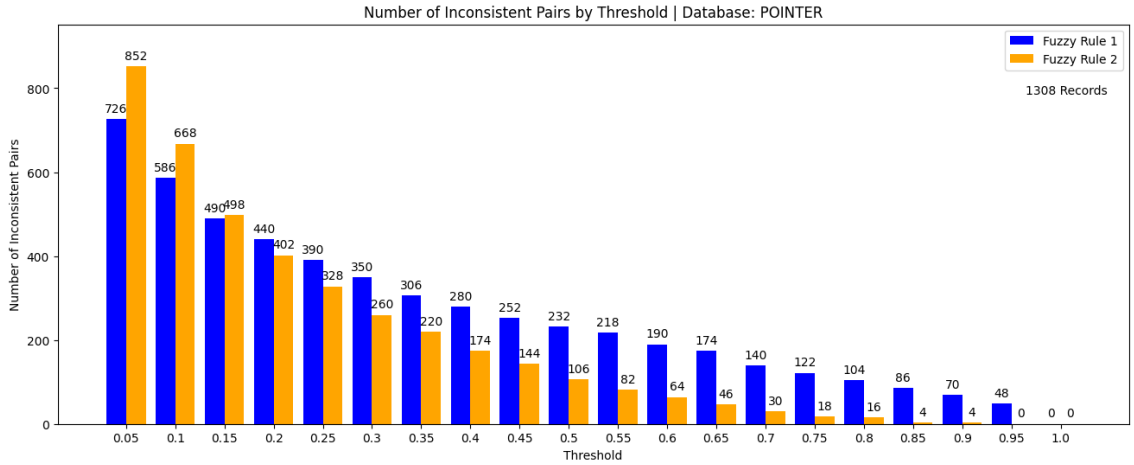

Figura 11 - Distribución del número pares inconsistentes por umbral para base de datos Pointer

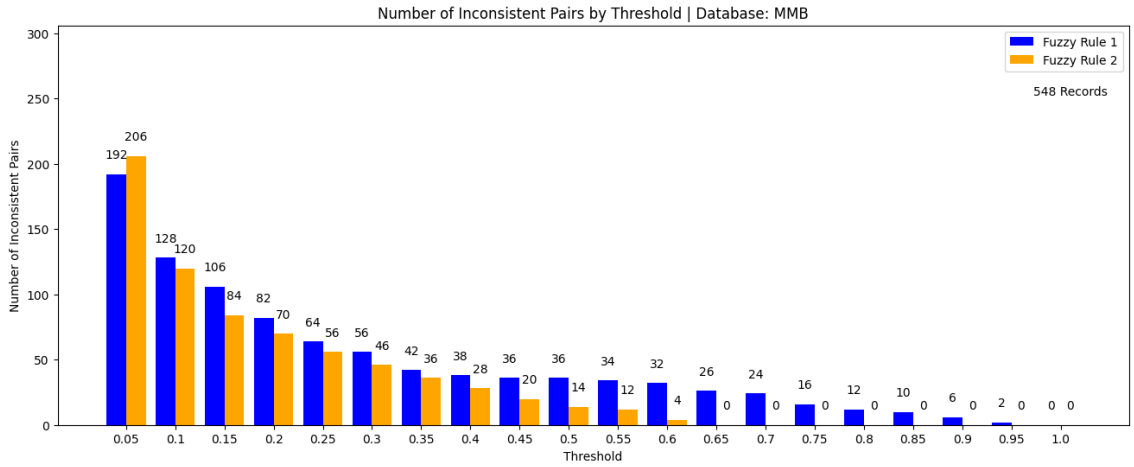

Figura 12 - Distribución del número pares inconsistentes por umbral para base de datos MMB

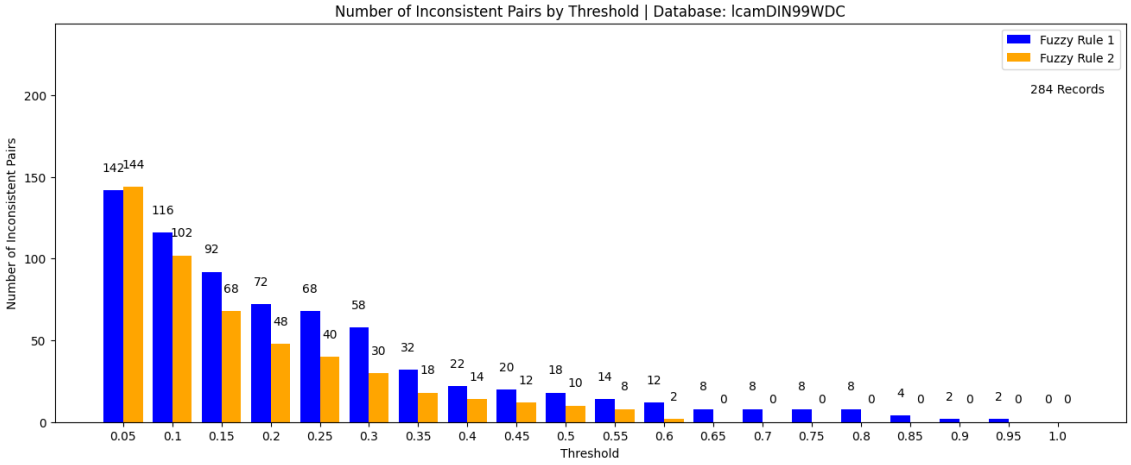

Figura 13 - Distribución del número pares inconsistentes por umbral para base de datos IcamDIN99WDC

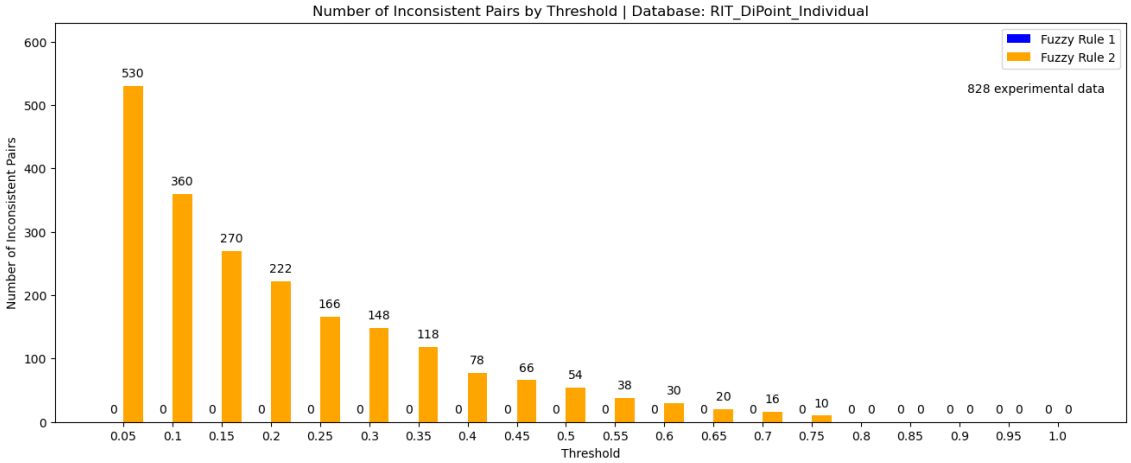

Figura 14 - Distribución del número pares inconsistentes por umbral para base de datos RIT_DiPoint_Individual

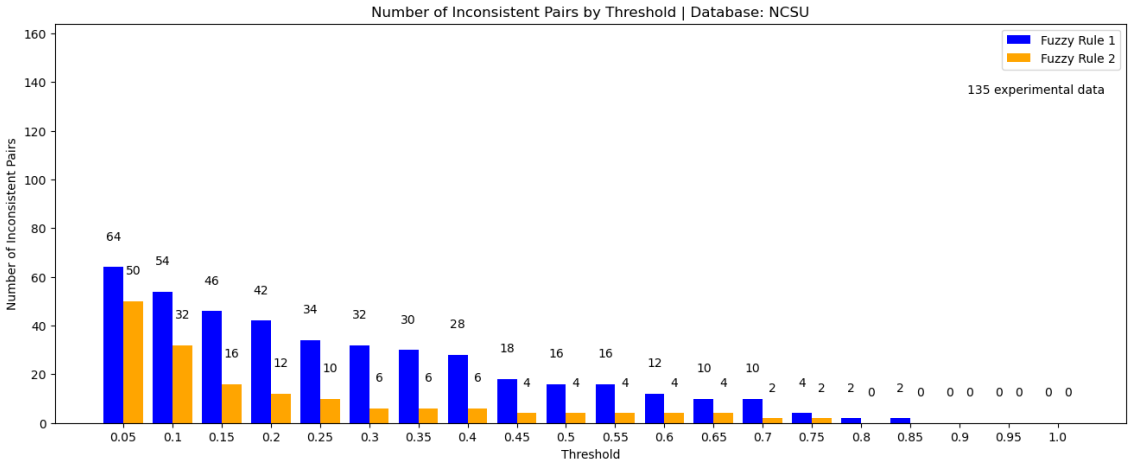

Figura 15 - Distribución del número pares inconsistentes por umbral para base de datos NCSU

En estos gráficos, el eje x muestra los valores del umbral, que varían desde 0.05 hasta 1 con un aumento gradual de 0.05 en cada paso. El umbral representa la medida de cuán similares deben ser dos pares de colores para considerarse consistentes. Cuando el umbral es bajo, como 0.05, se requiere una similitud muy alta entre los pares de colores para considerarlos consistentes. Por lo tanto, es más probable que se detecten pares inconsistentes en este punto.

A medida que aumentamos el umbral, se vuelve más tolerante a las diferencias de color, lo que significa que se necesitaría una mayor diferencia para considerar dos pares de colores como inconsistentes. Es por eso, por lo que, en los gráficos, observamos un aumento en el número de pares inconsistentes a medida que disminuye el umbral. A medida que aumentamos el umbral, es decir, nos movemos hacia la derecha en el gráfico, el número de pares inconsistentes tiende a disminuir.

Lo que es especialmente interesante es que, en algunos casos, el número de pares inconsistentes llega a cero a medida que el umbral se incrementa. Esto significa que, para ciertos valores de umbral, no se detectan pares de colores inconsistentes en los datos, lo que sugiere que los datos son altamente consistentes dentro de ese rango de umbral. Sin embargo, este punto puede variar según la base de datos específica, lo que demuestra la importancia de personalizar el umbral según el contexto y los requisitos del estudio.

En las Figuras 11 y 15, se observa que el número de pares inconsistentes para la base de datos Pointer y NCSU para la regla 1 tiende a seguir una tendencia lineal disminuyendo la cantidad de pares inconsistentes a medida que aumenta el umbral de manera gradual.

Por otro lado, en las Figuras 12 y 13, en las bases de datos MMB e IcamDIN99WDC, se nota una tendencia exponencial en el número de pares inconsistentes para la regla 1. Esto indica que pequeños cambios en los parámetros o valores pueden llevar a un aumento significativo y no lineal en la cantidad de pares inconsistentes.

Una diferencia significativa que vemos en la Figura 14 para la base de datos RIT_DuPont_Individual, es que solo se obtienen pares inconsistentes utilizando la regla 2, a diferencia de otras donde se obtienen pares inconsistentes en ambas reglas. Esta singularidad resalta que la naturaleza de la base de datos podría requerir un enfoque específico en el análisis y mejora de la consistencia de los datos, adaptándose a su particularidad dentro del conjunto de bases de datos estudiadas.

En el caso de la regla 2, se observa una tendencia clara en el número de pares inconsistentes que disminuye de manera exponencial a medida que se incrementa el umbral, como podemos ver en las figuras 11-15.

Sin embargo, en las bases de datos Qiao, RIT_COM_V5 y WangHan, esta tendencia parece diferir de las anteriores ya que no se detectan pares inconsistentes.

En conjunto, estas observaciones resaltan cómo la tendencia en el número de pares inconsistentes para la regla 1 puede variar significativamente entre diferentes bases de datos. Esto puede deberse a las características inherentes de los datos en cada base y sugiere la importancia de comprender estas tendencias para un análisis de datos más preciso y efectivo.

7.3. Mejora del valor de STRESS

Para adentrarnos en el tema que abordaremos, es fundamental explorar cómo la meticulosa eliminación de datos inconsistentes puede tener un impacto significativo en la mejora del valor de STRESS. Este proceso de limpieza de datos juega un papel crucial en la optimización de la calidad de nuestros resultados, y a lo largo de esta discusión, exploraremos cómo esta estrategia se traduce en mejoras palpables en la evaluación de las métricas y la precisión de nuestro estudio.

El criterio utilizado para elegir el umbral se basa en la mejora del nivel de STRESS a medida que se eliminan los datos experimentales inconsistentes. En general, la calidad de la base de datos es mejor a medida que se eliminan más datos inconsistentes.

Sin embargo, existe un punto crítico en este proceso. Llega un punto en el que los datos experimentales que permanecen en la base de datos después de la eliminación son igualmente consistentes en comparación con los que se eliminaron. En este punto, el valor del STRESS deja de mejorar y permanece prácticamente constante.

Este enfoque busca encontrar un equilibrio entre la mejora de la calidad de la base de datos al eliminar datos inconsistentes y preservar la naturaleza inicial de la base de datos. Si se eliminan demasiados datos, la base de datos podría perder su representación original y convertirse en un conjunto más reducido, pero altamente consistente. Por lo tanto, este criterio apunta a encontrar el punto óptimo en el que se maximiza la calidad de la base de datos sin perder su integridad inicial.

Pointer

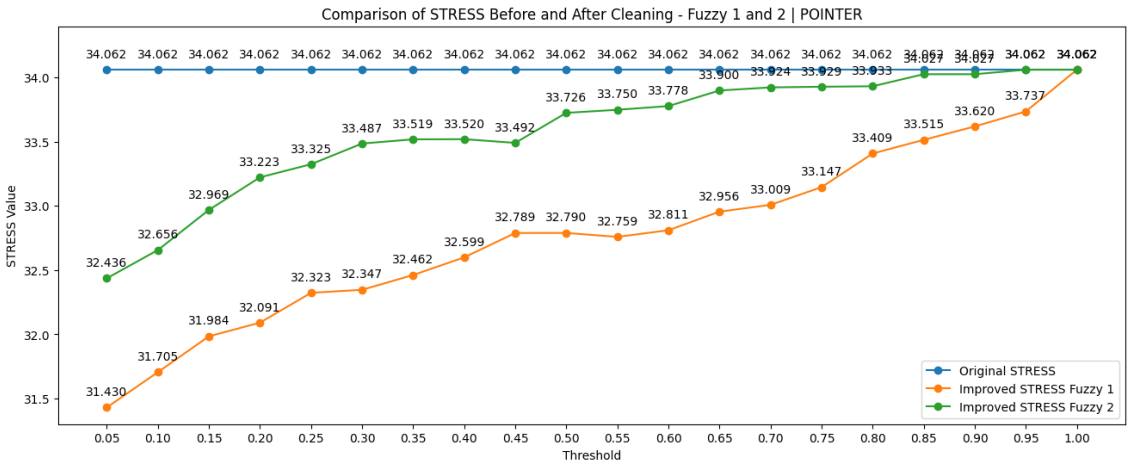

Figura 16 – Evolución de la métrica STRESS par cada valor de umbral en la base de datos Pointer

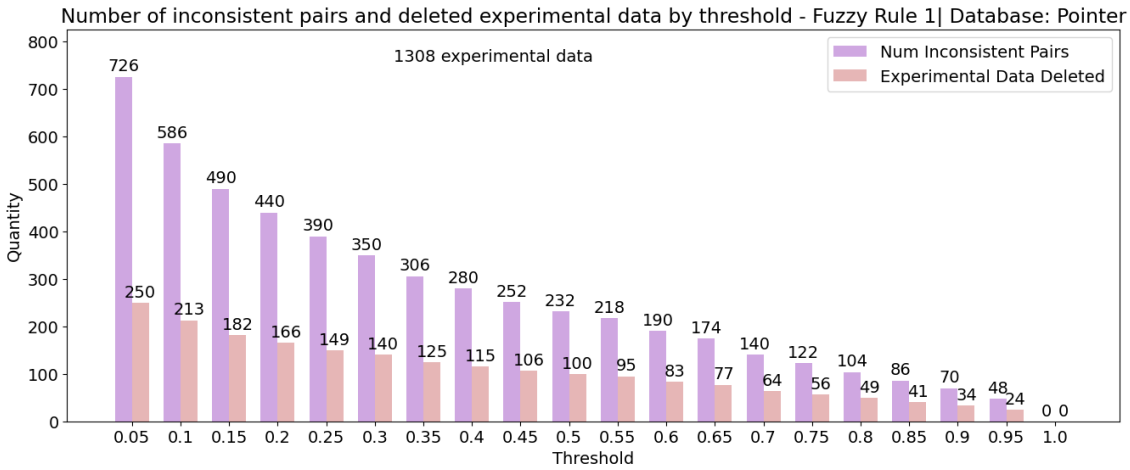

Figura 17 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 1. Base de datos Pointer

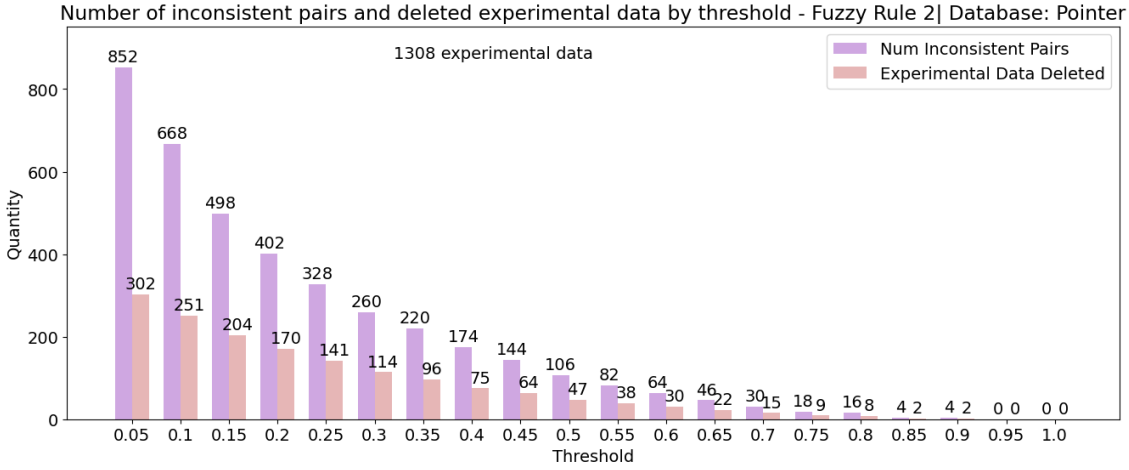

Figura 18 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 2. Base de datos Pointer

De cada pareja inconsistente, se elimina el que tenga una mayor inconsistencia: aquel dato experimental, de la pareja de pares inconsistentes cuyo diferencia entre ΔV y ΔE sea mayor.

Dentro del contexto de la base de datos Pointer, hemos realizado un análisis detallado para determinar el valor óptimo de ganancia en STRESS para cada umbral especificado en las reglas Difusa 1 y Difusa 2. En ambos casos, hemos encontrado que el umbral de 0.6 es un candidato adecuado. Esto significa que, al eliminar datos experimentales por debajo de este umbral, logramos una mejora sustancial en el nivel de STRESS.

Sin embargo, es importante destacar que, al reducir aún más el umbral a 0.55 y luego a 0.5, la mejora en el nivel de STRESS es mínima. Esto sugiere que los datos eliminados en estos niveles de umbral no difieren significativamente en inconsistencia con respecto a los que permanecen en la base de datos. Para la regla 1, eliminaríamos menos del 6.5% de los datos (Figura 17), y para la regla 2, eliminaríamos menos del 2.5% (Figura 18).

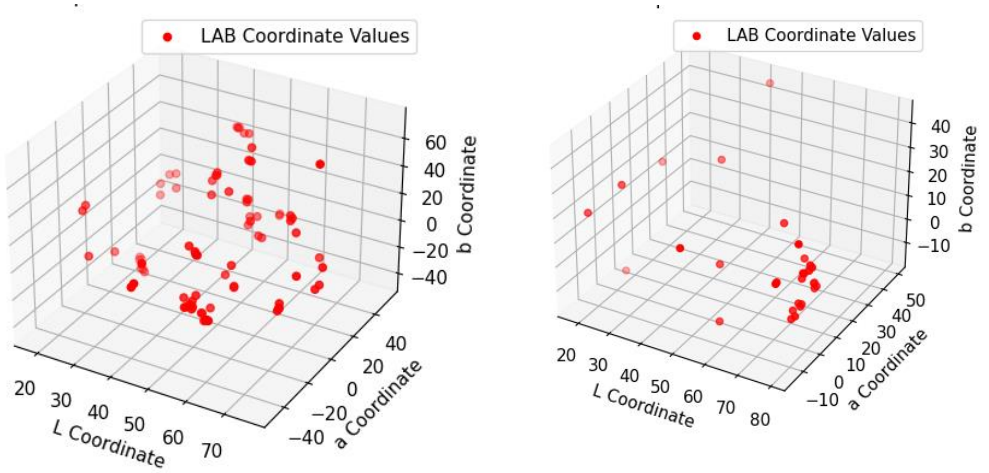

*Figura 19 – Coordenadas L*a*b de los datos eliminados para un umbral de 0.6 para las reglas 1 (izq.) y 2(der.) para la base de datos Pointer*

En la Figura 19, se hace evidente una agrupación de datos experimentales con similitudes en la representación de la derecha, mientras que en la representación de la izquierda, no se distingue ninguna agrupación discernible.

MMB

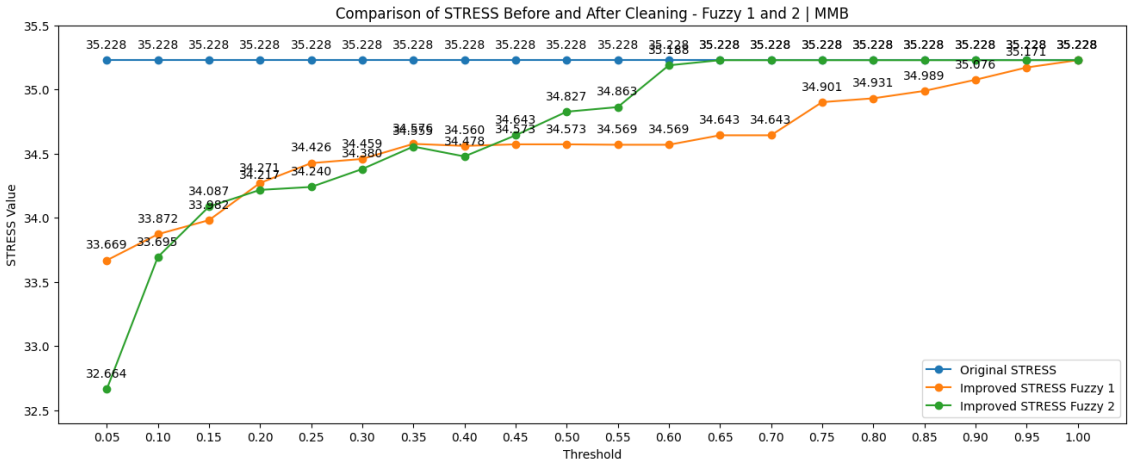

Figura 20 – Evolución de la métrica STRESS par cada valor de umbral en la base de datos MMB

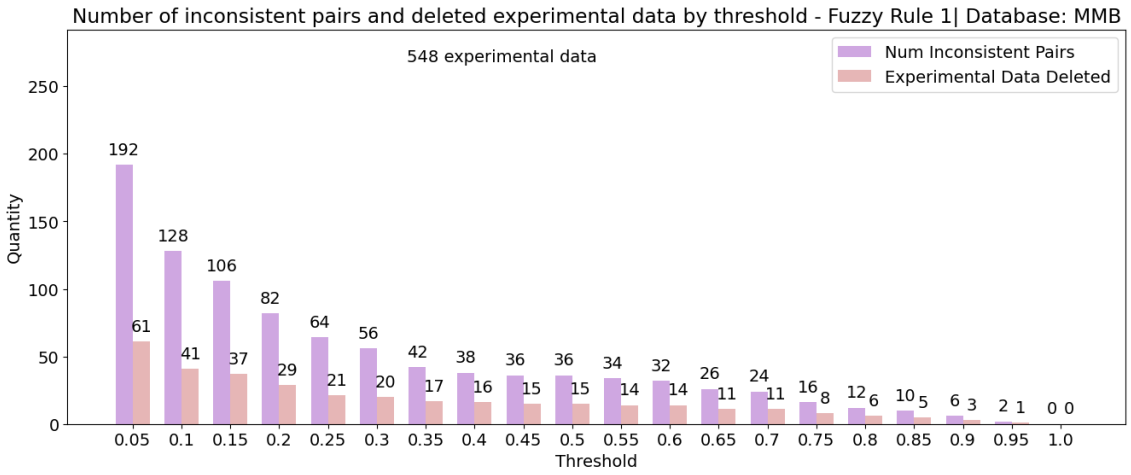

Figura 21 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 1. Base de datos MMB

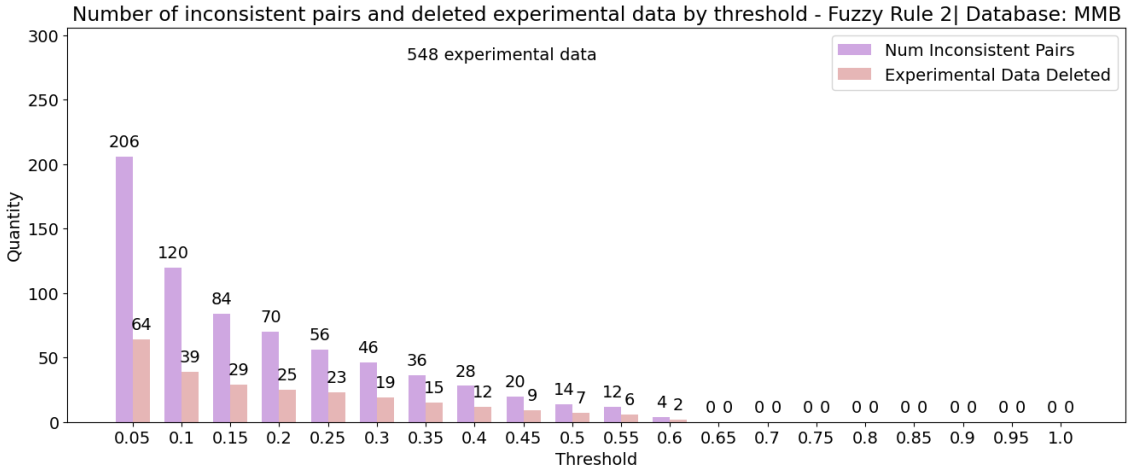

Figura 22 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 2. Base de datos MMB

En la base de datos MMB, como se muestra en la Figura 20, hemos identificado que un umbral de 0.70 o 0.65 es apropiado para la regla 1 y un umbral de 0.4 para la regla 2.

Para la regla 1 con un umbral de 0.70 o 0.65, al eliminar 11 datos (Figura 20) observamos una mejora considerable en el nivel de STRESS. Por otro lado, para la regla 2 con un umbral de 0.4, se produce un minimo en el valor del STRES (Figura 20) eliminando en cualquiera de los dos casos menos de un 2.5% de los datos (Figuras 21 y 22).

Este análisis personalizado resalta la importancia de adaptar los umbrales según las características de cada base de datos para obtener resultados precisos y significativos en nuestros estudios.

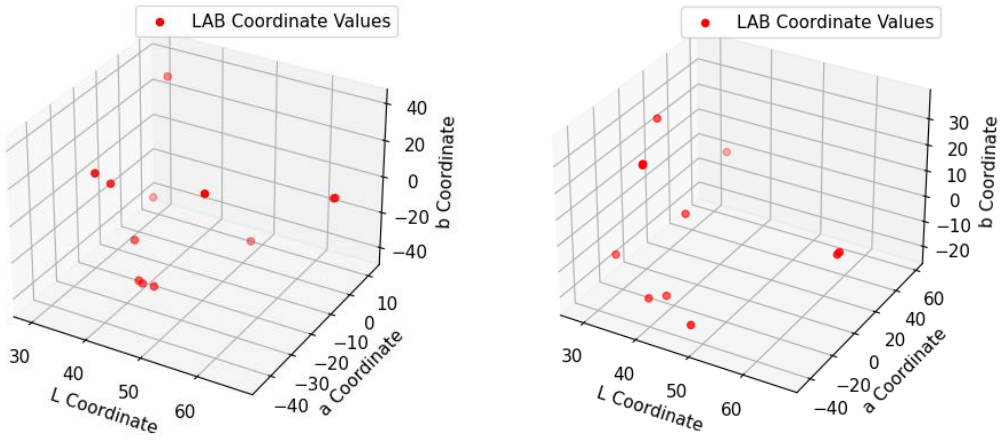

*Figura 23 – Coordenadas L*a*b de los datos eliminados para la regla 1 con umbral 0.7 (izq.) y 0.4 para la regla 2 (der.) para la base de datos MMB*

En la Figura 23, no se percibe una agrupación clara de datos, ya que la eliminación de estos no parece seguir un patrón discernible. Resulta difícil identificar un patrón evidente en la eliminación de estos datos.

lcamDIN99WDC

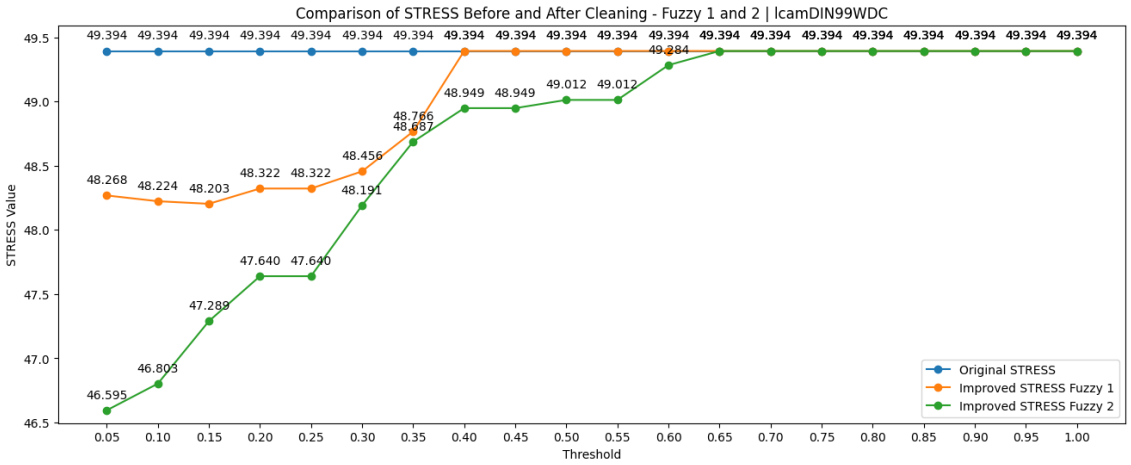

Figura 24 – Evolución de la métrica STRESS par cada valor de umbral en la base de datos lcamDIN99WDC

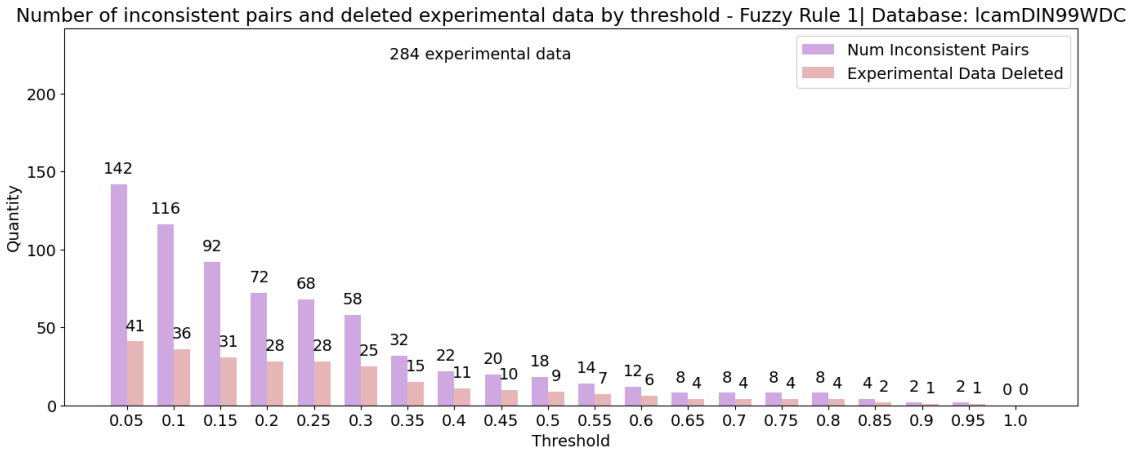

Figura 25 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 1. Base de datos IcamDIN99WDC

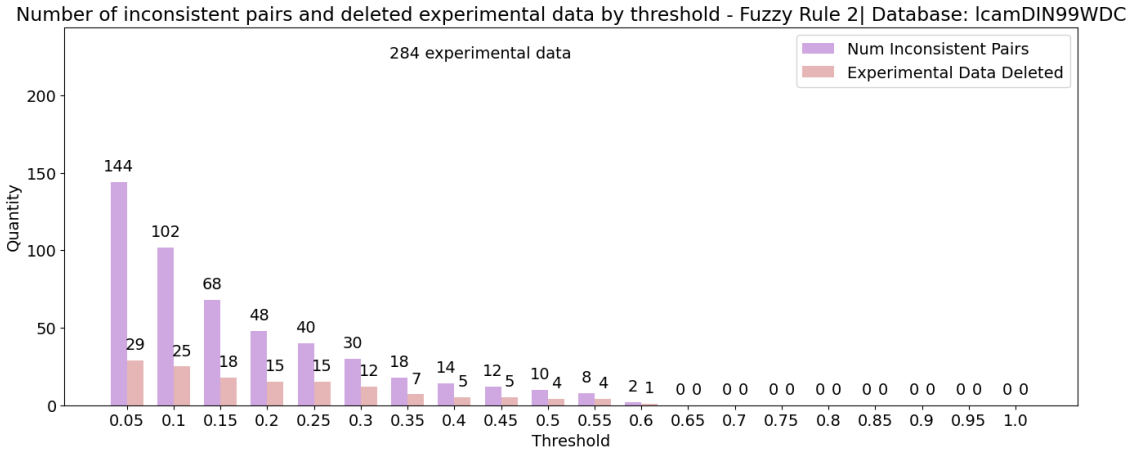

Figura 26 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 2. Base de datos IcamDIN99WDC

Para la base de datos IcamDIN99WDC, como se muestra en la Figura 24, hemos identificado que un umbral de 0.55 es apropiado para la regla 1 y un umbral de 0.4 para la regla 2.

Para la regla 1 con un umbral de 0.55, al eliminar 7 datos, un 2.5%, (Figura 25) observamos una mejora considerable en el nivel de STRESS. Por otro lado, para la Regla 2 con un umbral de 0.3, se consigue reducir valor de STRESS (Figura 24) eliminando 12 datox experimentales que suponen el un 4.2% (Figura 26).

BIGC

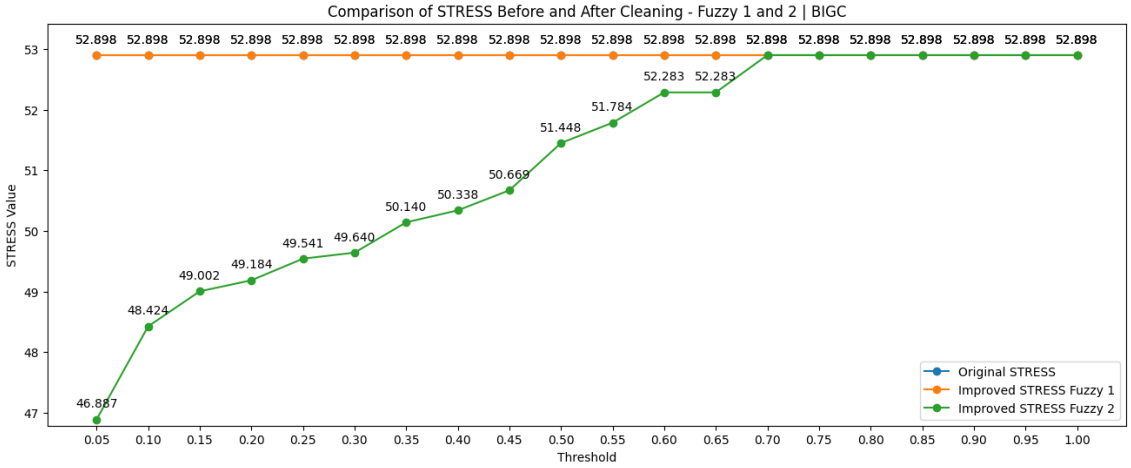

Figura 27 – Evolución de la métrica STRESS par cada valor de umbral en la base de datos BIGC

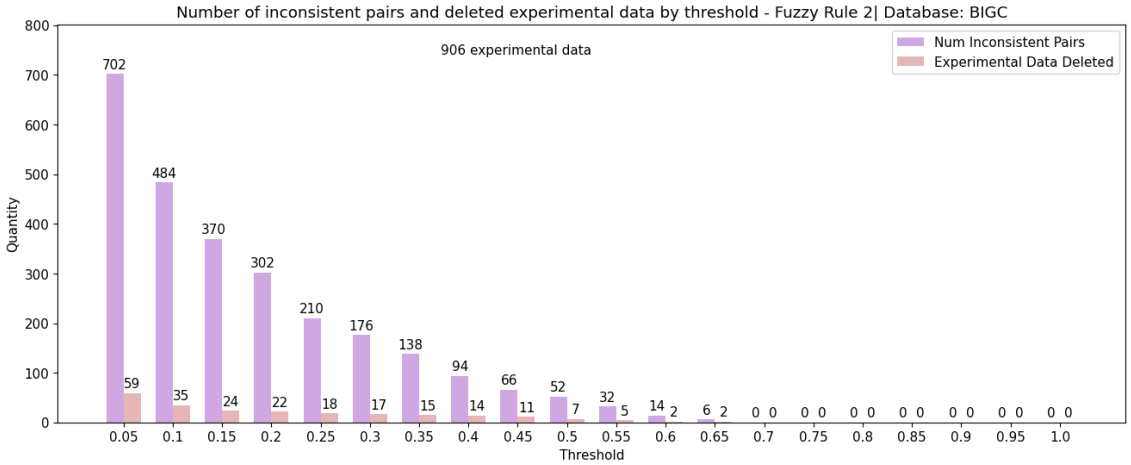

Figura 28 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa2. Base de datos BIGC

En lo que respecta a la base de datos BIGC, la regla 1 no logra detectar ningún par inconsistente para ningun valor de umbral, por lo que el valor de STRESS coincide con el valor del STRESS original (Figura 27). Por contraste, la regla 2 si que detecta pares inconsistentes en un amplio rango de valores de umbral. El valor de umbral óptimo para esta base de datos es 0.3, ya que en este punto se registra una mejora sustancial en el STRESS eliminando unicamente 17 datos, menos del 2%, como se muestra en la Figura 30. En otras palabras, la regla 2 es altamente efectiva para identificar y corregir pares inconsistentes en esta base de datos BIGC.

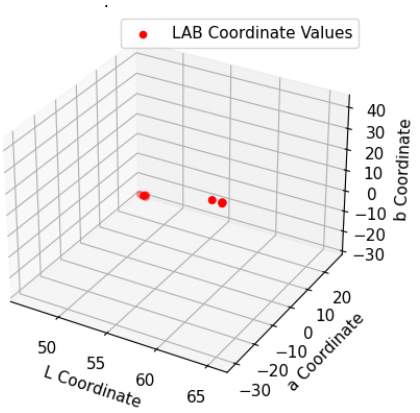

*Figura 29 – Coordenadas L*a*b de los datos eliminados para la regla 2 con umbral de 0.3 para la base de datos BIGC*

En la Figura 29, podemos observar que el numero de datos eliminados es mucho menor que el del resto de bases de datos, debido a la naturaleza de la base de datos.

Qiao, RIT_COM_V5 y WangHan

Para base de datos Qiao, RIT_COM_V5 y WangHan no se detecta ningún par de datos inconsistentes para ninguna de las dos reglas ni para ningún valor de umbral. Todas estas bases de datos, con un tamaño pequeño de datos experimentales (44, 249 y 266 respectivamente) tienen una gran consistencia.

RIT_DiPoint_Individual

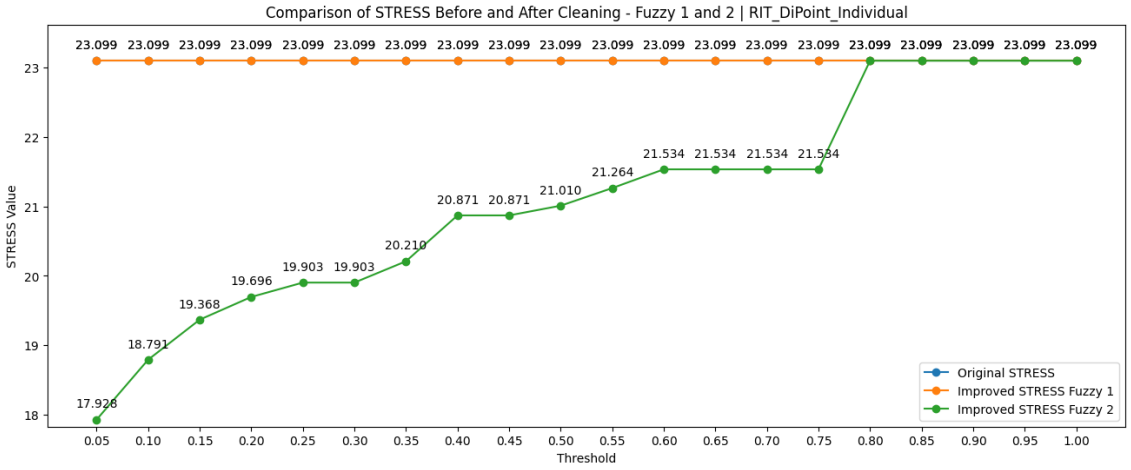

Figura 30 – Evolución de la métrica STRESS par cada valor de umbral en la base de datos RIT_DiPoint_Individual

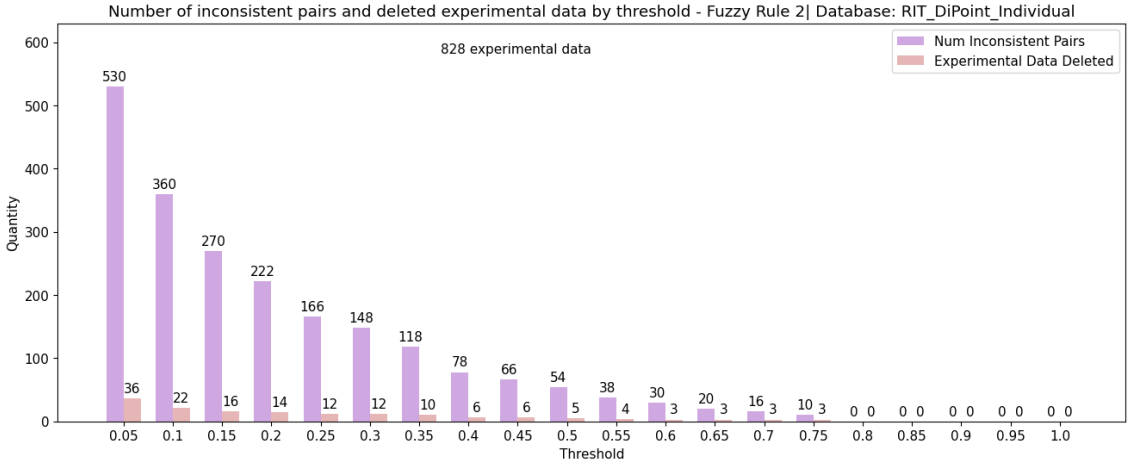

Figura 31 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 2. Base de datos RIT_DiPoint_Individual

En el caso de la base de datos RIT_DiPoint_Individual, como se puede ver en la Figura 30, la regla 1 no identifica pares inconsistentes en ningún valor de umbral evaluado. Por otro lado, la regla 2 demuestra una capacidad destacable para mejorar el conjunto de datos. Se observa que el valor de umbral óptimo es 0.75. Eliminando unicamente los 3 datos experimentales detectados en dicho valor de umbral, se consigue mejorar notablemente el valor del STRESS

NCSU

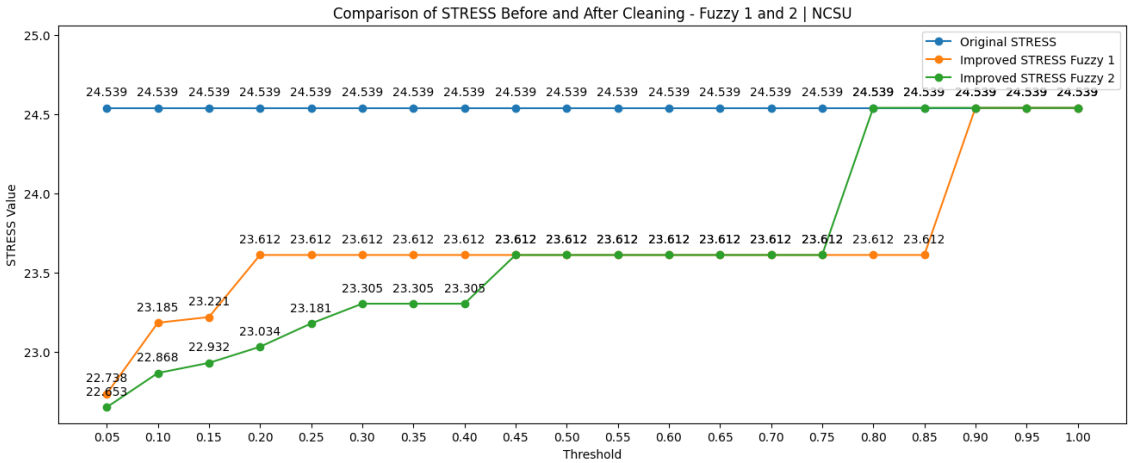

Figura 32 – Evolución de la métrica STRESS par cada valor de umbral en la base de datos NSCU

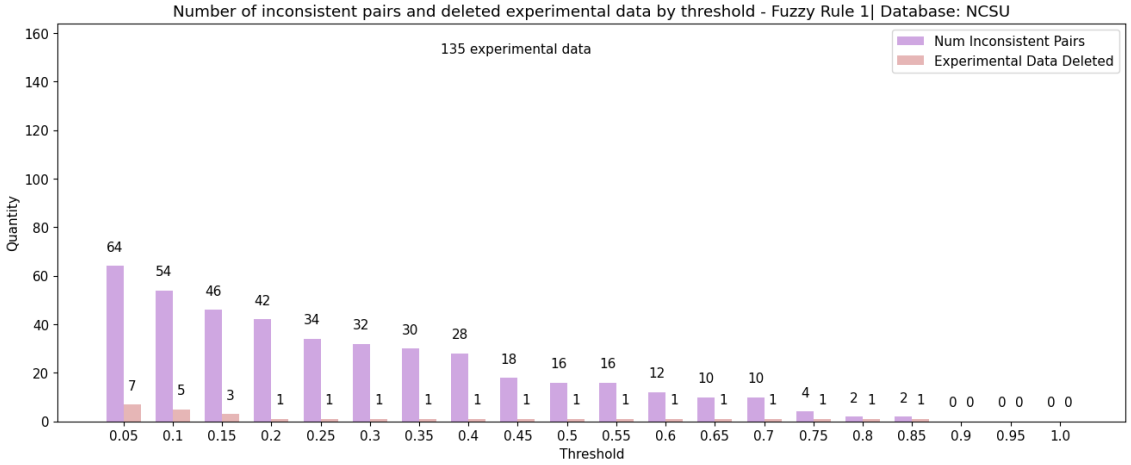

Figura 33 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 1. Base de datos NCSU

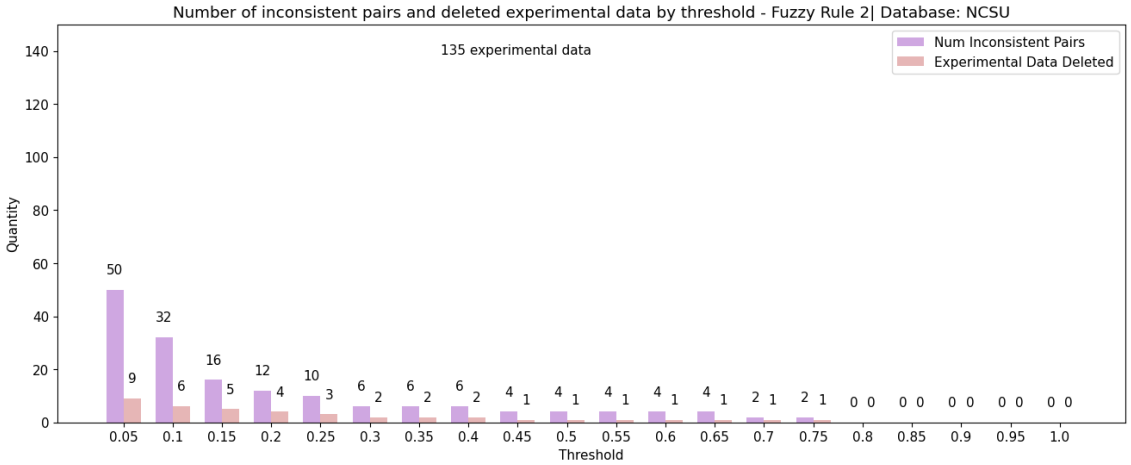

Figura 34 – Datos experimentales detectados y eliminados par cada umbral. Regla difusa 1. Base de datos NCSU

En la Figura 32 se muestra la mejora del STRESS en la base de datos NCSU ambas reglas muestran comportamientos similares eliminando unicamente un solo dato experimental. En particular, la regla 1 logra una gran mejora con un umbral de 0.85. Por otro lado, la regla 2 alcanza una mejora similar con un umbral ligeramente inferior de 0.75. Como se puede ver en las Figuras 33 y 35, este dato experimental supone menos del 1% del conjunto de datos de la base de datos. Ambas graficas de STRESS presentan una evolucion muy plana despues de eliminar estos datos, hacia valores de umbral más bajos. Estos resultados indican que ambas reglas son efectivas para identificar y corregir pares inconsistentes en la base de datos NCSU.

A pesar de haber generado las gráficas 3D correspondientes a esta base de datos, resulta evidente que al contar únicamente con dos puntos de datos, estas representaciones carecen de la suficiente información para ofrecer un análisis significativo. Por esta razón, he decidido no incorporarlas en el documento, ya que no aportarían valor adicional a la comprensión del tema que estamos tratando.

En términos globales, estos resultados destacan la importancia de ajustar los umbrales de eliminación de datos inconsistentes de manera específica para cada regla y base de datos, ya que no existe un valor universal que funcione para todos los casos. Este enfoque personalizado garantiza que los datos restantes sean más coherentes y confiables para su posterior análisis, lo que conduce a resultados más precisos y significativos en nuestros estudios.

8. CONCLUSIONES

Este estudio muestra la necesidad del análisis de la consistencia de datos perceptuales de diferencia de color, subrayando la importancia de adaptar este enfoque a la naturaleza específica de cada base de datos. Al hacerlo, hemos logrado obtener una comprensión más precisa y esclarecedora de los datos.

Como resultado de este análisis, hemos observado mejoras significativas en la consistencia de datos en múltiples bases de color respecto a su estado inicial. Estos hallazgos resaltan la necesidad de un enfoque más adaptativo y orientado a los datos en la gestión de información perceptual de diferencia de color.

El análisis realizado en varias bases de datos revela que la detección y corrección de pares inconsistentes mediante las reglas difusas 1 y 2 es un proceso altamente dependiente de la naturaleza de los datos. Los resultados indican que los valores óptimos de umbral varían significativamente según la base de datos y la regla aplicada.

En general, se observa que la aplicación de umbrales apropiados puede mejorar la calidad de los datos al eliminar las inconsistencias detectadas. Sin embargo, es crucial encontrar un equilibrio, ya que umbrales demasiado bajos pueden llevar a la eliminación excesiva de datos útiles, mientras que umbrales demasiado altos pueden no detectar las inconsistencias.

No obstante, en el caso de las bases de datos Qiao, RIT_COM_V5 y WangHan, no se ha logrado observar una mejora, ya que no se identificaron datos experimentales que difirieran de manera significativa del conjunto general de la base de datos.

Estos hallazgos destacan la importancia de adaptar las estrategias de análisis y corrección de datos a las características específicas de cada conjunto de datos. El uso de estas reglas difusas junto con umbrales óptimos puede ser una herramienta valiosa para mejorar la calidad de las bases de datos y garantizar la fiabilidad de los análisis posteriores.

9. FUTURAS LÍNEAS DE TRABAJO

Algunas posibles líneas de trabajo futuras que podrían surgir de este estudio incluyen:

- **Mejora de Métodos de Detección de Inconsistencias**: Investigar y desarrollar métodos más precisos y eficientes para la detección de inconsistencias en datos de diferencia de color. Esto podría incluir otros métodos de generación de reglas, o de aprendizaje máquina, para la detección

- **Validación Experimental Adicional**: Realizar experimentos adicionales para validar los resultados obtenidos en este estudio. Esto podría implicar la realización de pruebas de percepción de color en un entorno controlado para confirmar las inconsistencias detectadas y evaluar la relevancia práctica de estas inconsistencias.

- **Desarrollo de Nuevas Métricas de Consistencia:** Explorar y proponer nuevas métricas y métodos de evaluación de la consistencia de datos de diferencia de color. Estas métricas podrían abordar aspectos específicos de la percepción del color que no se consideraron en este estudio.

- **Optimización de Algoritmos Difusos**: Refinar y optimizar los algoritmos difusos utilizados en este estudio para mejorar su eficiencia y capacidad de detección de inconsistencias.

- **Exploración de Diferentes Espacios de Color**: Investigar cómo los resultados varían al utilizar diferentes espacios de color en lugar de CIELAB, lo que podría revelar información adicional sobre la percepción del color.

- **Aplicación en Industrias Específicas:** Aplicar los métodos y hallazgos de este estudio en industrias específicas, como la industria textil o la industria de la pintura, para evaluar la consistencia del color en productos y procesos reales.

- **Estudios de Percepción del Color:** Realizar estudios más detallados sobre la percepción del color humana para comprender mejor cómo se perciben las diferencias de color en diferentes contextos y por diferentes personas.

Estas líneas de trabajo futuras podrían contribuir al avance de la comprensión y gestión del color en diversas aplicaciones industriales y científicas.

10. REFERENCIAS BIBLIOGRÁFICAS

1. M. Melgosa, E. Hita, J. Romero, and L. Jiménez del Barco, “Some classical color differences calculated with new formulas,” J. Opt. Soc. Am. A 9, 1247–1254 (1992).
2. International Commission on Illumination (CIE), Parametric Effects in Colour Difference Evaluation (CIE Central Bureau, 1993).
3. Martinez-Uriegas E. Fundamentos de colorimetría. Color Res Appl. 2003;28(6).
4. Kuehni RG. COLOR-TOLERANCE DATA AND THE TENTATIVE CIE 1976 L*a*b* FORMULA. J Opt Soc Am. 1976;66(5).
5. Huang M, Xi Y, Pan J, He R, Li X. Colorimetric Observer Categories for Young and Aged Using Paired-Comparison Experiments. IEEE Access. 2020;
6. Morillas S, Gómez-Robledo L, Huertas R, Melgosa M. Method to determine the degrees of consistency in experimental datasets of perceptual color differences. Journal of the Optical Society of America A. 2016;33(12).
7. Morillas S, Gomez-Robledo L, Huertas R, Melgosa M. Fuzzy analysis for detection of inconsistent data in experimental datasets employed at the development of the CIEDE2000 colour-difference formula. J Mod Opt. 2009;56(13).
8. Latorre-Carmona P, Huertas R, Pedersen M, Morillas S. Proposal of a new fidelity measure between computed image quality and observers quality scores accounting for scores variability. J Vis Commun Image Represent. 2023;90.
9. CIE 217:2016, Recommended Method for Evaluating the Performance of Colour-Difference Formulae (CIE Central Bureau, 2016).
10. M. R. Luo, G. Cui, and B. Rigg, “The development of the CIE 2000 colour-difference formula: CIEDE2000,” Color Res. Appl. 26, 340–350 (2001).
11. M. Huang, G. Cui, M. Melgosa, M. Sánchez-Marañón, C. Li, M. R. Luo, and H. Liu, “Power functions improving the performance of colordifference formulas,” Opt. Express 23, 597–610 (2015).
12. Publications Briefly Mentioned: CIE 142-2001, improvement to industrial colour-difference evaluation. Color Res Appl. 2002;27(1).
13. Rouvray DH. Fuzzy sets and fuzzy logic: Theory and applications. Endeavour. 1996;20(1).
14. García PA, Huertas R, Melgosa M, Cui G. Measurement of the relationship between perceived and computed color differences. Journal of the Optical Society of America A. 2007;24(7).