# Metodología de visualización interactiva de datos de alta dimensión a partir de un modelo intuitivo de reducción de dimensión


**DIEGO FERNANDO PEÑA UNIGARRO**

**UNIVERSIDAD DE NARIÑO**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA**
**SAN JUAN DE PASTO**
**2017**

# Metodología de visualización interactiva de datos de alta dimensión a partir de un modelo intuitivo de reducción de dimensión

**DIEGO FERNANDO PEÑA UNIGARRO**

**TRABAJO DE GRADO PARA OPTAR POR EL TITULO DE INGENIERO ELECTRÓNICO**

**DIRECTOR**

**PhD. DIEGO HERNÁN PELUFFO ORDÓÑEZ**

**INGENIERO ELECTRÓNICO**

**UNIVERSIDAD DE NARIÑO**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA**
**SAN JUAN DE PASTO**
**2017**

## NOTA DE RESPONSABILIDAD

“La Universidad de Nariño no se hace responsable por las opiniones o resultados obtenidos en el presente trabajo y para su publicación priman las normas sobre el derecho de autor.”

Acuerdo 1. Artículo 324. Octubre 11 de 1966, emanado del honorable Consejo Directivo de la Universidad de Nariño.

**NOTA DE ACEPTACIÓN:**

_____
_____
_____
_____
_____
_____

_____
Firma del presidente del jurado

_____
Firma del jurado

_____
Firma del jurado

San Juan de Pasto, 3 de agosto de 2017

| Universidad de Nariño | FACULTAD DE INGENIERIA<br>DEPARTAMENTO DE ELECTRÓNICA<br>ACTAS SOCIALIZACIONES | Código: DEL-FOA-FR-03<br>Página: 1 de 6<br>Versión: 1<br>Vigente a Partir de 11/06/2009 |
|---|---|---|

**ACTA No. 002**

En San Juan de Pasto, a los dos dias del mes de marzo de 2017, se reunieron los Jurados Calificadores del Trabajo de Grado: "METODOLOGIA DE VISUALIZACIÓN INTERACTIVA DE DATOS DE ALTA DIMENSIÓN A PARTIR DE UN MODELO INTUITIVO DE REDUCCIÓN DE DIMENSIÓN" modalidad Trabajo de Investigación, los Ingenieros JUAN CARLOS ALVARADO PEREZ y JAVIER REVELO FUELAGÁN, elaborado por el estudiante DIEGO FERNANDO PEÑA UNIGARRO del Programa de Ingeniería Electrónica, bajo la dirección del Ingeniero DIEGO HERNÁN PELUFFO ORDÓÑEZ, actúa como Secretaria Ingeniera DORIS MARTINEZ RICAURTE..

ORDEN DEL DIA

1. Lectura de los objetivos consignados en el Trabajo de Grado
2. Lectura del concepto emitido por el Jurado Calificador
3. Lectura de la Calificación del Trabajo de Grado otorgada por el Jurado Calificador
4. Sustentación
5. Calificación

1. Se da lectura por parte de la Secretaria Académica de los documentos anunciados en el Primer y Segundo numeral del Orden del Dia.

   Acto seguido, se procede a la sustentación y Calificación Definitiva del Trabajo de Grado.

A. Calificación: Sustentación y Seminario:

Sr. (Srta). **DIEGO FERNANDO PEÑA UNIGARRO**

Jurado No. 1 40

Jurado No. 2 40

Promedio 40

Calificación del Desarrollo y Presentación del Trabajo de Grado:

Jurado No. 1 60

Jurado No. 2 60

Promedio 60

## DEDICATORIA

*“A toda mi familia por ser la base fundamental de mis proyectos de vida, especialmente a mi padre Edgar Peña, a mi madre Sandra Unigarro y a mi hermano Andrés Peña por ser un apoyo incondicional en este proceso y por acompañarme especialmente en los momentos más difíciles. A mis primos Carlos Rodríguez y Francisco Quinchoa, por cuidarme como a un hermano y ayudarme a crecer como persona. A mis abuelos, Luz Rodríguez, María Elena Paz y Jesús Unigarro, por su amor incondicional y sus enseñanzas que han sido muy importantes para alcanzar todas mis metas”.*

DIEGO FERNANDO PEÑA UNIGARRO

**AGRADECIMIENTO**

*“Agradezco a mi familia por ayudarme a cumplir mis objetivos y metas, a mis amigos por acompañarme y ser mi gran apoyo en todo este proceso académico, a mis profesores por compartir sus conocimientos y experiencias. Especialmente quiero agradecer a mi asesor el profesor Diego Hernán Peluffo Ordoñez por brindarme su confianza, amistad y acompañamiento para hacer este nuevo logro posible, además su paciencia y motivación han sido fundamentales para mi formación como investigador, igualmente quiero agradecer al profesor Wilson Olmedo Achicanoy Martínez por su apoyo y enseñanzas que fueron un aporte muy importante para el desarrollo de este trabajo”.*

DIEGO FERNANDO PEÑA UNIGARRO

## RESUMEN

Una consecuencia de la sobrecarga de información en la actualidad es que la capacidad tecnológica para recolectar, comunicar y guardar grandes volúmenes de información está incrementando más rápido que las capacidades humanas de análisis, dejando una brecha entre los usuarios y el conocimiento existente en las bases de datos. Esta brecha ha motivado el desarrollo de maneras gráficas de representar y analizar datos de alta dimensión, entendiéndose dimensión como el número de atributos o características que posee una muestra u objeto. Particularmente, en este trabajo se propone desarrollar una interfaz gráfica basada en un modelo cromático que permite la combinación intuitiva de métodos de reducción de dimensión (RD), con el fin de hacer la visualización de datos más inteligible para los usuarios y, por lo tanto, un usuario (no necesariamente experto) podrá fácilmente y de manera intuitiva hacer uso de métodos de reducción de dimensión.

El modelo interactivo propuesto está basado en el espacio de color RGB en donde los valores de intensidad en puntos dentro de una superficie formarán un vector de ponderación que, a su vez, definirá el grado con que un método en particular es utilizado. De este modo, un usuario podrá aplicar intuitivamente una mezcla de métodos de RD a través de la exploración del modelo cromático con el fin de obtener una representación de los datos que pueda brindar información que no fue detectada con anterioridad. Los métodos de RD seleccionados son implementados a través de aproximaciones kernel de forma que, la combinación de métodos de RD se verá reflejada en una combinación lineal de matrices kernel.

## ABSTRACT

Nowadays, a consequence of data overload is that world's technology capacity to collect, communicate, and store large volumes of data is increasing faster than human analysis skills. Such an issue has motivated the development of graphic ways to represent and analyze high-dimensional data, where dimensionality is defined as the number of measurements or observations that represents an object or instance. Particularly, in this work, we propose a graphical interface that allow the combination of dimensionality reduction (DR) methods using a chromatic model to make data visualization more intelligible for humans. In consequence, (even non-expert) users can intuitively either select a concrete DR method or carry out a mixture of methods.

This work presents an intuitive model that allows the combination of three dimensionality reduction methods taking into account properties such as interactivity and controllability. The model is based on the RGB color space where each primary color (red (R), green (G), blue (B)) represents a particular method and the full range of colors derived from the combination is reflected in the mixture of dimensionality reduction methods. Conventional DR methods are implementing by kernel approximations, which there are combined trought a lineal combination of kernel matrices.

# TABLA DE CONTENIDO

## LISTA DE FIGURAS

## LISTA DE ANEXOS

## GLOSARIO

**Big Data:** Big Data, macrodatos o datos masivos es un concepto que hace referencia al almacenamiento de grandes cantidades de datos y a los procedimientos usados para encontrar patrones repetitivos dentro de dichos datos.

**Dimensión:** En términos generales, la dimensión de una base de datos es definida como la cantidad de mediciones, características o atributos que tiene cada objeto o muestra.

**Reducción de dimensión:** Las técnicas de reducción de la dimensión tienen por objetivo final condensar la información de un conjunto de variables en un nuevo conjunto de variables (de menor número que el anterior), con la menor pérdida de información posible.

**Espacio embebido:** En este trabajo espacio embebido hace referencia al espacio de baja dimensión resultante cuando un método de reducción de dimensión es aplicado a una base de datos de alta dimensión.

**Diagrama de dispersión:** Un diagrama de dispersión o gráfica de dispersión o gráfico de dispersión es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos o tres variables para un conjunto de datos.

**Interpolación baricéntrica:** La interpolación es un proceso por el cual se define un valor en un punto cualquiera a partir de los valores conocidos en algunos puntos dados (en este caso vértices de un triángulo).

**Imagen RGB:** Son aquellas imágenes que están representadas a través de los colores primarios rojo, verde y azul. Este tipo de imágenes son las más adecuadas para ser mostradas en monitores y que, finalmente, serán impresas en impresoras de papel fotográfico.

**Pixel:** El píxel puede definirse como la más pequeña de las unidades homogéneas en color que componen una imagen de tipo digital.

**Resolución de intensidad:** En escala de grises la resolución de intensidad de una imagen se define como los niveles de grises que una imagen puede tener.

**Resolución de espacial:** Se define como y el número de filas (el alto) y columnas (el ancho) que tiene una imagen, es decir, el número de pixeles.

# INTRODUCCIÓN

Los avances en las capacidades de recolección y almacenamiento de datos han dejado una sobrecarga de información que excede en mucho las capacidades humanas de análisis, y teniendo en cuenta que bases de datos pueden ser generadas en diferentes áreas del conocimiento surge la necesidad de buscar técnicas y metodologías que permitan representar los datos de una manera más inteligible para el ser humano [1]. La transformación de datos de alta dimensión en datos de una dimensión menor (datos embebidos) que preserve la información original lo más adecuadamente posible, es un área de investigación ampliamente estudiada [2], [3], dada su habilidad de reducir el costo computacional y/o mejorar el desempeño de tareas de reconocimiento de patrones y de visualización de la información [4], [5]. A pesar de la existencia de herramientas que han alcanzado altos indicadores en términos de rendimiento computacional, exploración y representación de datos de alta dimensión, éstas carecen de propiedades importantes como interactividad y controlabilidad, lo que hace necesario la intervención de un experto con conocimientos a priori con el fin de aplicar técnicas y metodologías que permitan procesar de la mejor manera los datos e interpretar los resultados obtenidos, que en algunos casos pueden ser ambiguos y difíciles de analizar aun para el usuario experto [4], [6]. Cabe resaltar que debido a la naturaleza interdisciplinar del manejo de bases de datos y las herramientas de análisis demasiado complejas y abstractas para el usuario inexperto, se estaría formando una brecha entre los usuarios y el conocimiento disponible en una base de datos [3]. La reducción de dicha brecha es la premisa en la cual esta investigación está basada.

Este trabajo toma conceptos y técnicas existentes en las áreas de reducción de dimensión y visualización de la información con el fin de que un usuario pueda interactuar directamente con una base de datos y conseguir representaciones gráficas que le permitan obtener conclusiones y tomar decisiones [6], [7]. Para este fin, se presenta un modelo intuitivo que permite la combinación de tres diferentes métodos de RD de manera que los datos embebidos puedan ser obtenidos de manera controlada e interactiva. El modelo propuesto está basado en el espacio de color RGB, donde cada color primario (rojo (R), verde (G) y azul (B)) representa un método de RD en particular mientras que todo el rango de colores derivado de la combinación de estos colores se verá reflejado en la mezcla de métodos de RD los cuales, son implementados a través de aproximaciones kernel. Finalmente, la matriz kernel alimenta un algoritmo generalizado de análisis de componentes principales (KPCA) [6], [8] . El beneficio de esta aproximación es que el usuario podrá utilizar métodos de RD sobre los datos, incluso sin tener conocimientos acerca de los fundamentos teóricos detrás de ellos, de esta forma el usuario controlará los datos embebidos resultantes mediante la exploración intuitiva del modelo cromático. En

otras palabras, el modelo interactivo propuesto utiliza los puntos de color dentro de la superficie los cuales definen el grado o nivel con que un método de RD (matriz kernel) es utilizado.

Este enfoque permite evaluar visualmente el comportamiento de los datos originales cuando se aplica un método de RD en particular o una combinación de métodos de manera que el usuario sea capaz de escoger la representación que más se ajuste a sus necesidades. El modelo cromático propuesto en este trabajo es evaluado usando tres métodos clásicos de reducción de dimensión : ***Locally Linear Embedding*** (LLE) [9], ***Classical Multidimensional Scaling*** (CMD) [10], [11] y ***Laplacian Eigenmaps*** (LE) [12]. Además, los experimentos son desarrollados con cuatro bases de datos: Dos bases de datos reales (imágenes de objetos - COIL 20 e imágenes de dígitos -MNIST) y dos bases de datos artificiales (Cascarón esférico en 3D y el rollo suizo). El desempeño de los métodos de reducción de dimensión es evaluado mediante una versión escalada de la tasa promedio del acuerdo entre los $k$-vecinos más cercanos como se explica en [13].

# 1. DESCRIPCION DEL PROBLEMA

## 1.1. PLANTEMIENTO DEL PROBLEMA

Los avances electrónicos e informáticos presentes en la actualidad han tenido un gran impacto en la forma cómo se maneja la información, debido a que terabytes de información representados en bases de datos son generados a diario [10], [14]. Si bien las capacidades de almacenamiento están avanzando de forma acelerada para acaparar los grandes volúmenes de información que se generan actualmente, las capacidades para análisis y representación de datos están evolucionando de una manera más lenta, dejando una brecha entre el usuario y la información [1], [7], [10]. De esta manera la tarea de presentar y/o representar datos de forma comprensible e intuitiva es un área de estudio que ha adquirido importancia en los últimos años, debido a que en muchas ocasiones los datos no son fácilmente interpretables lo que hace necesario la intervención de un experto con conocimientos a priori [3] [14]. Uno de los mayores problemas que enfrenta la representación de los datos es la alta dimensión debido a que la percepción humana está limitada a tres dimensiones, e inclusive, en tres dimensiones los datos pueden ser ambiguos y no muy claros [2], [3], [14]. En términos generales, la dimensión en una base de datos está definida por el número de mediciones o atributos que posee cada muestra u objeto, en consecuencia, si un registro tiene más de tres mediciones no podrá ser representado de manera tradicional en un diagrama de dispersión en el plano cartesiano, dificultando el proceso de descubrimiento de nuevo conocimiento en una base de datos [3]. La visualización interactiva de información permite a un ser humano el desarrollo de operaciones mentales que hacen posible el rápido acceso a grandes bases de datos (*Big Data*), convirtiendo a la máquina en una herramienta importante en el proceso de descubrimiento de patrones y nueva información que no fueron previstos con anterioridad [1], [7], [14].

La percepción de un patrón a menudo puede ser la base de una nueva visión, por lo tanto, la visualización facilita la comprensión de las características tanto a gran escala como a pequeña escala de los datos, por esta razón la visualización de datos es en muchos casos imprescindible, especialmente en las etapas de análisis donde se hacen hipótesis significativas sobre los datos. El descubrimiento de conocimiento en base de datos (DCBD) es en sí un proceso con el cual se extraen patrones a partir de conjuntos de datos con el fin de encontrar información de interés o formular predicciones de algún evento en particular. El DCBD ya ha sido ampliamente explorado y desarrollado, y tiene una gran diversidad de aplicaciones como la determinación de perfiles de clientes fraudulentos [15], descubrir una relación implícita que exista entre síntomas y enfermedades [16], [17] análisis de mercado,

ventas y soluciones a clientes [14], [15], entre otras aplicaciones, es por esto que se requiere que las técnicas de DCBD se sigan explorando y desarrollando.

En la actualidad existen herramientas que implementan técnicas que implican el pre procesamiento, el uso de métodos de minería de datos y/o visualización [18], [19], estas herramientas realizan el proceso de DCBD pero presentan diversos problemas entre los que se encuentra una respuesta aceptable pero con un gasto computacional alto y, además, con un bajo rendimiento en velocidad, o una respuesta rápida pero con resultados muy difícilmente interpretables, es decir, para poder analizar los resultados es necesario contar con un conocimiento previo o tener experiencia en dicha interpretación [20], también se considera que estas herramientas no mantienen una apropiada interactividad con el usuario debido a que no permite adecuar los parámetros existentes en los grandes volúmenes de datos y no toman a consideración las necesidades del usuario [21], [22]. Las herramientas de visualización de grandes volúmenes de datos se encargan de aplicar, métodos de minería de datos para la extracción de patrones en formas de reglas o funciones, o métodos de reducción de dimensión orientados a la visualización, pero sea cual sea el caso, realizar la visualización de dichas dimensiones reducidas puede llegar a ser algo abstracto [18], [19], [20]. Por lo tanto, se aprecia que existe la necesidad de que el usuario cuente con una herramienta interactiva que permita manipular los métodos de DCBD acorde a sus necesidades para obtener los resultados deseados y que éstos sean fácilmente interpretables sin un conocimiento previo de dichos métodos [6], [18]. Teniendo esto en cuenta, se puede decir que la representación visual cumple un papel muy importante dentro del procesamiento de datos cuando se fusiona y complementa con los procesos de minería de datos [7], [14].

## 1.2. JUSTIFICACIÓN

La capacidad de almacenar datos ha tenido un crecimiento exponencial en los años recientes, dificultando cada vez más el procesamiento y análisis de información [2], [10], [18]. La minería de datos como herramienta fundamental para encontrar conocimiento tiene una naturaleza interdisciplinar debido a que puede abarcar diferentes campos que usualmente son ajenos a metodologías del proceso DCBD, de este modo surge la necesidad de crear herramientas cercanas a los usuarios en el sentido de que sean de uso intuitivo y de fácil manejo [14], [23] [24]. En la actualidad, las tareas de reconocimiento de patrones y la minería de datos involucran generalmente bases de datos de alta dimensión que no pueden ser representados de manera tradicional en el diagrama de dispersión [3], por esta razón la reducción de dimensión se convierte en una herramienta importante para conocer la naturaleza de una base de datos en particular, de esta manera las técnicas RD pueden ayudar a un usuario a elegir metodologías apropiadas para realizar tareas de clasificación, predicción, extracción de nuevo conocimiento entre otras [6], [8], [4]. El principal objetivo de las técnicas de RD es extraer un espacio

de representación de baja dimensión en donde se encuentre la información más relevante eliminando datos redundantes, esto se verá reflejado en un costo computacional más bajo [3], [8], [25]. Las aplicaciones de técnicas de RD requieren a menudo de personal experto que sintonicé diferentes parámetros para obtener una buena representación en un espacio de baja dimensión, que involucra un incremento en tiempo y costos, para el desarrollo del proceso de descubrimiento de conocimiento en bases de datos. Por tanto, integrar métodos de RD con técnicas de visualización es una necesidad latente, que ha sido objeto de múltiples estudios, sin embargo, el diseño de una herramienta de análisis visual que sea intuitiva y que se adapte a las necesidades de un usuario en particular sigue siendo aún un problema abierto.

El desarrollo de este trabajo busca la unión de múltiples métodos de RD mediante una interfaz gráfica, la cual debe ser intuitiva, interactiva y pueda exponerle al usuario el efecto que tienen las mezclas de métodos de RD en una base de datos de entrada. Teniendo en cuenta que la metodología de visualización propuesta puede ser usada en cualquier área del conocimiento donde se generen datos, el usuario tiene que ser capaz de aplicar métodos de RD a través de la exploración del modelo cromático, además se busca reducir el costo computacional explorando un entorno de programación alternativo a MATLAB (Processing, en este caso).

## 1.3. CONTRIBUCIONES DE ESTA TESIS

La búsqueda de nuevos espacios que puedan representar de una mejor manera los datos de alta dimensión es un área de estudio que ha tomado fuerza en los últimos años debido que en muchos casos se necesitan más de tres mediciones para describir un objeto o fenómeno en particular, en consecuencia, las formas tradicionales de representación empiezan a perder utilidad si se tiene un número elevado de mediciones. Por lo tanto, se deberá recurrir a métodos de visualización más complejos con una elevada dificultad de interpretación para usuarios inexpertos. Este trabajo intenta disminuir la brecha entre los usuarios y las bases de datos mediante el desarrollo de una metodología interactiva de visualización de datos de alta dimensión.

La interfaz gráfica estará definida por medio de un modelo intuitivo que está basado en el espacio de color RGB que permite la combinación de métodos de RD en base al rango de colores generados en este espacio de color, por lo tanto, un usuario será capaz escoger un método de RD en particular o una mezcla de estos con el fin de observar patrones de interés que no fueron detectados con anterioridad. Esta nueva metodología aporta soluciones a problemáticas existentes en la representación y visualización de datos lo que permitirá que bases de datos puedan ser interpretadas fácilmente en tiempos de procesamiento prudentes. Un último aspecto importante son las estrategias interactivas de análisis visual, las cuales corresponden a formas gráficas de sintonizar los parámetros de métodos de

reducción de dimensión, de modo que usuarios no expertos puedan aplicarlas de manera intuitiva e interactiva.

Otra contribución importante tiene que ver con el aporte científico y la divulgación de resultados a través de la publicación de artículos científicos obtenidos con el desarrollo de este trabajo de grado. En total fueron tres los artículos publicados, uno como autor principal (STSIVA'16) y dos como coautor (CIARP'16 y LA-CCI'16). Las temáticas abordadas en los tres artículos están relacionadas con visualización de datos, además, de formas interactivas y eficientes de aplicar métodos de reducción de dimensión a una base de datos en particular.

## 1.4. ORGANIZACIÓN DEL DOCUMENTO

Este trabajo está dividido en 7 secciones principales nombradas de la siguiente manera: Introducción, descripción del problema, objetivos, marco teórico, metodología, resultados y conclusiones.

En la sección 1 se presenta el planteamiento del problema, la justificación de este trabajo y las contribuciones científicas de esta investigación.

En la sección 2 se presenta el objetivo general y los objetivos específicos, que fueron planteados al inicio de este trabajo de grado.

En la sección 3 se presenta una revisión bibliográfica que incluyen conceptos básicos de *Big Data*, alta dimensión en bases de datos y las diferentes etapas que conforman el proceso de descubrimiento de nuevo conocimiento en bases de datos como, por ejemplo: minería de datos, diferentes técnicas de visualización de datos, reducción de dimensión, aprendizaje de máquina y reconocimiento de patrones.

En la sección 4 se describe la metodología de visualización propuesta, así como la implementación del modelo cromático, la medida de calidad utilizada y las bases de datos empleadas. Del mismo modo, Los resultados de los experimentos se discuten en la sección 5.

Finalmente, en la sección 6 se presentan las conclusiones que se obtuvieron a partir de este trabajo, así como los trabajos futuros que pueden mejorar la metodología de visualización propuesta y definir nuevos trabajos de investigación.

# 2. OBJETIVOS

En esta sección se dan a conocer los objetivos planteados en el presente trabajo de grado.

## 2.1. OBJETIVO GENERAL

Desarrollar e implementar una metodología de visualización interactiva de datos de alta dimensión, a partir de un modelo intuitivo de mezcla de métodos de reducción de dimensión, que involucre un bajo costo computacional.

## 2.2. OBJETIVOS ESPECÍFICOS

- Implementar algoritmos de métodos de reducción de dimensión en el entorno de desarrollo integrado Processing, con el fin de reducir el tiempo de respuesta presentado en algunas plataformas de alto nivel.

- Proponer un modelo de visualización intuitivo para la percepción humana que permita la mezcla interactiva de los métodos de reducción de dimensión.

- Evaluar el desempeño de la metodología de visualización mediante la implementación de una interfaz garantizando criterios de bajo costo computacional, manejo intuitivo e interactividad.

# 3. MARCO TEÓRICO

## 3.1. BIG DATA Y DATOS DE ALTA DIMENSIÓN

El gran poder de procesamiento de las máquinas y su bajo costo de almacenamiento ha permitido un gran crecimiento en las capacidades de generar y agrupar datos en cantidades enormes, estos grupos de datos contienen una gran cantidad de información oculta la cual es sumamente importante [22], [26], (**Figura 1**). Dicha información puede ser estructurada, semiestructurada o no estructurada y puede aportar enorme valor a cualquier entidad. Sin embargo, trabajar con estos grandes volúmenes de información supone un consumo excesivo de recursos humanos e informáticos para su correcta manipulación e interpretación [14], [27].

El término *Big Data* hace referencia a todo aquello que tiene que ver con grandes volúmenes de información provenientes de diversas fuentes y escenarios (Economía, educación, salud, astronomía, etc.). Los conceptos fundamentales que son agrupados en el área del *Big Data* son: Volumen, visualización, variabilidad y velocidad; todos suponen un reto al momento de procesar y almacenar la información. La variedad de su origen y la rapidez con la que se incrementa su volumen, son algunos de los factores que dan lugar a esta área emergente de investigación. El avance de la tecnología ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos que tomaría demasiado tiempo y sería muy costoso cargarlos a una base de datos relacional para su análisis. Entonces se puede decir que el concepto de *Big Data* aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales [28]. Además del gran volumen de información, existe una gran variedad de datos que pueden ser representados de diversas maneras, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso [14], [28].

*Big Data* representa una clase especial de bases de datos la cual es muy grande para utilizar métodos estándar de procesamiento de datos. Los problemas en este tipo de datos están presentes en la recolección, el almacenamiento, búsqueda, visualización y análisis [28]. Una ventaja importante de *Big Data* es la información adicional que puede ser obtenida de grandes bases de datos en lugar de un análisis de bases de datos pequeñas y separadas entre sí. Una parte fundamental de *Big Data* es la visualización debido a la enorme importancia de exponer a un usuario en particular, una representación inteligible de los datos que posee, de tal forma que, pueda obtener a partir de patrones de interés nuevo conocimiento que le permita tomar decisiones de una mejor manera [28]. Muchas aplicaciones de visualización involucran procesamiento y análisis de la información contenida en bases de datos de alta dimensión, por ejemplo, clasificación de documentos, reconocimiento de patrones, detección de intrusos, etc. La metodología de estas aplicaciones depende en gran medida de la eficacia de procesamiento y extracción de patrones significativos de los conjuntos de datos y la precisión de la búsqueda [7], [22].


**Figura 1.** Diferentes fuentes de información en la actualidad que han hecho del *Big Data* un campo de estudio ampliamente utilizado en las entidades y organizaciones. **Fuente:** https://innovainternetmx.com/2015/01/uso-big-data/.

Tradicionalmente la estructura de un conjunto de datos se presenta como una matriz de **m** columnas y **n** filas, representando cada fila información sobre **n** variables medidas en cada unidad (individuo, empresa, inmueble, calle de una gran ciudad, procedimiento judicial, etc.) [14], [15]. De este modo los volúmenes de datos generados actualmente (*Big Data*) crecen en cantidad y dimensión, lo que hace, que un registro pueda fácilmente tener docenas de atributos y el dominio de cada atributo un rango bastante amplio, por este motivo, es importante obtener

representaciones fácilmente interpretables en bases de datos de alta dimensión [5], [6]. Sin embargo, el problema de los datos de alta dimensión es a menudo abordado por el usuario delimitando el análisis sólo a unas cuantas características o mediciones. En consecuencia, la definición de un sub-espacio que sólo contenga algunas de las mediciones efectuadas en cada registro puede ser propenso a errores [6], [27], por consiguiente, necesario encontrar un espacio de características reducido que posea la información más relevante del espacio original.

## 3.2. DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

Los grandes volúmenes de datos presentes en la actualidad, vienen acompañados de la necesidad de poderosas herramientas de análisis y representación, debido a que existen casos en donde se puede encontrar un denso repositorio de datos, pero sin los instrumentos adecuados la información que se obtiene de este puede no ser de mucha utilidad [7]. En consecuencia, existe un desaprovechamiento de valiosa información que puede estar incrustada en grandes volúmenes de datos, de este modo un usuario sin las herramientas apropiadas será propenso a formar conjeturas o tomar decisiones sin tener en cuenta la información que puede poseer a su disposición. Entonces surge la necesidad de encontrar diferentes técnicas, métodos y algoritmos que ayuden a los investigadores, analistas, gerentes entre otros, a la obtención de patrones útiles para los grandes volúmenes de datos. Estas técnicas y herramientas son el sujeto de un emergente campo de investigación conocido como **Descubrimiento de Conocimiento en Bases de Datos** (DCBD o KDD -por sus siglas en inglés-) y es básicamente un proceso automático en el que partiendo de una base de datos de entrada se aplican técnicas y metodologías de forma que se pueda aprovechar y explorar de una mejor manera dicha base de datos (**Figura 2**). Este proceso conlleva a extraer patrones en forma de reglas o funciones, con el fin de que el usuario realice el respectivo análisis y pueda encontrar nuevo conocimiento [14], [15].

El proceso DCBD puede ser dividido en las siguientes etapas: **limpieza de los datos,** en esta etapa se limpian los datos “sucios”, es decir, datos incompletos (donde hay atributos o valores de atributos perdidos), el ruido (valores incorrectos o inesperados) y datos inconsistentes (conteniendo valores y atributos con nombres diferentes). Los datos “sucios” en algunos casos deben ser eliminados ya que pueden contribuir a un análisis inexacto y resultados incorrectos. **Integración de los datos,** combina datos de múltiples procedencias incluyendo múltiples bases de datos, que podrían tener diferentes contenidos y formatos. **Selección de los datos,** consiste en buscar el objetivo y las herramientas del proceso de minería, identificando los datos que han de ser extraídos, buscando los atributos apropiados de entrada y la información de salida para representar una tarea. **Transformación de los datos** en esta etapa los datos son transformados y consolidados en formas apropiadas para la minería de datos, algunas veces la transformación y la consolidación de los datos son desarrollados antes del proceso de selección de

datos, la reducción de los datos puede también ser desarrollada en esta etapa para obtener una representación más acorde a una tarea en particular. **Minería de datos,** en este proceso se aplican métodos inteligentes para la extracción de patrones inmersos en un conjunto de datos. **Evaluación del patrón,** se identifican los verdaderos patrones de interés que representan conocimiento. Por último, se tiene una etapa denominada **representación del conocimiento** en donde técnicas de visualización y representación son usadas para presentar el conocimiento en forma de patrones al usuario [14], [23].

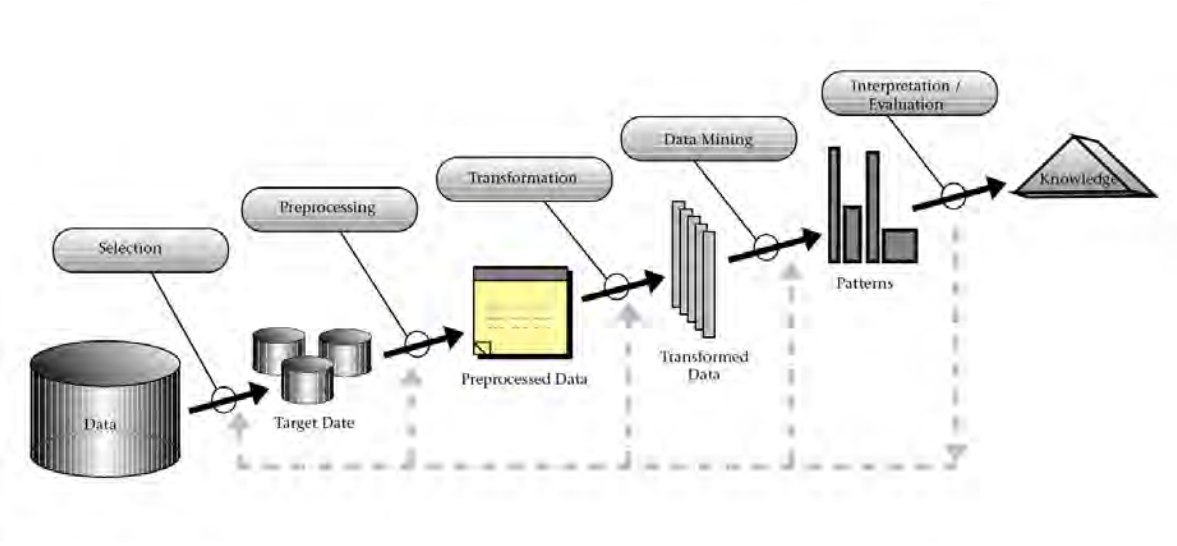



**Figura 2.** Esquema general de las etapas presentes en el proceso de descubrimiento de conocimiento en bases de datos (DCBD): integración de los datos, selección de los datos, transformación de los datos, minería de datos, evaluación del patrón, representación del conocimiento y por último interpretación y evaluación para encontrar nuevo conocimiento. **Fuente:** [23].

A pesar de que el proceso DCBD contiene técnicas y metodologías útiles al momento de analizar y aprovechar bases de datos, el crecimiento de los volúmenes de información presentes en la actualidad ha generado una alta demanda en el desarrollo de procesos que permitan entender estos volúmenes de información, esto se realiza eficazmente mediante la minería de datos, pero grandes volúmenes de datos pueden generar similares conjuntos de reglas o patrones. Estas formas de representación del conocimiento requieren de analistas con habilidades en la interpretación de patrones y reglas para extraer verdaderamente el conocimiento subyacente [24]. Lo anterior, es una de las razones por las que surgen las técnicas de reducción de dimensión las cuales permiten mitigar en cierta forma el problema de la alta dimensión, permitiendo reducir por ejemplo de 5000 variables a tan solo 5 o 4, pero, aun así, tales variables pueden ser abstractas, por lo que estas técnicas también necesitan de un experto para su interpretación [7], [19], [22]. En la actualidad, se han desarrollado herramientas de visualización y exploración

inteligente de datos que permiten comprender de una mejor forma los resultados obtenidos cuando se aplican técnicas de minería de datos, mientras se interactúa con múltiples presentaciones visuales de la información [22].

La información visual juega un papel muy importante en la minería de datos, ya que el objetivo de esta área es lograr descubrir conocimiento inmerso en datos, tal conocimiento no se puede determinar sino por métodos de minería, pero si este conocimiento no es fácilmente interpretable, se aumenta la inversión de tiempo, dinero y entendimiento (que supone la presencia de un experto en el tema). En la actualidad existen herramientas que, en general, implican etapas de pre procesamiento, uso de métodos de minería de datos, post procesamiento y/o la visualización. Sin embargo, no todas las herramientas integran todas las etapas mencionadas, terminando en resultados abstractos de la información. Asimismo, las herramientas que integran todas las etapas no tienen especial énfasis en la visualización, por lo que los resultados, a pesar de que involucran un análisis visual, tienden también a ser abstractos [7], [20], [22].

Este trabajo se enfoca principalmente en las etapas de: transformación de los datos originales y la visualización de manera que, se pueda reducir el tamaño de los datos (número de mediciones), encontrando las nuevas características más significativas en otro espacio. Este proceso es conocido como reducción de dimensión y busca mapear o proyectar datos en un espacio de características menor donde se conserven algunas propiedades del espacio original bajo ciertos criterios, y puedan ser representados en el diagrama de dispersión en el plano cartesiano (**Figura 3**).

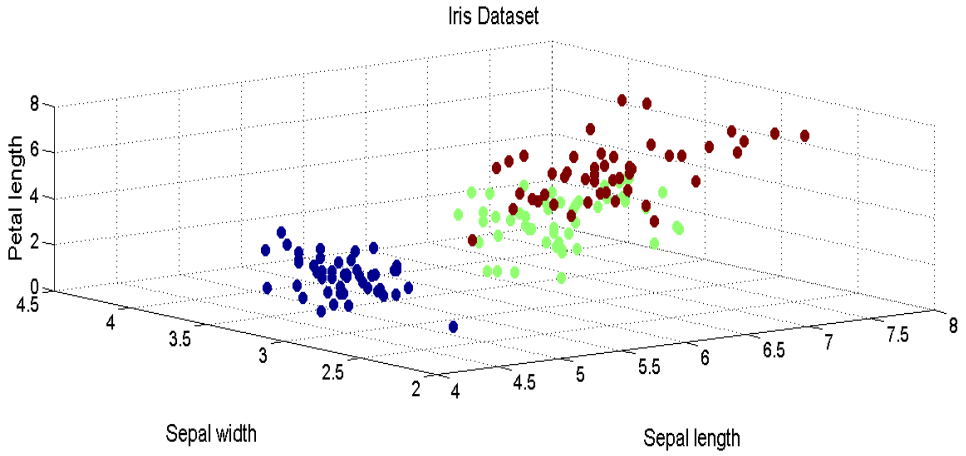



**Figura 3.** Diagrama de dispersión (*scatter plot*) de la base de datos iris la cual es quizás la base de datos más conocida que se encuentra en la literatura de reconocimiento de patrones. El conjunto de datos contiene 3 clases (color del punto) de 50 casos cada uno (número de puntos por clase), donde cada clase se refiere a un tipo de planta iris **Fuente:** Esta investigación.

## 3.3. REDUCCIÓN DE DIMENSIÓN.

La reducción de dimensión es considerada dentro del proceso DCBD como una etapa de pre procesamiento debido a que puede mapear o proyectar los datos a un espacio en donde los datos originales sean representados con menos atributos o características, con el fin, de mejorar tareas como la minería de datos y el reconocimiento de patrones. Además, al poder representar la base de datos original con dos o tres mediciones ayudaría en gran medida a la representación de los datos con el fin de que sean fácilmente interpretables por parte del usuario [4], [6], [27]. La aplicación de métodos de reducción de dimensión puede generar una representación de los datos de alta dimensión originales en un espacio de dimensión menor, el cual es formado por una combinación lineal o no lineal de unos atributos dados, un claro ejemplo de la aplicación de diferentes tipos de métodos de RD y sus efectos puede ser observado en la **Figura 4**.

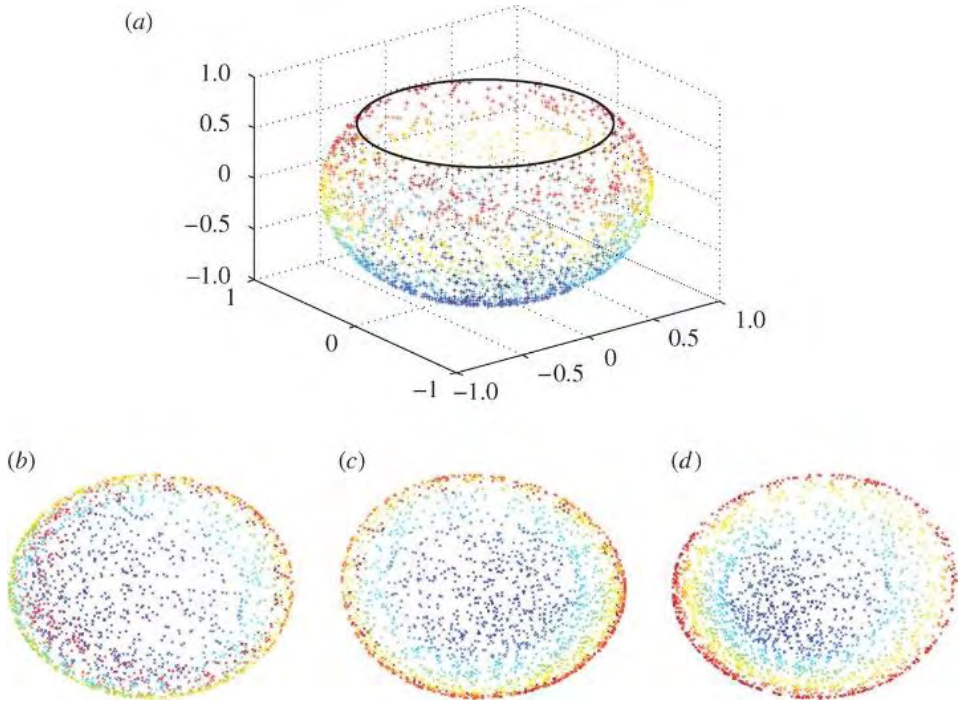



**Figura 4.** (a) Datos 'pecera' (esfera con la tapa superior retirada). (b) Datos embebidos bidimensionales obtenidos a través de PCA. (b y d) Las representan los espacios embebidos obtenidos por métodos de RD no lineales conocidos como mapas de difusión y *Laplacian Eigenmaps*, respectivamente. **Fuente:** [29].

El nuevo espacio de baja dimensión tiene la propiedad de conservar en cierta medida las distancias presentes en el espacio original de alta dimensión, tomando un subconjunto de variables de forma que el espacio de representación original de los datos sea reducido de manera óptima de acuerdo con ciertos criterios de calidad,

cuya finalidad será diferenciar el subconjunto que permita representar de la mejor manera el espacio inicial [29]. Los datos iniciales corresponden a muestras u objetos representados en características o variables. La inclusión de un gran número de variables dentro del proceso de exploración de los datos puede incrementar costos y tiempo de procesado e incluso puede generar datos con información ruidosa e irrelevante que pueden ser eliminadas con un método de RD [4]. Como se ha mencionado antes, una forma intuitiva de visualizar datos es mediante gráficos 2-D o 3D, lo que resulta en una visualización natural e inteligible para los seres humanos. En este sentido, la reducción de dimensión toma lugar, siendo una etapa importante en los sistemas de visualización de datos [14], [22]. Técnicamente, la reducción de dimensión (RD) tiene por objetivo alcanzar una representación de los datos dentro de un espacio de baja dimensión, sobre el cual, se puede mejorar el desempeño de las tareas de minería de datos, y a la vez hace que la representación de los datos, considerando su naturaleza intrínseca, sea más adecuada e inteligible para el ser humano [6]. Por esta razón las técnicas RD se convierten en herramientas importantes de análisis que permiten aprovechar todo el conjunto de mediciones realizadas por cada registro. En la **Figura 5** se puede encontrar una clasificación general de los métodos de RD.

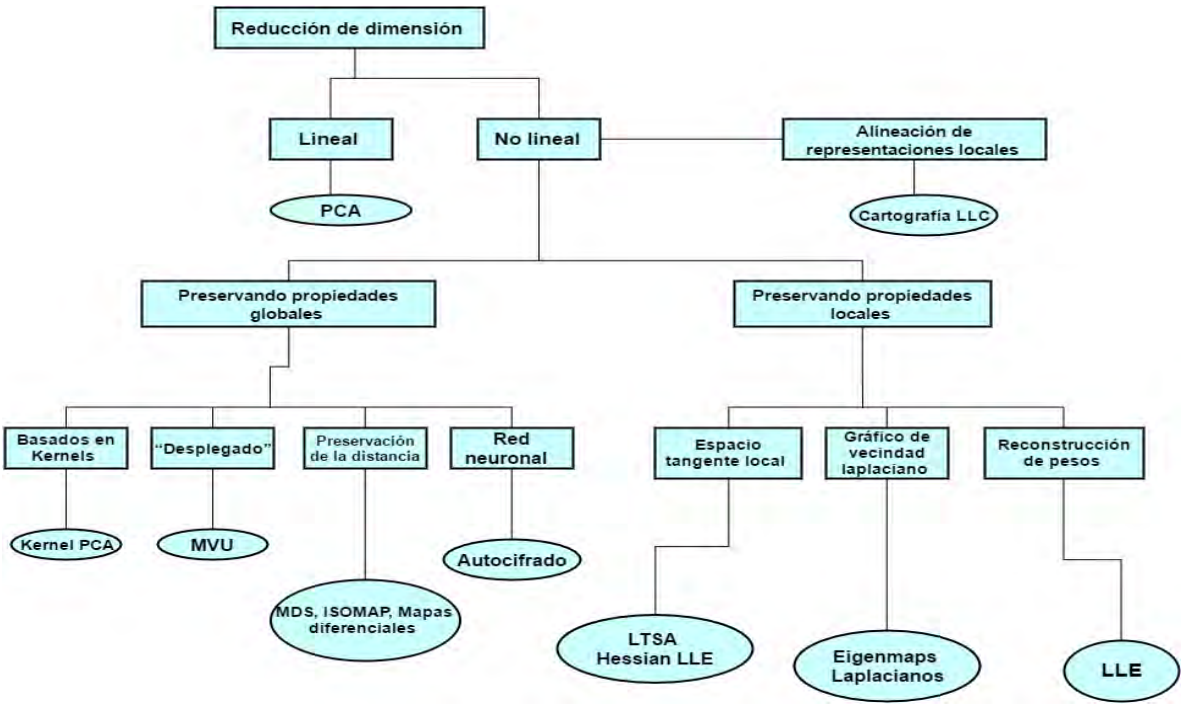



**Figura 5.** Clasificación de los diferentes métodos de reducción de dimensión existentes. En grandes rasgos como se indica en la figura los métodos de RD pueden clasificarse como métodos lineales y no lineales. **Fuente:** [27].

## 3.4. MINERÍA DE DATOS

El crecimiento exponencial en la generación de datos y bases de datos ha creado una urgente necesidad por nuevas técnicas y herramientas que puedan procesar de manera inteligente y automática grandes volúmenes de datos y transformarlos en información útil (conocimiento). De este modo la minería de datos se ha convertido en un área de investigación con una importancia cada vez mayor, debido a que formula un conjunto de técnicas y herramientas que permiten el correcto procesamiento de bases de datos [14], [15], [23].

La minería de datos es una parte esencial del proceso DCBD y se define como un proceso de descubrimiento de patrones, tendencias y significativas relaciones al examinar grandes volúmenes de datos para determinar información inmersa (también conocida como información oculta) que no pudo ser detectada con anterioridad permitiéndole al usuario realizar predicciones que resuelven problemas y permiten tomar decisiones. Más específicamente, las técnicas de minería de datos tienen como finalidad el descubrir patrones, perfiles y tendencias de interés a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos según el requerimiento de usuarios. En la actualidad, la madurez de la minería de datos ha logrado que estas técnicas y tecnologías incursionen directamente en entornos de base de datos actuales [14]. La minería de datos importa conceptos del *Machine learning*, la Inteligencia Artificial y la estadística multivariada para analizar los patrones en las bases de datos, donde estas últimas se representan en forma de matrices con información estructurada.

## 3.5. VISUALIZACIÓN DE DATOS

La visualización de los datos tiene como objetivo presentar los datos de manera clara y efectiva a través de representaciones gráficas con el fin de acercar las bases de datos a los usuarios, de este modo, la visualización de datos se convierte en una herramienta efectiva para explorar los grandes volúmenes de datos que pueden ser generados en diferentes entidades y organizaciones [22], [29]. Por esta razón la visualización de los datos ha sido en los últimos años ampliamente utilizada en muchas aplicaciones como, por ejemplo, reportes de trabajo, gestión de las operaciones comerciales y seguimiento de algún proceso. Pero aún más importante las técnicas de visualización pueden ser usadas para descubrir posibles relaciones que de otra forma no son fácilmente observables en los datos sin ninguna clase de pre-procesamiento [14]. En consecuencia, se puede decir que la visualización de datos está fuertemente ligada a la minería de datos dada su capacidad para representar los patrones encontrados en una base de datos. Según [10] hay dos

problemáticas que existen en los sistemas de visualización cuando se quiere representar grandes volúmenes de datos: primero los sistemas actuales tienen que gastar considerables cantidades de tiempo y recursos computacionales manipulando los datos, lo cual es bastante problemático cuando se intenta analizar datos en grandes cantidades. El segundo problema, radica en que los sistemas actuales de representación carecen de herramientas efectivas para representar estos datos de manera que se evite la superposición de los datos en un gráfico, es decir, en algunos casos existen demasiados registros para graficar, haciendo la visualización demasiado densa para ser útil para el usuario [3].

En esta sección se explicará brevemente algunos métodos de visualización de datos, con el fin, de abordar el problema de la representación efectiva de los datos.

### 3.5.1. Técnicas de visualización basadas en píxeles.

Una forma sencilla de visualizar una característica o atributo de una base de datos es por medio del uso de pixeles debido a que ofrece un análisis netamente visual basado en el nivel de intensidad de un pixel. Para una base de datos de dimensión **n**, existirán **n** ventanas en donde cara pixel representará un registro y el nivel de intensidad de cada pixel representará el valor de la característica. Dentro de las ventanas los datos son ordenados de alguna manera en específica según la tarea que se quiera realizar con el fin de observar patrones o tendencias, un ejemplo de este proceso puede ser visualizado en la **Figura 6** en donde se tiene una base de datos de una empresa de electrónica, con cuatro características (dimensión igual a cuatro). Se clasifican a todos los clientes en un orden predeterminado con el fin de exponer los datos de los clientes en las cuatro ventanas de visualización. Los colores de los píxeles se eligen de manera que cuanto más pequeño sea el valor, más claro es el sombreado.

Sin embargo, si se quiere representar en una ventana datos de una manera lineal puede no funcionar bien para una ventana con un número considerable de columnas. Debido a que, el primer pixel en una fila está lejos del último píxel en la fila anterior, no obstante, están uno al lado del otro en el orden global, por otra parte, un píxel es cercano al pixel de arriba en la ventana a pesar de que, los dos no son vecinos en el orden global. Para resolver este problema, se pueden ordenar los datos en una curva con espacios vacíos para rellenar las ventanas como se indica en la **Figura 7.** Es importante tener en cuenta que, las ventanas no tienen que ser necesariamente de una forma rectangular. Por ejemplo, la técnica de segmento circular utiliza ventanas en forma de segmentos de un círculo, como se ilustra en la **Figura 8**. Esta técnica puede facilitar la comparación de las dimensiones o mediciones porque las ventanas de dimensiones están una al lado de la otra, para formar una circunferencia [14].

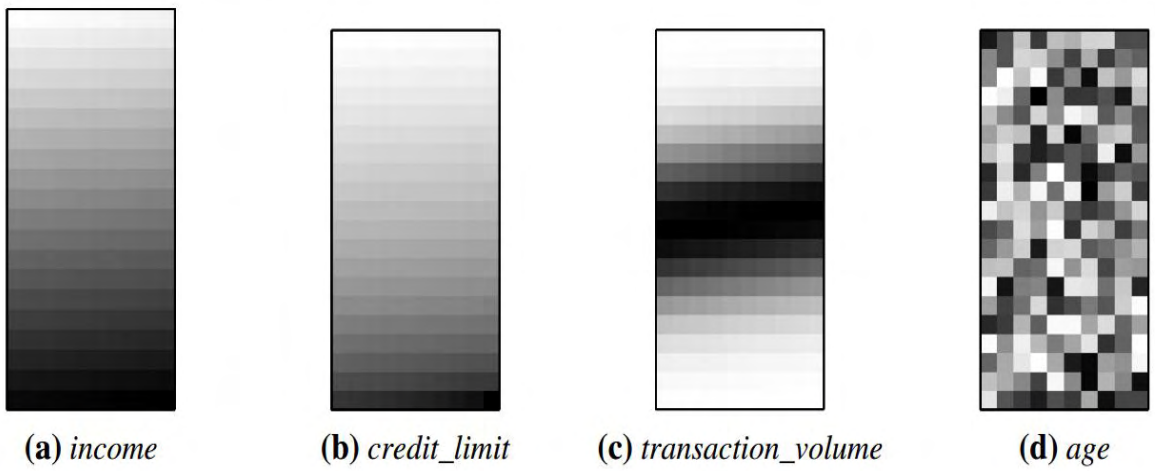



**Figura 6.** Visualización de datos por medio de pixeles. En la figura se hace una clasificación de los clientes en orden ascendente de ingresos (a). En la figura se puede ver tendencias como, por ejemplo, el límite de crédito (b) aumenta a medida que aumenta el ingreso; los clientes con ingresos en el rango medio son más propensos a comprar mucho más (c), además no existe una clara correlación entre el ingreso y la edad (d). **Fuente:** [14]**.**

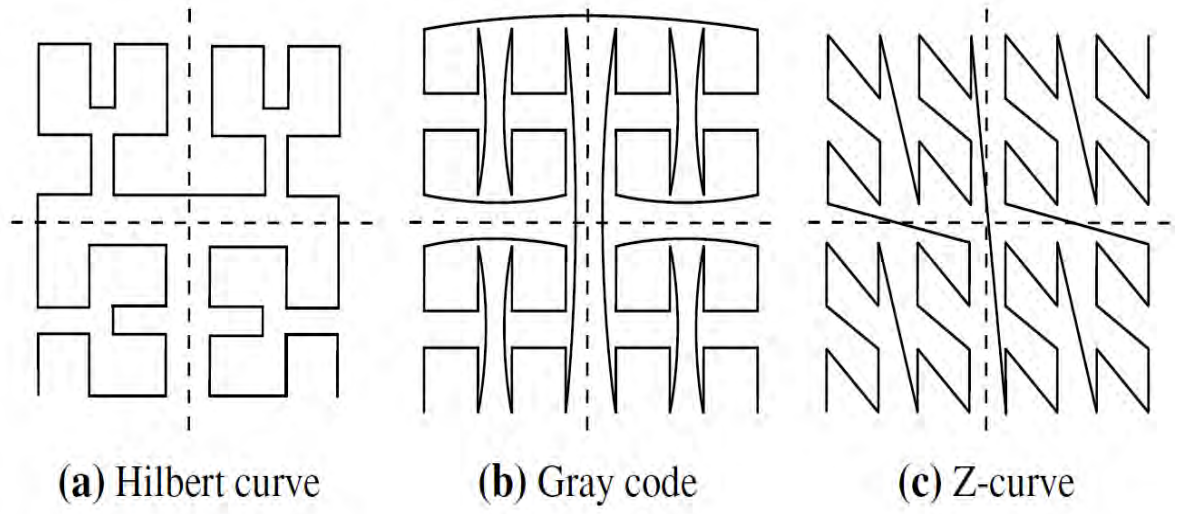



**Figura 7.** Algunos ejemplos de curvas 2-D alternativas a la superficie rectangular, las cuales son frecuentemente utilizadas para la representación de datos multidimensionales. **Fuente:** [14].

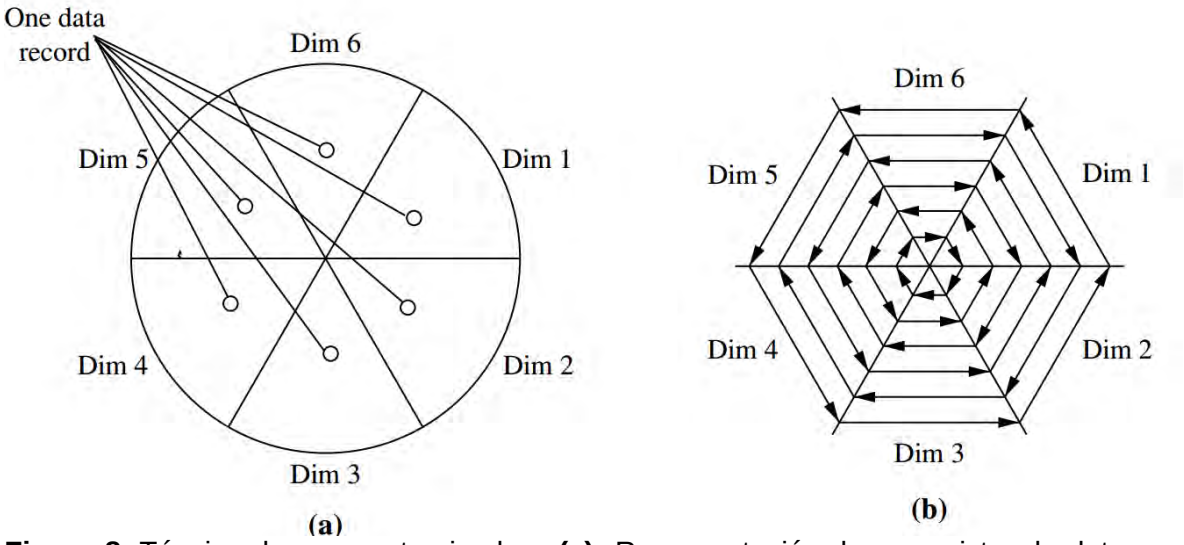



**Figura 8.** Técnica de segmento circular. **(a)** Representación de un registro de datos en segmentos de un círculo. **(b)** Los pixeles están en un arreglo fuera del círculo. **Fuente:** [14].

### 3.5.2. Técnicas de visualización de proyección geométrica

Las técnicas de visualización basadas en proyecciones geométricas ayudan al usuario a encontrar agrupamientos o patrones dentro de datos multidimensionales. Sin embargo, el reto central de estas técnicas es intentar visualizar un **espacio** de alta dimensión en un plano. Un diagrama de dispersión (*scatter plot*) en 2-D grafica puntos usando coordenadas cartesianas y el valor de la proyección en el eje $x$ y eje $y$ estará definido por el valor del atributo o característica a la cual representa [3]. Una tercera dimensión puede ser agregada, manteniendo el espacio bidimensional usando colores o formas para representar diferentes puntos. Un ejemplo de este tipo de gráficos puede ser observado en la **Figura 9.**

Un diagrama de dispersión en 3-D (**Figura 3**) usa tres ejes en el sistema de coordenadas cartesianas y al igual que en el plano bidimensional se puede agregar una cuarta dimensión utilizando formas y colores, sin embargo, para bases de datos con más de cuatro dimensiones, los diagramas de dispersión comienzan a perder utilidad y la información que presentan puede ser ambigua incluso para el usuario experto. Existe otra técnica popular de visualización llamada **coordenadas paralelas (Figura 10),** la cual puede manejar datos de alta dimensión para visualizar puntos n-dimensionales. La técnica de coordenadas paralelas dibuja **n** ejes igualmente espaciados, uno por cada dimensión, paralelos a uno de los ejes de visualización. Un dato es representado por una recta poligonal que intersecta cada eje en el punto correspondiente al valor de cada dimensión (atributo o característica).

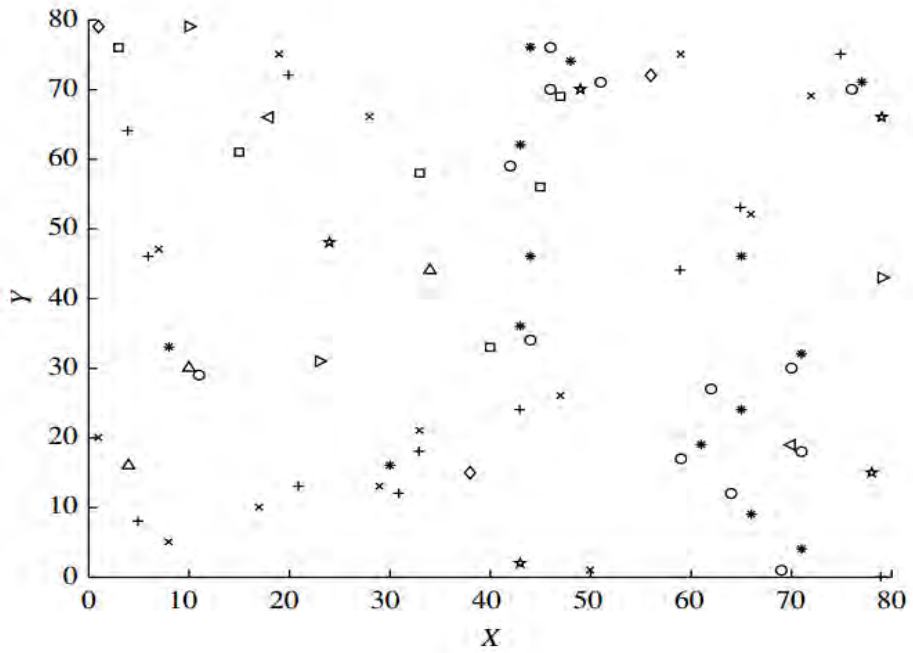



**Figura 9.** Se muestra un ejemplo donde los valores del eje x y el eje y son de dos atributos y la tercera dimensión es representada por diferentes formas, de modo que se tiene una visión tridimensional de las bases de datos en un plano bidimensional. **Fuente:** [14].

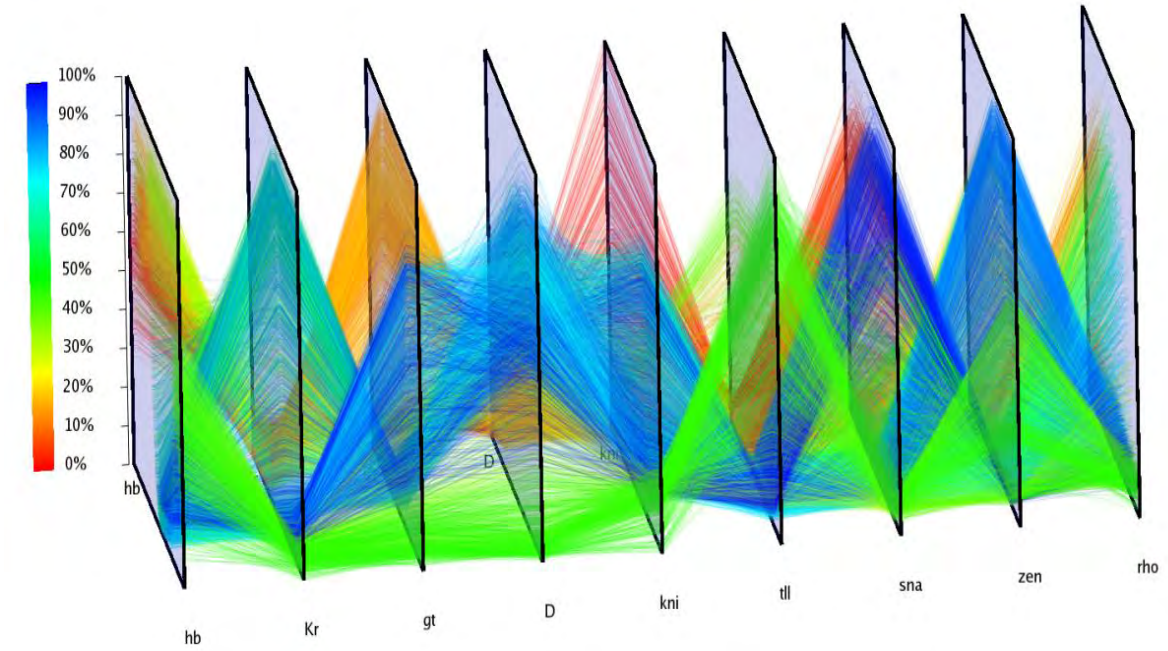



**Figura 10.** Un ejemplo de representación 3D de coordenadas paralelas de una base de datos de un conjunto de células y nueve genes (9 características). **Fuente:** http://www-vis.lbl.gov/Events/SC07/Drosophila/3DParallelCoordinates.png.

### 3.5.3. Técnicas de visualización basadas en iconos

Estas técnicas de visualización hacen uso de pequeños íconos para representar datos multidimensionales. Dos ejemplos populares de estas técnicas son, las caras de Chernoff y las figuras de bastón. Las caras de Chernoff (**Figura 11**) presentan aun usuario datos multidimensionales de hasta 18 variables (dimensiones) a través de caras en caricatura. Las caras de Chernoff ayudan a revelar tendencias en los datos con ayuda de las partes que conforman el rostro como por ejemplo los ojos, los oídos, la boca y la nariz, en consecuencia, la forma y dimensión de estas partes dependerán directamente de los valores de las características. La visualización de grandes tablas de datos puede ser tedioso y poco práctico, de manera que las caras de Chernoff hacen uso de la capacidad de la mente humana para reconocer las pequeñas características faciales y asimilarlas en conjunto.

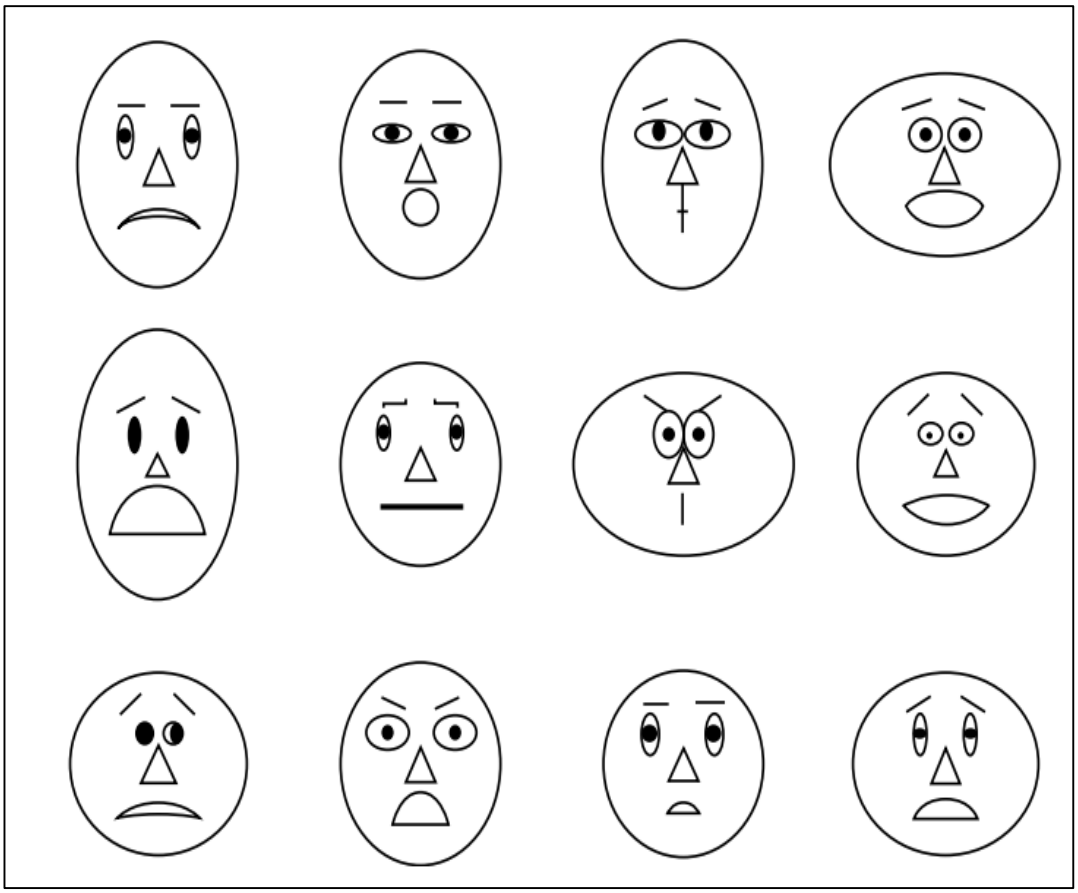


**Figura 11**. Caras de Chernoff. Cada cara representa un registro n-dimensional ($n \leq 18$) de manera que la forma y simetría de las caras será equivalente de alguna manera al valor de la característica o atributo de cada objeto. **Fuente:** [14].

La técnica de visualización basada en figuras de bastón (**Figura 12**) mapea datos multidimensionales a 5 figuras diferentes, en donde cada figura tiene cuatro extremidades y un cuerpo. Un plano bidimensional es formado con dos ejes ($\boldsymbol{x}$ y $\boldsymbol{y}$) y las dimensiones restantes son mapeadas en el ángulo o el largo de las extremidades. Si los elementos de datos son relativamente densos con respecto a las dos dimensiones de la pantalla, la visualización resultante muestra patrones de textura, lo que refleja tendencias en los datos.

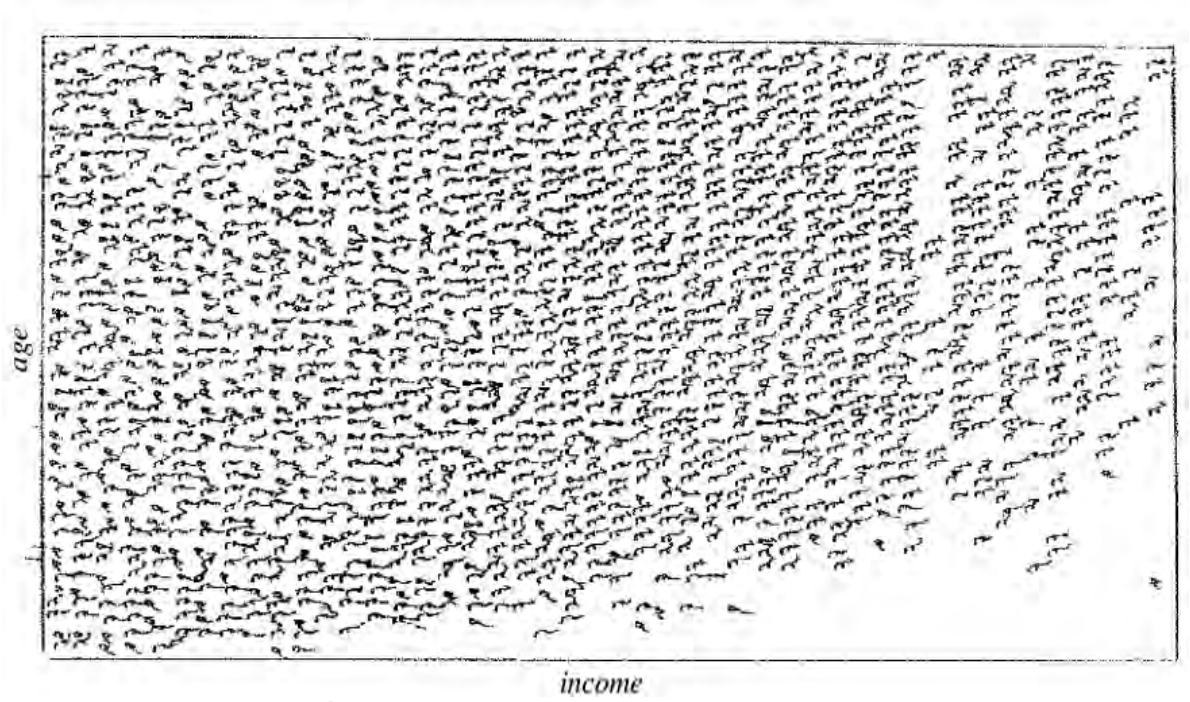



**Figura 12.** Representación de datos generados por un censo mediante el uso de figuras de bastón. En esta figura se representa datos del censo donde la edad y el ingreso son mapeados en el plano y las características restantes (el género, la educación, e hijos) son representados mediante figuras de bastón. **Fuente:** Professor G. Grinstein, Department of Computer Science, University of Massachusetts at Lowell.

### 3.5.4. Técnicas de visualización jerárquicas

Las técnicas de visualización generalmente presentan problemas cuando se intenta representar una base de alta dimensión teniendo en cuenta todos sus atributos. Las técnicas de visualización jerárquica intentan solucionar este problema dividiendo las características o dimensiones en sub-conjuntos (sub-espacios) de tal forma que los sub-espacios puedan ser visualizados de manera jerárquica. Una de los métodos más representativos de visualización jerárquico es el conocido como **"Worlds-within-Worlds,"** o **n-Vision** el cual puede ser visualizado en la **Figura 13.**

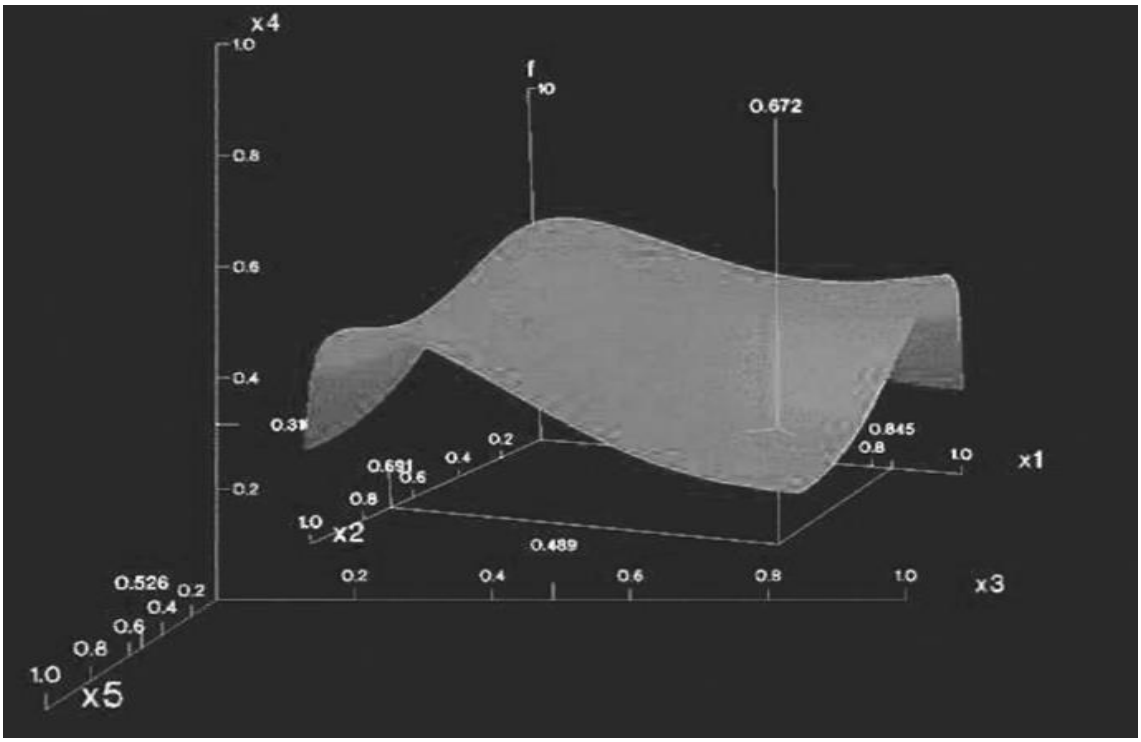



**Figura 13.** En la imagen se puede observar un ejemplo de la técnica de visualización conocida como “Worlds-within-Worlds,” o n-Vision con un plano tridimensional el cual subdivide las n dimensiones del espacio de datos y las organiza en un sub-espacio de manera jerárquica **Fuente:** [14].

### 3.5.5. Reducción de dimensión como técnica de visualización

Como se ha dicho anteriormente la visualización es la primera etapa del análisis de los datos, donde, el objetivo principal es darle sentido a la información que se tiene antes de proceder con otras fases como, por ejemplo, modelamiento, clasificación y análisis [2], [4]. Dada la existencia de una base de datos con una gran cantidad de variables medidas una idea obvia es reducir los atributos o características en variables más condensadas que representen la base de datos original con la menor pérdida de información posible [6]. Las técnicas de reducción de dimensión permiten realizar este procedimiento convirtiéndose en una herramienta clave de pre-procesamiento de una base de datos multidimensional.

La visualización de datos por medio del diagrama de dispersión (**Figura 3**) es una de las formas más naturales y básicas de representación de datos, sin embargo, esta técnica de visualización está limitada en gran medida por el número de características (dimensiones) que una base de datos pueda tener. Si bien, las técnicas nombradas anteriormente ayudan a mitigar de alguna manera los problemas producidos por la alta dimensión, los resultados gráficos que se obtienen pueden no ser lo suficientemente explícitos para un usuario inexperto. En

consecuencia, se podrían presentar ambigüedades al momento de la interpretación de una base de datos [3], [16]. De este modo, la reducción de dimensión como técnica de visualización se convierte en una gran herramienta que puede sacar provecho de todas las mediciones (características) que son realizadas a los registros, eliminando los datos redundantes y poco trascendentes con el fin de que puedan ser representados en un plano Cartesiano bidimensional o tridimensional que representa de manera concreta la naturaleza de los datos y de esta manera posibles tendencias o patrones (**Figura 14**) [18].

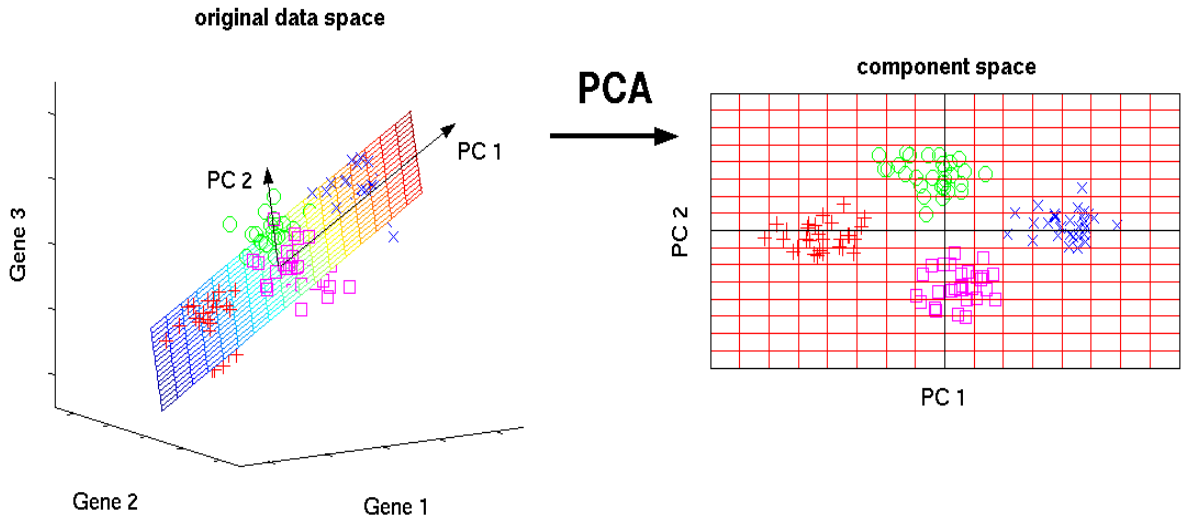



**Figura 14.** Efectos del método de reducción de dimensión conocido como, análisis de componentes principales (PCA). En este ejemplo se puede apreciar como un espacio de tres dimensiones es reducido a uno de dos dimensiones, en donde la base es más fácil de analizar **Fuente:** [16].

## 3.6. APRENDIZAJE DE MÁQUINA

Es un campo de estudio interdisciplinar, relacionado con el desarrollo de programas de computadora que mejoran el desempeño de ciertas tareas mediante la experiencia. De esta forma, los algoritmos para el aprendizaje de máquina (*machine learning*) han probado su importancia al tratar de crear programas capaces de generalizar comportamientos a partir de una información estructurada o no estructurada suministrada en forma de ejemplos. El aprendizaje de máquina estudia técnicas automáticas para el aprendizaje con el fin de encontrar reglas de clasificación o predicción de manera precisa con el fin, de imitar el razonamiento humano con base en la experiencia lo más parecido posible [24].

### 3.6.1. Reconocimiento de patrones

El reconocimiento de patrones es la ciencia que se encarga de la descripción y clasificación (reconocimiento) de objetos, personas, señales, representaciones, etc. Esté campo del aprendizaje de máquina (*machine learning*) trabaja con base en un conjunto previamente establecido de todos los posibles objetos (patrones) individuales a reconocer. Los márgenes de aplicaciones del reconocimiento de patrones son de naturaleza interdisciplinar y pueden proponer soluciones a problemas en diversas áreas del conocimiento [23], [24]. A grandes rasgos se puede decir que el reconocimiento de patrones busca imitar las capacidades cognoscitivas humanas para diferenciar algún objeto o fenómeno en particular de otro, teniendo en cuenta información proveniente del mundo exterior (señales). Un esquema de un sistema de reconocimiento de patrones consta de varias etapas relacionadas entre sí. La **Figura 15** muestra un esquema general de un sistema de reconocimiento de patrones, el cual está compuesto en primera instancia por un sensor o un transductor que es la fuente de información principal, este elemento comunica el sistema de reconocimiento de patrones con el objeto o fenómeno que se desea procesar. Una vez se tenga el patrón se realiza una etapa de extracción de características en donde a partir del patrón de representación, se extrae la información discriminatoria eliminando la información redundante e irrelevante, en otras palabras, es una etapa de pre-procesamiento de la información que se obtienen a partir del sensor. El clasificador es la etapa en donde se toman las decisiones del sistema. Su rol es asignar los patrones de clase desconocida a la categoría apropiada previamente definida [24].


**Figura 15.** Esquema general de un sistema de reconocimiento de patrones, en donde a una señal de entrada se asignará una clase previamente definida. **Fuente:** Esta investigación.

El objetivo de estas etapas es ajustar el sistema para que sea capaz de clasificar señales u objetos de entrada en una de las clases predefinidas. Para ello se deberá analizar un cierto número de características y para poder clasificar satisfactoriamente señales de entrada, es necesario un proceso de aprendizaje en el cual el sistema crea un modelo de cada una de las clases a partir de una

secuencia de entrenamiento o conjunto de vectores de características de cada una de las clases, como por ejemplo en la **Figura 16** se puede observar la representación de una base de datos en el plano cartesiano, las cuales son lineal mente separables una de las otras mediante un modelo basado en un recta $\boldsymbol{y = mx + b}$ en donde de acuerdo al vector de características de entrada se le asignara una clase [14]**,** [23], [24]. Generalmente se acepta que la secuencia de muestras de entrenamiento debe contener para cada una de las clases un mínimo de elementos igual a diez veces la dimensión de los vectores de características. El sistema de reconocimiento de patrones debe tener en cuenta las fuentes de variabilidad como son el ruido, rotaciones, cambio de escala y deformaciones que se logra incluyendo en la secuencia de entrenamiento patrones que hayan experimentado estas modificaciones [24].

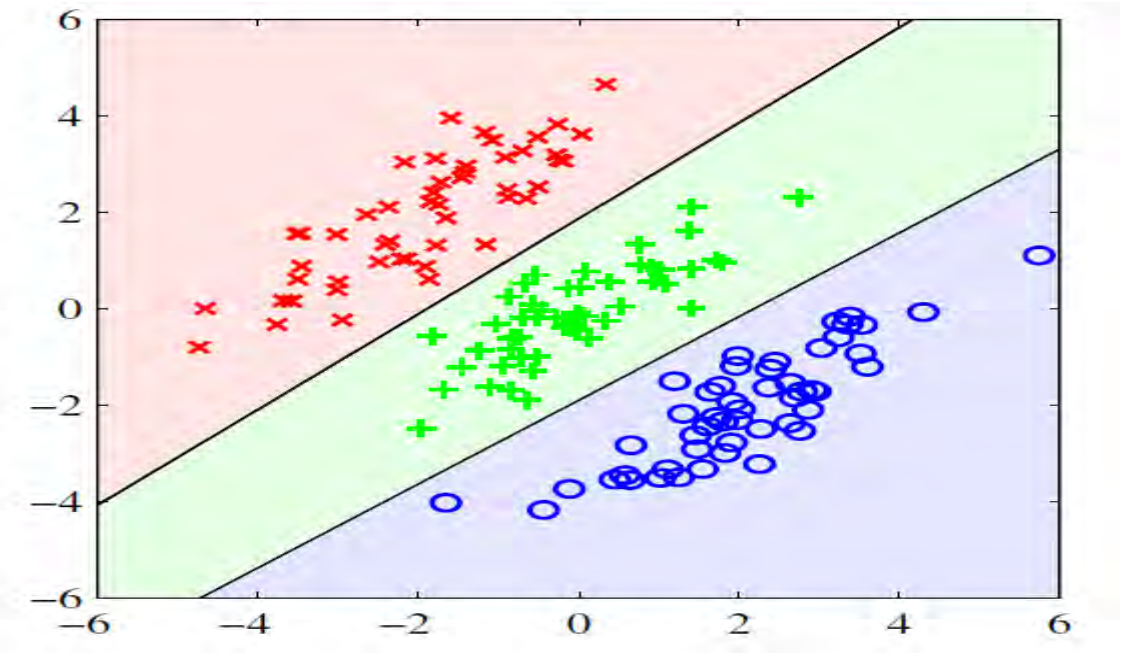



**Figura 16.** En esta imagen se puede apreciar la representación de una base de datos en el diagrama de dispersión. La base de datos está conformada por tres clases diferentes, cada clase está representada por un color (Rojo, verde y azul), una de las principales tareas del reconocimiento de patrones es asignar a un dato cualquiera una clase en particular. En este caso se puede observar que las clases son linealmente separables debido a que existe una recta que separa correctamente las clases. **Fuente:** [24].

# 4. METODOLOGÍA

En términos matemáticos, el objetivo de la reducción de dimensión es mapear o proyectar (si es una transformación lineal) datos de un espacio de alta dimensión $\boldsymbol{Y} \in \mathbb{R}^{DxN}$ a un espacio de baja dimensión $\boldsymbol{X} \in \mathbb{R}^{dxN}$, donde $d < D$, de modo que, los datos originales y los datos reducidos estarán conformados por $N$ puntos o registros, denotado respectivamente por $y_i \in \mathbb{R}^D$ y $x_i \in \mathbb{R}^d$, con $i \in \{1, \ldots, N\}$ [6], [8]. Esto significa que una matriz de datos de alta dimensión al ser mapeada o proyectada a un espacio embebido podría ser representada con menos atributos o características sin afectar el número de muestras presentes en la matriz de datos original. Para este trabajo se tienen en cuenta únicamente las dos primeras características del espacio embebido resultante, las cuales presentan la mayor parte de la información del espacio original, facilitando la tarea de representar los datos en un plano cartesiano bidimensional [27].

Como se explica más adelante este trabajo utiliza representaciones kernel de métodos de RD y un modelo cromático para su combinación. A continuación, en la **Figura 17** se indica un esquema general de la metodología de visualización propuesta.

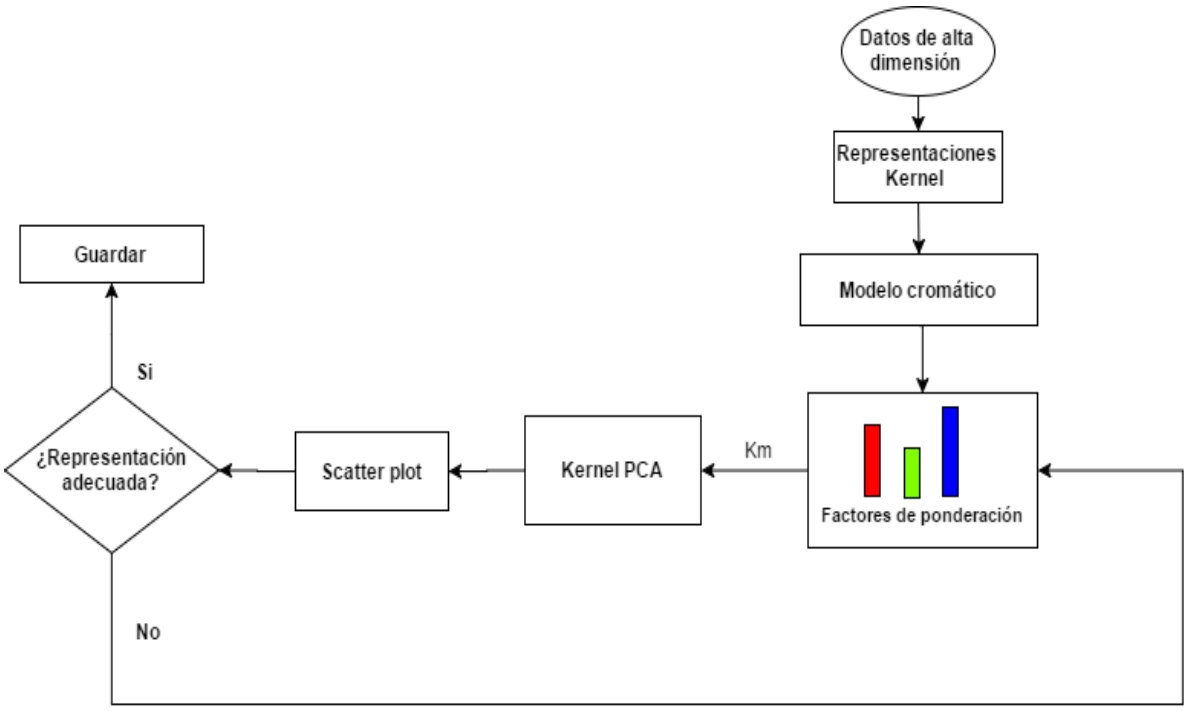



**Figura 17.** Esquema general de la metodología de visualización propuesta a través de aproximaciones kernel de métodos de RD y un modelo cromático. **Fuente:** Esta investigación.

## 4.1. KERNEL PCA

Como se ha dicho anteriormente, los métodos de RD tienen como objetivo encontrar a partir de una matriz $\mathbf{Y} \in \mathbb{R}^{D\times N}$ , un espacio embebido $\mathbf{X} \in \mathbb{R}^{d\times N}$ con $d < D$ que preserve la estructura o propiedades de $\mathbf{Y}$ tanto como sea posible bajo un criterio establecido [4], [6]. El método de RD conocido como análisis de componentes principales (PCA), es una proyección lineal que intenta preservar la varianza a partir de los valores y vectores propios de la matriz de covarianza [8], [27]. Además, cuando una matriz de datos es centrada, es decir que el valor medio de las filas (características) es igual a cero, la preservación de la varianza puede ser vista como una preservación del producto interno euclidiano [8].

Kernel PCA, al igual que PCA maximiza el criterio de varianza, pero en este caso de una matriz kernel, la cual es básicamente un producto interno de un espacio desconocido de alta dimensión. Sea $\mathbf{\Phi} \in \mathbb{R}^{D_h\times N}$ un espacio de alta dimensión con $D_h \gg D$, el cual es completamente desconocido excepto por su producto interno que puede ser estimado [8]. Para hacer uso de las propiedades de este nuevo espacio de alta dimensión y de su producto interno es necesario definir una función $\phi(\cdot)$ que pueda mapear los datos del espacio original a uno de más alta dimensión ($\mathbf{\Phi}$) de la siguiente manera:

$$\phi(\cdot): \quad \mathbb{R}^D \longrightarrow \mathbb{R}^{D_h},$$
$$y_i \longrightarrow \phi(y_i),$$

donde el $i$-esimo vector columna de la matriz $\mathbf{\Phi}$ estará dado por $\Phi i = \phi(y_i)$.

Teniendo en cuenta las condiciones de Mercer [30] y que la matriz $\mathbf{\Phi}$ este centrada, el producto interno de la función kernel $\mathrm{k}(\cdot,\cdot)$ puede ser calculado de la siguiente manera, $\phi(y_i)^T\phi(y_j) = k(y_i, y_j) = k_{ij}$. Además, al organizar todos los productos internos posibles en un arreglo $\mathbf{K} = [k_{ij}]$ se tendrá como resultado una matriz kernel:

$$\mathbf{K}_{N\times N} = \mathbf{\Phi}^T_{N\times D_h}\,\mathbf{\Phi}_{D_h\times N}, \tag{1}$$

donde $k_{ij} = k(y_i, y_j)$.

Para proyectar los datos de alta dimensión se utiliza una combinación lineal a través de una matriz $\mathbf{W} \in \mathbb{R}^{D_h\times d}$ (matriz de rotación), que es una base ortonormal, definida de la siguiente manera $\mathbf{W} = [\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(d)}]$ y $\mathbf{W^T W} = \mathbf{I}_d$, donde $\mathbf{w}^{(\ell)} \in \mathbb{R}^{D_h}$ y $\mathbf{I}_d$ es una matriz identidad de dimensión $d$. Entonces, la matriz proyectada a otro espacio de características $\mathbf{X} \in \mathbb{R}^{d\times N}$ puede ser obtenida con la expresión:

$$\mathbf{X} = \mathbf{W^T \Phi}. \tag{2}$$

Generalmente, la proyección es desarrollada sobre un espacio con una dimensión más baja que el espacio original, esto significa que los datos son proyectados con una representación de bajo rango de la matriz de rotación $(d < D)$. Sin embargo, los datos pueden ser proyectados en su totalidad al establecer $d = D$. Además, teniendo en cuenta la ecuación (2), una matriz de datos de bajo rango $\widehat{\mathbf{\Phi}} \in \mathbb{R}^{D_h \times d}$ puede ser obtenida cuando $d < D$ por:

$$\widehat{\mathbf{\Phi}} = \mathbf{WX}. \tag{3}$$

Reemplazando la ecuación (2) en la ecuación (3) se puede escribir $\widehat{\mathbf{\Phi}} = \mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{\Phi}$. Por lo tanto, el criterio de varianza puede ser expresado de la siguiente manera $E_{\mathbf{\Phi}}\{\|\mathbf{\Phi_i} - \mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{\Phi}\|_2^2\}$ donde $\|\cdot\|_2$ y $E_{\mathbf{\Phi}}\{\cdot\}$, denotan la norma euclidiana y el operador del valor esperado con respecto a $\mathbf{\Phi}$, respectivamente. Considerando a $E_{\mathbf{\Phi}}$ como un simple promedio, la función objetivo del error medio cuadrado puede ser escrita como:

$$\frac{1}{N}\sum_{i=1}^{n}\left\|\mathbf{\Phi_i} - \mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{\Phi}\right\|_2^2 = \left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_F^2, \tag{4}$$

donde $\|\cdot\|_F$ representa la norma de Frobenius. La representación optima de un espacio de alta dimensión en un espacio de baja dimensión, puede formularse como el siguiente problema de optimización:

$$\begin{gathered} \min_{\mathbf{W}} \left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_F^2 \\ \text{s.t. } \mathbf{W}^{\mathbf{T}}\mathbf{W} = \mathbf{I}_d, d < D \\ \mathbf{X} = \mathbf{W}^{\mathbf{T}}\mathbf{\Phi}, \end{gathered} \tag{5}$$

puede ser extendido como se muestra a continuación:

$$\left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_F^2 = \mathbf{tr}(\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi}) - \mathbf{2tr}\left(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\mathbf{\Phi}\right) + \mathbf{tr}(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\widehat{\mathbf{\Phi}}). \tag{6}$$

Teniendo en cuenta que el término $\mathbf{tr}(\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi}) = \|\mathbf{\Phi}\|_F^2$ es constante y que además $\mathbf{tr}\left(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\mathbf{\Phi}\right) = \mathbf{tr}(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\widehat{\mathbf{\Phi}})$, es válido asumir la siguiente dualidad:

$$\|\mathbf{\Phi}\|_F^2 = \mathbf{tr}\left(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\mathbf{\Phi}\right) + \left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_F^2, \tag{7}$$

donde el problema de minimizar $\left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_F^2$ puede ser expresado como un problema dual de maximización de $\mathbf{tr}\left(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\mathbf{\Phi}\right)$. Además, al sustituir en $\mathbf{tr}\left(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\mathbf{\Phi}\right)$ la ecuación (3) se tiene que:

$$\mathbf{tr}\left(\widehat{\mathbf{\Phi}}^{\mathbf{T}}\mathbf{\Phi}\right) = \mathbf{tr}(\mathbf{\Phi}^{\mathbf{T}}\mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{\Phi}) = \mathbf{tr}(\mathbf{W}^{\mathbf{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathbf{T}}\mathbf{W}). \tag{8}$$

En consecuencia, el nuevo problema de optimización será:

$$\max_{\mathbf{W}} \mathbf{tr}(\mathbf{W^T \Phi \Phi^T W}) \quad (9)$$

$$\text{s.t. } \mathbf{W^T W} = \mathbf{I}_d.$$

Para resolver el anterior problema, se puede escribir el correspondiente lagrangiano de la siguiente forma:

$$\mathcal{L}(W|\mathbf{\Phi}) = \mathbf{tr}(\mathbf{W^T \Phi \Phi^T W}) - \mathbf{tr}\Big(\mathbf{\Lambda}(\mathbf{W^T W} - \mathbf{I}_d)\Big), \quad (10)$$

donde $\mathbf{\Lambda} = \mathbf{Diag}(\boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda}_{D_h})$ son los multiplicadores de Lagrange. Resolviendo la condición de primer orden del lagrangiano, se obtiene el siguiente problema dual.

$$\mathbf{\Phi \Phi^T W} = \mathbf{W \Lambda} \Rightarrow \mathbf{W^T \Phi \Phi^T W} = \mathbf{\Lambda}. \quad (11)$$

Por lo tanto, el problema de optimización puede ser solucionado definiendo las matrices $\mathbf{\Lambda}$ y $\mathbf{W}$ como los valores y los vectores propios de $\mathbf{\Phi \Phi^T}$, respectivamente. Además, como el problema es de maximización, los vectores propios asociados a los $d$ mayores valores propios deben ser seleccionados. De manera similar, pre-multiplicando la ecuación (11) por $\mathbf{\Phi^T}$, se obtiene:

$$\mathbf{\Phi^T \Phi \Phi^T W} = \mathbf{\Phi^T W \Lambda} \Rightarrow \mathbf{K X^T} = \mathbf{X^T \Lambda}, \quad (12)$$

De esta manera, un espacio embebido de $\mathbf{X}$ puede ser calculado como los vectores propios de la matriz $\mathbf{K}$.

## 4.2. MÉTODOS DE REDUCCIÓN DE DIMENSIÓN CON APROXIMACIONES KERNEL

La ventaja de trabajar con el espacio de alta dimensión $\mathbf{\Phi}$ es que puede mejorar en gran medida la representación y la visualización de los datos embebidos del espacio original mapeado al espacio de alta dimensión, a partir del cálculo de los valores y vectores propios de su producto interno. Una estimación del producto interno o interior (kernel) puede ser diseñado a partir de la función y aplicación que uno quiera desarrollar [31], en este caso las matrices kernel representarán funciones de distancia asociadas a un método de reducción de dimensión en particular. Para el desarrollo de la metodología de visualización propuesta se tienen en cuenta tres métodos espectrales de reducción de dimensión llamados: *Classical Multidimensional Scalling* (CMDS), *Locally linear Embedding* (LLE), y *Laplacian Eigenmaps* (LE), los cuales son ampliamente explicados en [26], [31].

La representación kernel para el método de reducción CMDS se define como la matriz de distancia $\mathbf{D} \in \mathbb{R}^{N \times N}$ doblemente centrada, es decir haciendo que la media de las filas y las columnas sea cero, así :

$$\boldsymbol{K_{CMDS}} = -\frac{\mathbf{1}}{\mathbf{2}}\big(\boldsymbol{I} - \mathbf{1_N 1_N^T}\big)\boldsymbol{D}\big(\boldsymbol{I} - \mathbf{1_N 1_N^T}\big), \quad (13)$$

donde la $ij$-esima entrada de $\boldsymbol{D}$ es dada por la distancia euclidiana $d_{ij} = \|y_i - y_j\|_2^2$.

El kernel para el método LLE puede ser aproximado a partir de la forma cuadrática en términos de la matriz $\boldsymbol{\mathcal{W}}$ con coeficientes lineales que suman 1 y pueden de manera óptima reconstruir los datos originales. Sea una matriz $\mathbf{M} \in \mathbb{R}^{N\times N}$ definida por la expresión $\boldsymbol{M} = (\boldsymbol{I_N}\text{-}\boldsymbol{\mathcal{W}})\ (\boldsymbol{I_N}\text{-}\boldsymbol{\mathcal{W}^T})$ y $\boldsymbol{\lambda_{max}}$ como el valor propio más grande de $\mathbf{M}$:

$$\boldsymbol{K_{LLE}} = \boldsymbol{\lambda_{max}}\boldsymbol{I_N} - \boldsymbol{M} \tag{14}$$

Debido a que kernel PCA es un problema de maximización de la covarianza de alta dimensión representada por un kernel, LE puede ser representado como la matriz pseudo-inversa del grafo $L$, como se muestra en la siguiente expresión.

$$\boldsymbol{K_{LE}} = \boldsymbol{L}^{\dagger}, \tag{15}$$

Donde $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, tal que $\boldsymbol{S}$ es una matriz de disimilitud y $\boldsymbol{D}$=**Diag** $(\mathbf{S}\mathbf{1}_N)$ la matriz del grado de $\boldsymbol{S}$. La matriz de similitud $\boldsymbol{S}$ está formada de tal manera que el parámetro del ancho relativo se estima manteniendo la entropía de la distribución con el vecino más cercano con aproximadamente $\log(\mathrm{K})$, donde k es el número dado de vecinos como se explica en [32]. El número de vecinos se establece como el entero más cercano al 15% de la cantidad de datos.

Kernel PCA es desarrollado bajo la condición de que la matriz $\boldsymbol{\Phi}$ tenga media cero, por lo tanto, se debe asegurar esta condición centrando la matriz kernel de la siguiente manera [8], [32].

$$\begin{aligned}\mathbf{K} \leftarrow \mathbf{K} - \tfrac{1}{N}\boldsymbol{K}\mathbf{1}_N\mathbf{1}_N^T - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\boldsymbol{K} + \tfrac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^T\boldsymbol{K}\mathbf{1}_N\mathbf{1}_N^T \\ = (\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N^T)\boldsymbol{K}(\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N^T),\end{aligned} \tag{16}$$

donde $\mathbf{1}_N$ es un vector N-dimensional con todos sus elementos iguales a 1 y $\boldsymbol{I}_N$ es la matriz identidad. El propósito de obtener esta representación kernel de métodos de reducción de dimensión es conseguir un kernel generalizado $\widehat{K} \in \mathbb{R}^{NxN}$ de una combinación lineal de kernel como se indica en la ecuación número (16).

### 4.3. MODELO CROMÁTICO PROPUESTO

Esta sección explica el modelo cromático propuesto, el cual está basado en el espacio de color RGB permitiendo una combinación interactiva de tres diferentes métodos de RD espectrales no supervisados, de manera que un usuario no necesariamente experto pueda hacer uso de un método de RD en específico o una combinación de estos. Una aproximación versátil de métodos espectrales de reducción de dimensión son las aproximaciones kernel.

### 4.3.1. IMÁGENES RGB.

Una imagen normalizada puede ser definida como un arreglo matricial descrito por la función $I: \mathbb{N}^3 \rightarrow [0,1]$ , donde cada par de números $x, y: \mathbb{N}^2$ son conocidos como pixeles y cada valor de intensidad $I(x, y, c)$ es asociado a un pixel $(x, y)$ del canal $c$. Existen diferentes formas de representar la luz reflejada por los objetos, pero la manera más común es la descomposición en colores primarios rojo, verde, azul (RGB). La descomposición es asociada con los valores de intensidad de los canales, los cuales están entre 0 y 1 (si la imagen está normalizada), de modo que un valor de 0 indica la completa ausencia de color (color negro) y el valor 1 es relacionado con el máximo valor de intensidad (color blanco) [33].

$(a)$ Imagen RGB

$(b)$ Nivel de rojo $(c)$ Nivel de verde $(d)$ Nivel azul

**Figura 18.** En (a) se puede observar una imagen de tres canales (imagen RGB). Donde (b) representa el nivel de intensidad del color rojo (canal 1), (c) el nivel de intensidad del color verde (canal 2) y por ultimo (d) representa el numero los valores de intensidad del color azul (canal 3). **Fuente:** [33].

En [22], se habla de la naturaleza visual de los humanos, se explica que a través del sentido de la vista se capta la mayoría de información del mundo exterior, es decir que el ancho de banda de la visión es mayor que el de todos los sentidos combinados. Teniendo en cuenta este enfoque, el modelo cromático saca ventaja

del contenido visual que ofrece la variedad de colores que pueden ser generados por el espacio de color RGB de manera que la mezcla de métodos de RD no sea definida con valores numéricos directamente, de modo que la interacción con el modelo sea más intuitiva he interactiva. Para este propósito se saca provecho de dos propiedades fundamentales de una imagen, la resolución espacial y la resolución de intensidad. La resolución espacial es definida como el número de pixeles que una imagen tiene, este valor puede ser calculado a través de la siguiente expresión $pixeles = m * n$ donde $m$ representa el número de filas y $n$ el número de columnas (el ancho y alto de la imagen). La resolución de intensidad es el rango de valores de intensidad que cada pixel puede tener, para este trabajo la imagen tiene una resolución de 8 bits es decir $2^8 - 1 = 255$ debido a que el valor de intensidad 0 es considerado [33].

### 4.3.2. COMBINACIÓN DE DOS MÉTODOS DE RD A TRAVÉS DE IMÁGENES.

Para entender de mejor manera el concepto del modelo cromático propuesto es necesario explicar primero la mezcla de dos métodos de RD. En la **Figura 19**, se indica una imagen con dos canales ($c_1$ y $c_2$) con una resolución espacial de $m = 256$ filas por $n = 100$ columnas, con esto en mente se puede observar que el número de filas (altura de la imagen) es igual a la resolución de intensidad, lo que significa que cada fila puede tener $256$ valores de intensidad de $0\ a\ 255$, sin embargo, cada canal tiene un cambio de intensidad diferente.

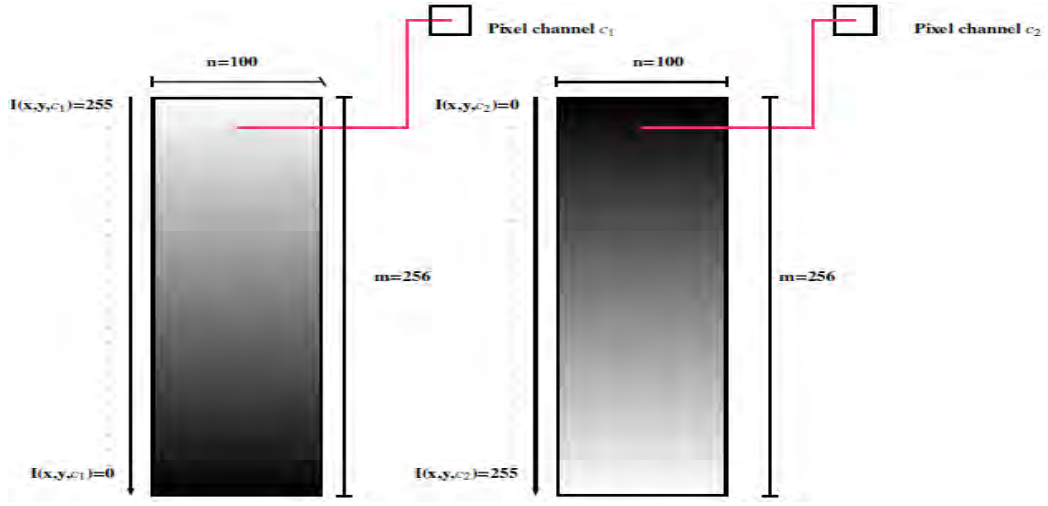



**Figura 19.** Imagen de dos canales, con un valor de intensidad descendente para el canal 1 ($c_1$) y un valor ascendente para el canal 2 ($c_2$) de tal forma que la suma de intensidad de ambos canales sea igual a 255. **Fuente:** Esta investigación**.**

Como se muestra en la **Figura 19,** $c_1$ tiene un cambio decreciente mientras que $c_2$ tiene un cambio creciente. En consecuencia, un pixel de la imagen tendrá dos valores de intensidad debido a sus dos canales. Además, si los valores de

intensidad de los dos canales son sumados el resultado siempre será igual a 255 (1 si una normalización es realizada).

Como se ha dicho anteriormente este trabajo se basa en el espacio de color RGB, esto implica la existencia de tres canales $c_1 = R$, $c_2 = G$, $c_3 = B$ donde cada canal representa un método de reducción de dimensión que es representado a través de una matriz kernel. De forma que, el color rojo representa el primer método RD ($RD_1$), el color verde representa el segundo método RD ($RD_2$) y finalmente el azul representará el método de RD número tres ($RD_3$). La interfaz propuesta permite al usuario escoger múltiples combinaciones de métodos de RD que se verán reflejadas en el rango de colores de una combinación escogida. Para la combinación de dos métodos se define una imagen RGB en la cual todos los pixeles tiene un valor igual a 0 para uno de sus tres canales, en consecuencia, esta imagen puede ser considerada como una imagen con dos canales y la propiedad explicada anteriormente puede ser aplicada. Para la combinación de dos métodos de reducción de dimensión existen tres posibles combinaciones como se evidencia en la **Figura 20**.

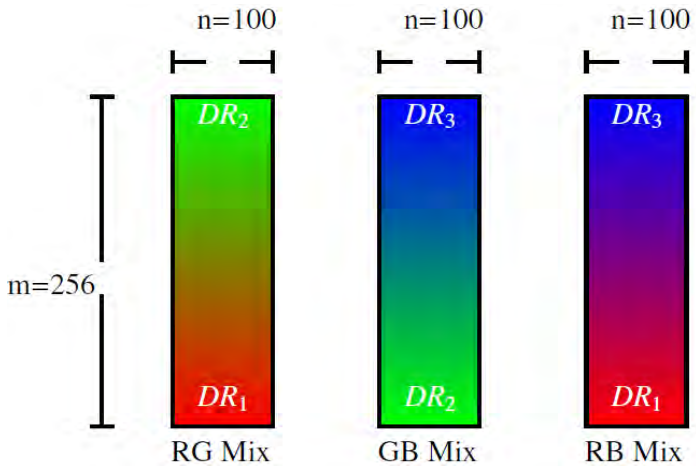



**Figura 20.** Combinaciones posibles para la mezcla ponderada de dos métodos de RD. En este caso uno de los canales de cada imagen tiene valores de intensidad igual a 0. **Fuente:** Esta investigación**.**

### 4.3.3. INTERPOLACIÓN BARICÉNTRICA

Finalmente, para la combinación de tres métodos de RD la misma metodología debe ser aplicada, sin embargo, tres direcciones de cambio deben ser encontradas, para este propósito el modelo está basado en el triángulo de maxwell en donde cada vértice representará un color primario que se verá reflejado en un método RD. Para encontrar un punto dentro del plano triangular se utilizó la técnica conocida como **interpolación baricéntrica** debido a que permite encontrar los valores de

intensidad de los tres canales en un punto dentro del triángulo teniendo en cuenta la distancia de los tres vértices a la cual se encuentra el punto [34]. Cada vértice del triángulo es un método RD en particular, si el usuario se mueve únicamente por el contorno del triángulo uno de los canales de la imagen siempre será igual a 0 por lo tanto, se tendría el caso de la **Figura 20**, ahora si un punto se encuentra dentro de la superficie del triángulo el valor de intensidad de los tres canales dependerá de las áreas de los triángulos que se generan cuando se proyecta una línea del vértice al punto elegido como se muestra en la **Figura 21**.

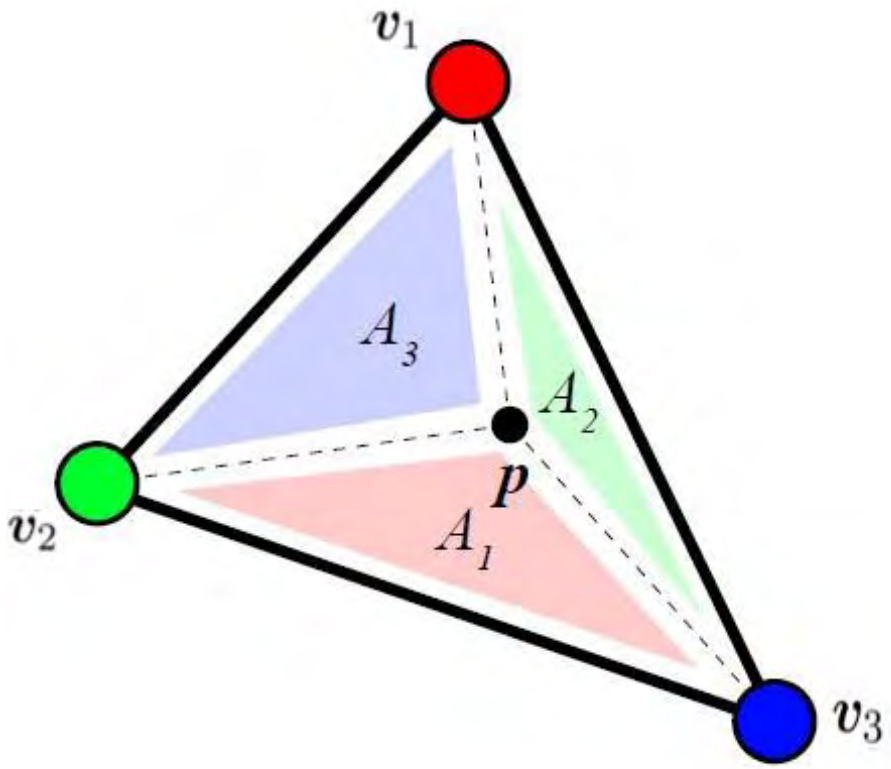



**Figura 21**. Esquema general de la técnica conocida como Interpolación baricéntrica. Donde cada vértice del triángulo es asociado a un color primario y un punto dentro de la superficie representará una combinación de estos con base en las áreas que se generan a partir de líneas proyectadas de los vértices a un punto en específico. **Fuente:** https://classes.soe.ucsc.edu/cmps160/Fall10/resources/barycentricInterpolation.pdf.

Donde $v_1$ es el vértice (pixel) que contiene el máximo valor del color rojo, en este punto los demás colores tienen un valor de intensidad igual a cero. Igualmente esta lógica es aplicada en un punto ubicado en los dos vértices restantes $v_2$ (verde) y $v_3$ (azul). Matemáticamente, los valores de intensidad de un punto dentro de la superficie triangular estarán definidos por la siguiente expresión:

$$R = \frac{A_1}{A} = \alpha_1, G = \frac{A_2}{A} = \alpha_2, B = \frac{A_3}{A} = \alpha_3, \quad (17)$$

con

$$\sum_{i=1}^{3} \alpha_i = 1.$$

El área del triangulo $A_n$ puede ser calculada a través de las coordenadas de sus vértices como se explica a continuación.

Sean $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}$ tres vértices que definen un triángulo con $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3} \in \mathbb{R}^2$ y $\boldsymbol{v_n} = \{v_{na}, v_{nb}\}$, además se tiene un punto $\boldsymbol{p}$ conocido dentro de la superficie triangular de manera que $\boldsymbol{p} \in \mathbb{R}^2$ y $\boldsymbol{p} = \{p_a, p_b\}$**,** entonces las áreas triangulares que se forman a partir de las distancias de los vértices hacia un punto estará definida por la siguiente expresión:

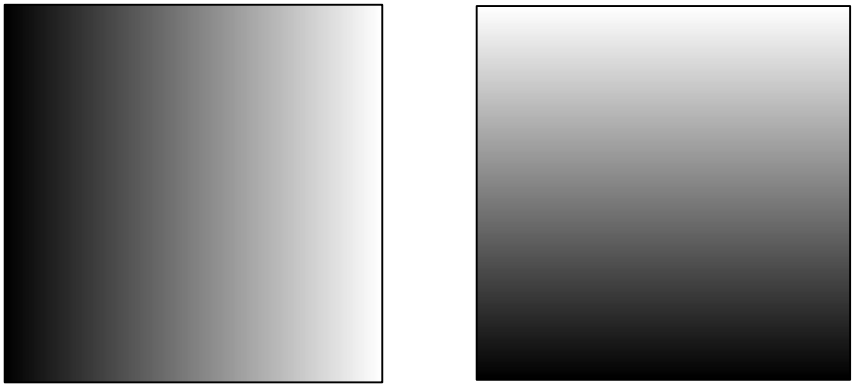
$$A_1 = \frac{1}{2}\left|det\begin{pmatrix} 1 & p_a & p_b \\ 1 & v_{2a} & v_{2b} \\ 1 & v_{3a} & v_{3b} \end{pmatrix}\right|, A_2 = \frac{1}{2}\left|det\begin{pmatrix} 1 & p_b & p_a \\ 1 & v_{1a} & v_{1b} \\ 1 & v_{3a} & v_{3b} \end{pmatrix}\right|, A_3 = \frac{1}{2}\left|det\begin{pmatrix} 1 & p_a & p_b \\ 1 & v_{1a} & v_{1b} \\ 1 & v_{2a} & v_{2b} \end{pmatrix}\right|, \quad (18)$$

donde el operador $det(\cdot)$ esta relacionado con el determinante de una matriz y $|\cdot|$ significa valor absoluto.

De esta forma, con las áreas $\{A_n\}$ generadas a partir de un punto $\boldsymbol{p}$ y el área total $A$ es posiblé obtener los valores RGB de un punto dentro de la superficie.

### 4.3.4. IMPLEMENTACIÓN DEL MODELO CROMATICO PROPUESTO

Como se ha dicho anteriormente el modelo está basado en el espacio de color RGB, el cual es representado a través de una superficie triangular en donde los colores primarios se encuentran en los vértices del triángulo. La implementación del modelo cromático es aplicada a través del enfoque del procesamiento de imágenes aprovechando parámetros básicos la como resolución espacial y la resolución de intensidad [33]. Para la creación del modelo en primera instancia se crean dos cuadriculas de 256 filas por 256 columnas que representan todas las posibles parejas ordenadas de los puntos $x, y$ como se puede observar en la **Figura 22**, donde el plano cartesiano es representado con una imagen y cada punto dentro de este se verá reflejado en un pixel.



$(a)$ Cambios en el eje $x$ $(b)$ Cambios en el eje $y$

**Figura 22**. Plano cartesiano discreto representado por dos imágenes en blanco y negro. El eje $x$ representado por $(a)$ y el eje $y$ representado por $(b)$ tienen valores entre 0-255. **Fuente:** Esta investigación.

Una vez que la cuadrícula es generada, se procede a realizar una máscara binaria de 256 filas por 256 columnas de forma que los valores dentro de la superficie tengan un valor igual a uno y valores fuera de la superficie sean nulos, en consecuencia, solo se tendrán en cuenta los valores dentro de la superficie cuando esta ventana es multiplicada punto a punto con otra imagen. Los vértices del triángulo son $\boldsymbol{v_1} = \{0{,}255\}, \boldsymbol{v_2} = \{255{,}0\}, \boldsymbol{v_3} = \{127{,}255\}$**,** cabe resaltar que debido a que se está trabajando con los pixeles de una imagen estos no tendrán valores continuos si no discretos, por lo tanto, los factores de ponderación en el contorno de la superficie tendrán una variación con respecto a los valores que pueden tomar cuando se trabaja con la combinación de dos métodos de RD.

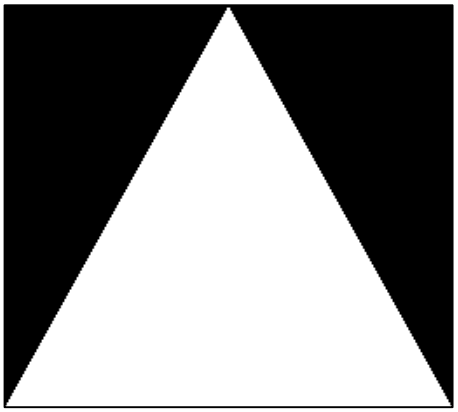
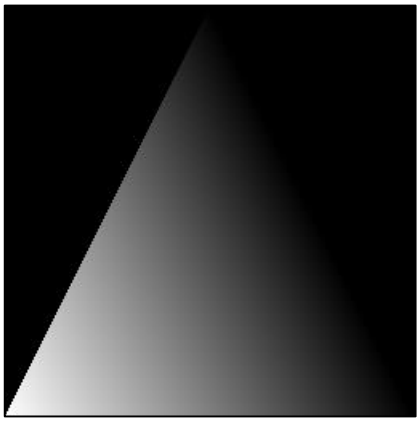


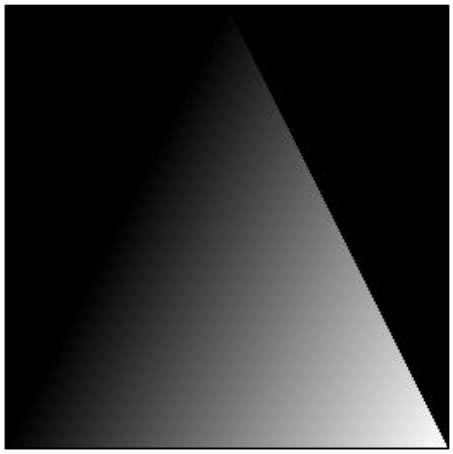
**Figura 23.** Máscara binaria que encierra la superficie triangular que contienen el modelo cromático y toda la gama de colores que conforman el espacio de color RGB. **Fuente:** Esta investigación.

Posteriormente, se aplica la ecuación (18) en toda la cuadricula que representa el plano cartesiano (**Figura 22**) teniendo en cuenta los vértices del triángulo $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}$**,**

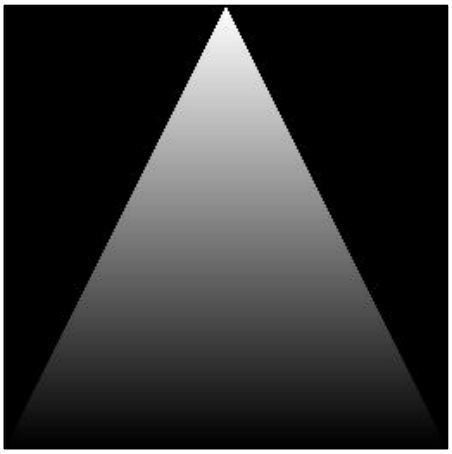
es decir, para un punto $\boldsymbol{p_n}$ dentro de la cuadricula se tendrán tres áreas triangulares que se verán reflejadas en los factores de ponderación, luego se multiplica la máscara de la **Figura 23** por la cuadricula creada y se tendrá como resultado los tres canales que componen el espacio de color RGB como se puede apreciar en la **Figura 24**.



$(a)$ Canal rojo



$(b)$ Canal verde

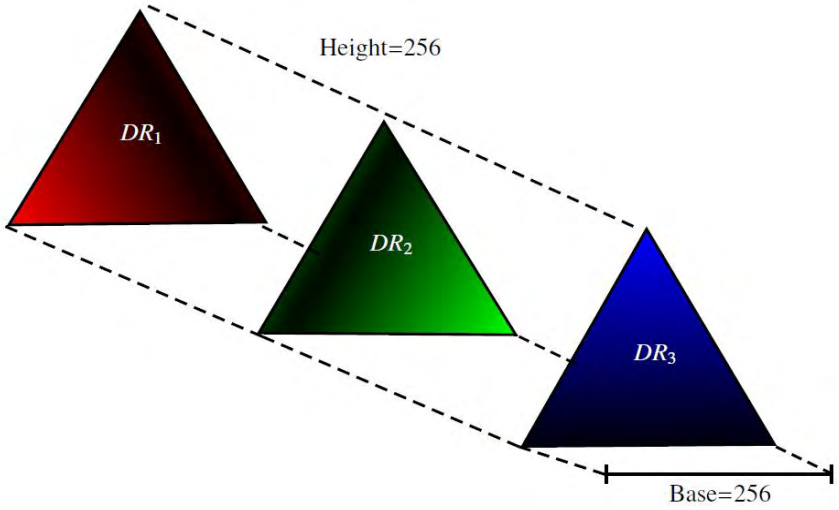


$(c)$ Canal Azul

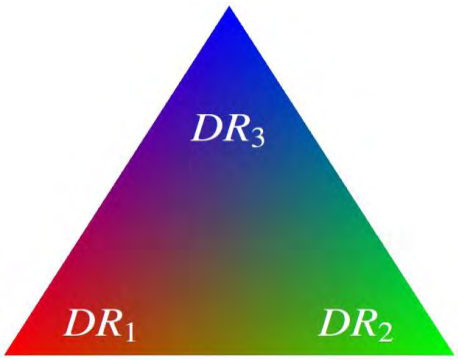
**Figura 24.** Superficies triangulares que representa el canal de los colores primarios $(a)$ rojo, $(b)$ verde y $(c)$ azul (RGB). Los canales son obtenidos al multiplicar una máscara que contenga la superficie triangular y el plano cartesiano ponderado con las áreas de los triángulos que se generan a partir de la relación de distancia entre un punto $\boldsymbol{P_n}$ con respecto a los vértices. **Fuente:** Esta investigación.

Finalmente, al superponer los tres canales (**Figura 25a**) se tendrá como resultado el modelo cromático propuesto con tres colores primarios en los vértices y toda la gama de colores que se pueden generar dentro de la superficie. Como la suma de los factores de ponderación tiene que ser igual a 1 en el modelo cromático no se tendrán en cuenta colores que impliquen la total presencia de dos colores primarios

como por ejemplo el amarillo, el cian y el magenta los cuales tienen un valor de intensidad normalizado de 1 en dos canales. En la **Figura 25** se puede observar el modelo cromático propuesto que será utilizado para la combinación de métodos de RD.




$(a)$ Superposición de canales




$(b)$ Modelo cromático

**Figura 25.** A partir de la superposición de los tres canales (a) se tiene como resultado el modelo cromático propuesto. (b) en donde los vértices representarán un método de RD en particular o una combinación de métodos si se elige un punto dentro de la superficie. **Fuente:** Esta investigación.

### 4.3.5. MEZCLA DE MÉTODOS DE RD

En términos de visualización de datos a través de métodos de RD, los parámetros a ser combinados son las matrices kernel, cada matriz corresponde a cada uno de los $M$ métodos de RD considerados, esto es $\{K^{(1)}, \dots, K^{(M)}\}$. Por consiguiente, se

obtiene una matriz kernel $\widehat{\boldsymbol{K}}$ resultante de la mezcla de las $M$ matrices kernel, tal que:

$$\widehat{K} = \sum_{m=1}^{M} \alpha_m K^{(m)}, \tag{19}$$

definiendo a $\alpha_m$ como el factor de ponderación correspondiente al método $m$ y a $\alpha = \{\alpha_1, \ldots., \alpha_m\}$ como el vector de ponderación. Estos dos parámetros estarán definidos por el color de un punto dentro de la superficie del modelo cromático (**Figura 25b**). Una vez obtenida la combinación lineal (19) se realiza un proceso de maximización de la covarianza (PCA) de los datos de alta dimensión representados por las aproximaciones kernels [26], [31].

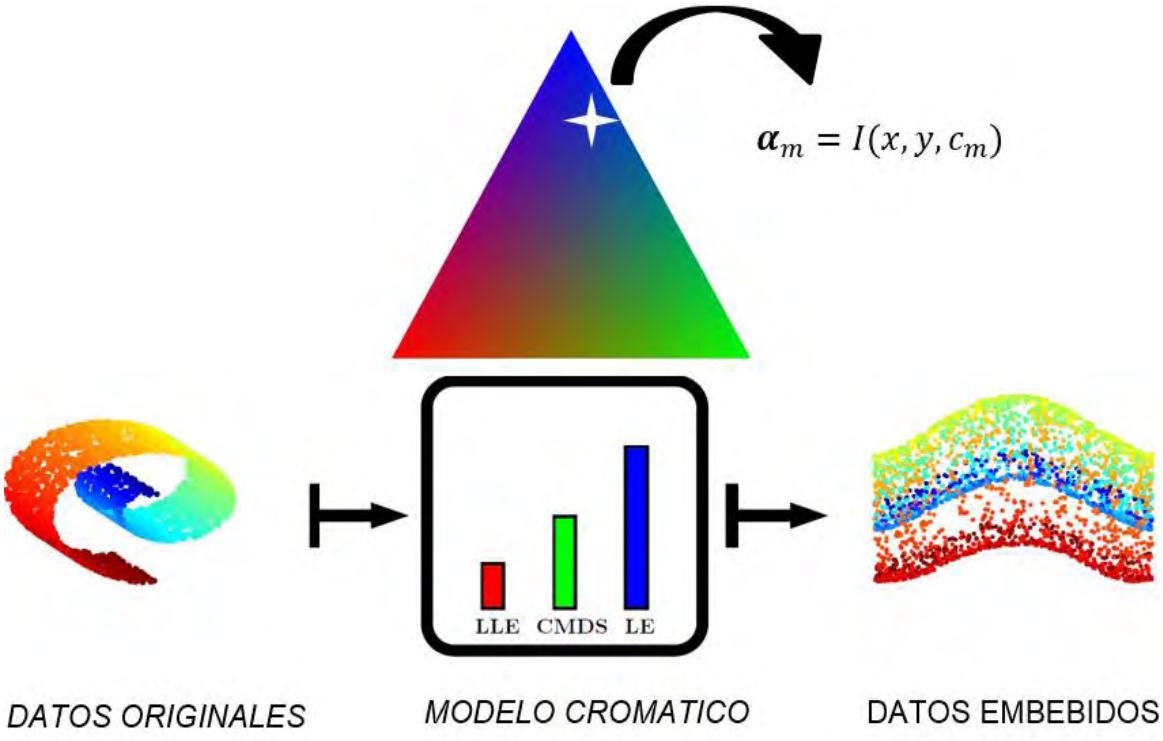



**Figura 26.** Gráfico ilustrativo de cómo se obtienen los factores ponderados que permiten hacer la combinación de tres métodos de RD ($M = 3$). **Fuente:** Esta investigación.

## 4.4. MEDIDA DE CALIDAD

Una parte importante de este trabajo es encontrar una medida de calidad para que los métodos de reducción de dimensión y su combinación puedan ser evaluados. Para este propósito se utiliza un criterio de calidad mediante la conservación de los $k$-ésimos vecinos más cercanos desarrollada en [13]. El objetivo de integrar esta medida de calidad a la metodología de visualización propuesta es determinar el

desempeño de los métodos de RD en cuanto a la preservación de la topología de los datos en el espacio de baja dimensión con respecto al de alta dimensión.

La medida de evaluación propuesta en [13] permite evaluar el espacio embebido de la siguiente manera. El rango de $\xi_j$ respecto a $\xi_i$ en el espacio de alta dimensión se denota como $p_{ij} = |\{k: \delta_{ik} < \delta_{ij} \text{ o } (\delta_{ik} = \delta_{ij} \text{ y } 1 \leq k < j \leq N)\}|$, donde $|\cdot|$ denota la cardinalidad del conjunto. Similarmente, en [13] define que el rango de $x_j$ respecto a $x_i$ en el espacio de baja dimensión es $r_{ij} = |\{k: d_{ik} < d_{ij} \text{ o } (d_{ik} = d_{ij} \text{ y } 1 \leq k < j \leq N)\}|$. Los k-ésimos vecinos de $\xi_i$ y de $x_i$ son los conjuntos definidos por $v_i^k = \{j: 1 \leq p_{ij} < K\}$ y por $n_i^k = \{j: 1 \leq r_{ij} < K\}$, respectivamente. Un primer índice de rendimiento puede ser denotado como:

$$Q_{NX}(K) = \sum_{i=1}^{N} \frac{|v_i^k \cap n_i^k|}{KN}. \quad (20)$$

La ecuación (20) resulta en valores comprendidos entre 0 y 1 y mide el promedio normalizado de acuerdo a los k-ésimos vecinos correspondientes entre los espacios de alta dimensión y baja dimensión. Definiendo de esta manera una matriz de co-clasificación $[Q = q_{NX}]$, $j \leq N - 1$ con $q_{kl} = |\{(i,j): p_{ij} = k \text{ y } r_{ij} = l|$. Por lo tanto, $Q_{NX}(K)$ cuenta k-por-k bloques de Q, el rango preservado (en la diagonal principal) y las permutaciones dentro de los vecinos (en cada lado de la diagonal) [32].

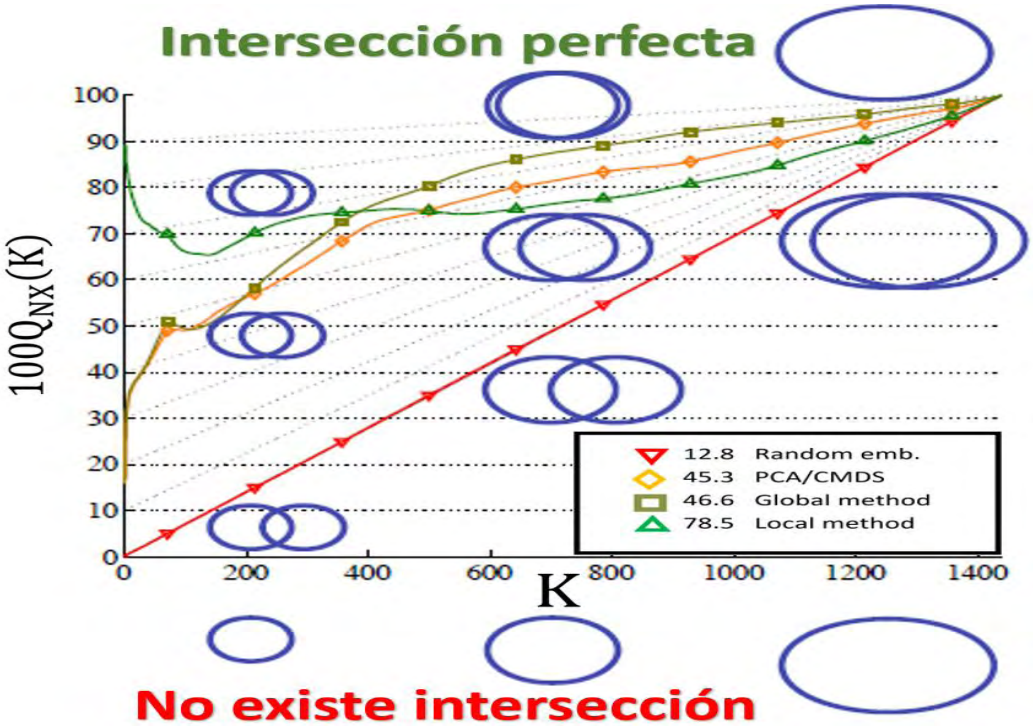



**Figura 27.** Ejemplo de la curva $Q_{NX}(K)$, que permite un primer acercamiento a una medida de calidad para el espacio embebido generado. **Fuente:** Esta investigación.

La **Figura 27** presenta una primera aproximación para evaluar la preservación de la topología de los datos en el espacio embebido debido a un método de RD en particular. Sin embargo, en [32] se realiza un ajuste a la curva $\mathrm{Q_{NX}(K)}$ con el fin de que el área bajo la curva sea un buen indicador de la preservación de la topología de los datos embebidos generados, por lo tanto, la curva de calidad que se integra a la metodología de visualización está dada por:

$$\mathrm{R_{NX}(K)} = \frac{\mathrm{(N-1)Q_{NX}(K)-K}}{\mathrm{N-1-K}}, \tag{21}$$

con el fin de dar una noción al usuario acerca de la calidad de la representación escogida. Un ejemplo de este tipo de curva puede ser observada en la **Figura 28.**

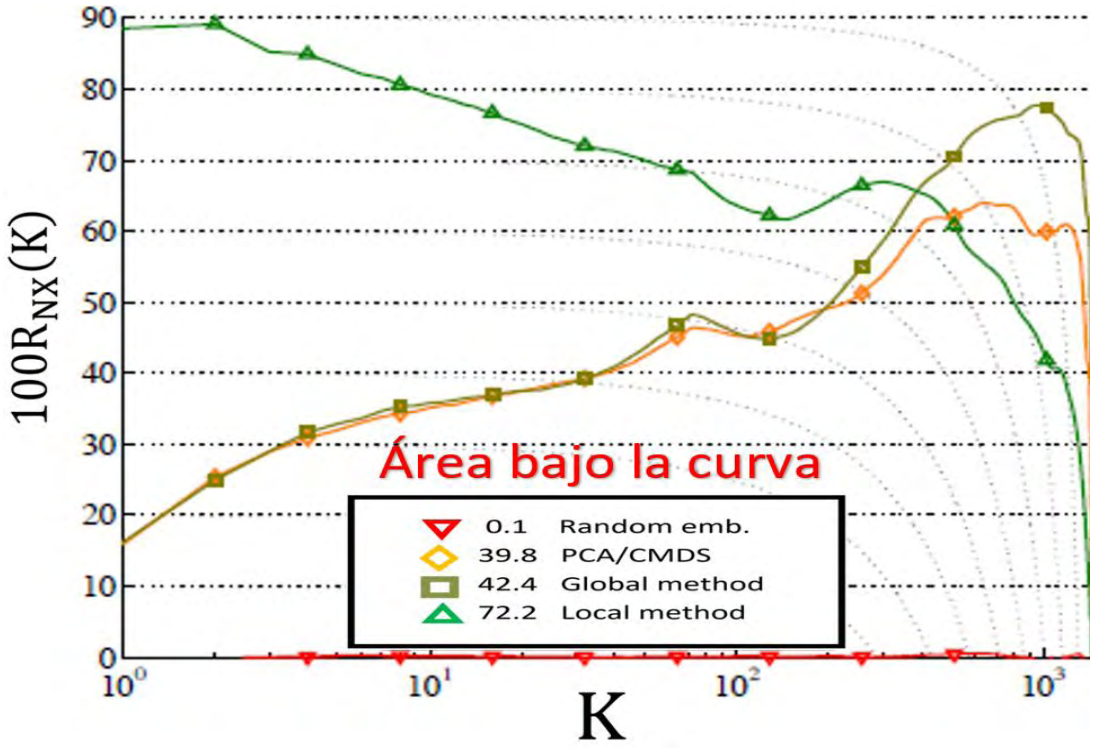
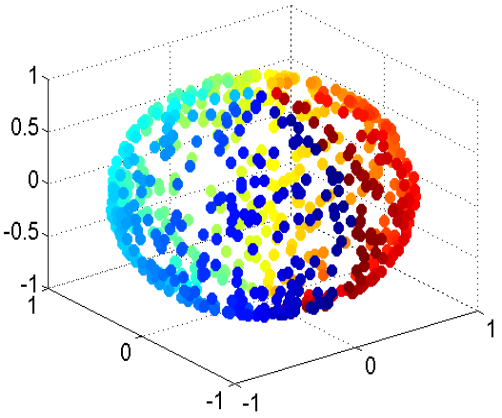



**Figura 28.** Medida de calidad $R_{NX}(K)$ utilizada en la metodología de visualización propuesta, que utiliza el área bajo la curva como un indicador del desempeño del espacio embebido. **Fuente:** Esta investigación.

## 4.5. BASES DE DATOS

En este trabajo las pruebas de la metodología de visualización son llevadas a cabo en cuatro bases de datos, dos bases de datos reales y dos bases de datos artificiales (**Figura 29**). La primera base de datos es una esfera conformada por 700 puntos y tres características (**Figura 29a**). La segunda base de datos es conocida

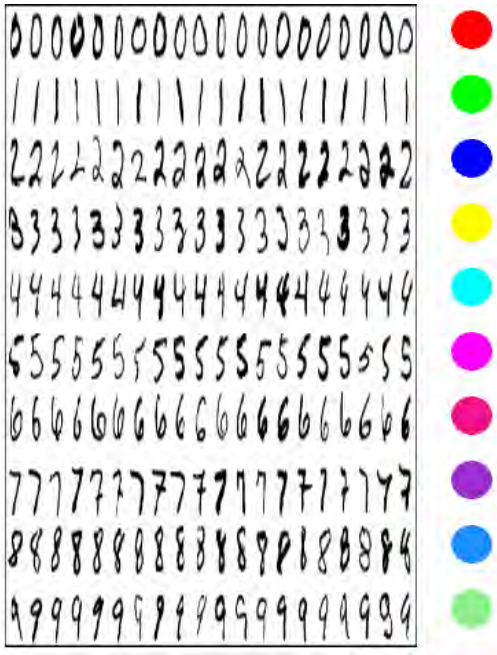
como rollo suizo (**Figura 29b**) y está formada por 700 puntos con tres características. La tercera base de datos COIL-20 (**Figura 29c**) es un banco de imágenes que contiene, 72 imágenes en escala de gris con 20 objetos diferentes (para el experimento N =400 registros –20 objetos en 20 ángulos diferentes con D=16384 –número de pixeles–). El cuarto conjunto de datos es un subconjunto seleccionado al azar del banco de imágenes MNIST (**Figura 29d**), que está formado por 6000 imágenes en escala de grises de cada uno de los 10 dígitos (N = 700 puntos de datos -70 Casos para todos dígitos 10- y D = 784). Figura 5 representa ejemplos de los conjuntos de datos considerados.




$(a)$ Cascaron esférico

$(b)$ Rollo suizo

$(c)$ Coil-20

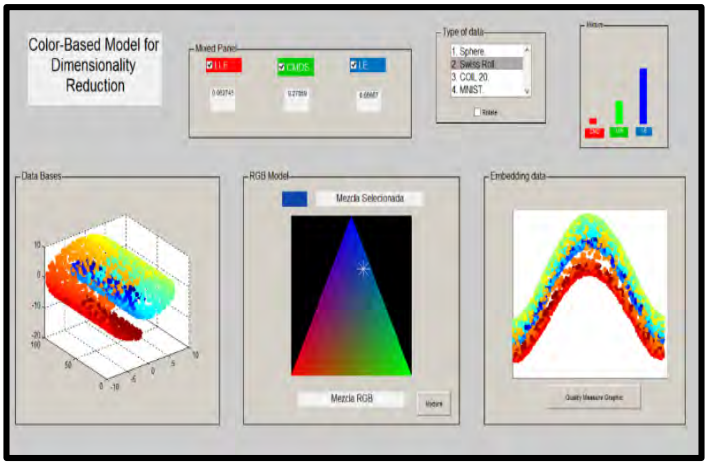


$(d)$ MNIST

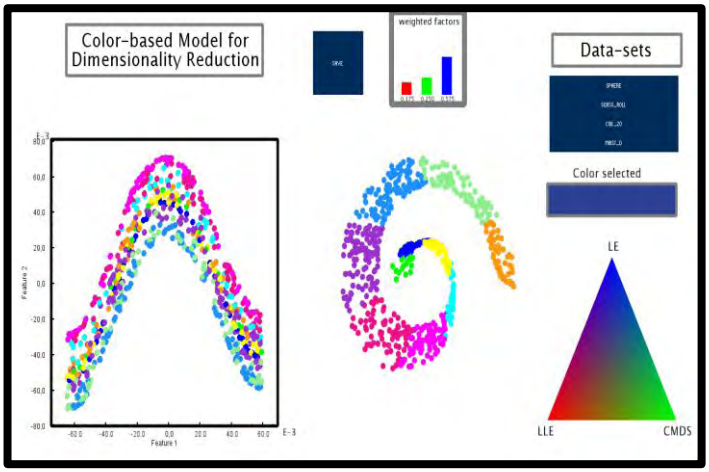
**Figura 29.** Las cuatro bases de datos consideradas para probar la metodología de visualización propuesta. **Fuente:** https://archive.ics.uci.edu/ml/datasets.html

# 5. RESULTADOS

El desarrollo de la metodología de visualización propuesta fue probada a través de la implementación de una interfaz gráfica con un modelo basado en el espacio de color RGB, en dos diferentes entornos de programación MATLAB (**Figura 30**) y Processing [35] (**Figura 30b**) con el fin de observar un posible costo computacional y un tiempo de respuesta más bajo. La evaluación del espacio embebido resultante se realiza a través del criterio descrito en la **sección 4.4**.




$(a)$ Interfaz de Matlab

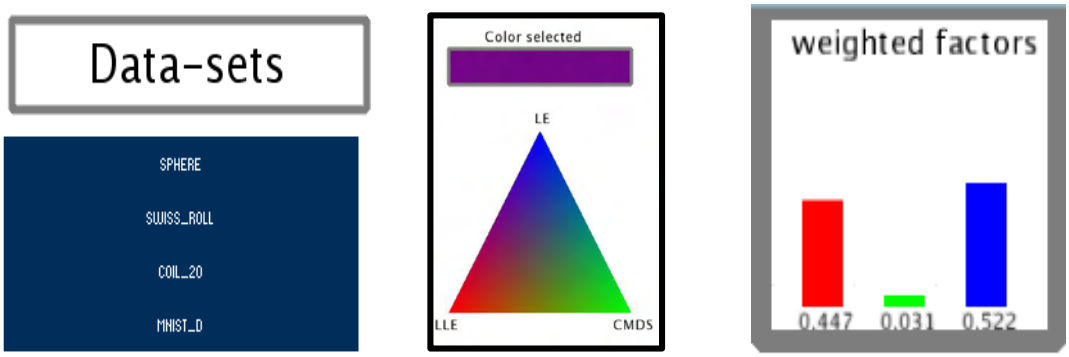



$(b)$ Interfaz de Processing

**Figura 30.** Se indican las interfaces gráficas implementadas en dos entornos de programación diferentes. $(a)$ Es una interfaz gráfica implementada en MATLAB, mientras que la $(b)$ es una interfaz gráfica implementada en el entorno de desarrollo libre Processing. **Fuente:** Esta investigación.

## 5.1. INTERACTIVIDAD DE LA INTERFAZ PROPUESTA.

Como se explica en la **sección 1** existe una necesidad latente de crear herramientas y metodologías de visualización que permitan una alianza entre la máquina y el ser humano, para encontrar conocimiento en forma de patrones que permitan formular conjeturas y tomar decisiones en base a el conocimiento disponible. Las interfaces de usuario creadas en MATLAB y en Processing están diseñadas de tal manera que sean de un uso amigable, es decir, que el usuario pueda aplicar un método de reducción en particular o una mezcla de estos, a través de los colores disponibles en el espacio de color RGB, los cuales serán traducidos en unos factores de ponderación que definen el espacio embebido resultante. Las interfaces desarrolladas en MATLAB y Processing cuentan con las mismas funciones principales.




(a) Base de datos (b) Modelo cromático (c) Factores de ponderación

(d) Diagrama de dispersión (e) Botón de guardado

**Figura 31.** Se indican las diferentes funciones presentes en la interfaz gráfica que permiten el desarrollo de la metodología de visualización propuesta. Entre las principales funciones tenemos cuatro botones para elegir la base de datos $(a)$, el modelo cromático $(b)$, barras de visualización $(c)$, un plano cartesiano $(d)$ y un botón de guardado $(e)$. **Fuente:** Esta investigación.

En primera instancia se tienen cuatro botones (**Figura 31a**) que le permiten al usuario cargar una de las cuatro bases de datos de prueba (**sección 4.5**). Como función principal de la interfaz gráfica se encuentra el modelo cromático (**Figura 31b**) y un rectángulo en la parte superior que indica el color seleccionado dentro del triángulo, además con el fin de ser lo más explícito posible existen unas barras que indican el grado de rojo, verde y azul del color seleccionado además del valor numérico de los factores de ponderación (**Figura 31c**). Uno de los objetivos principales de este trabajo es representar datos de alta dimensión en un diagrama de dispersión bidimensional, de este modo la interfaz cuenta con un plano cartesiano (**Figura 31d**) en donde los datos embebidos pueden ser representados. Por ultimo existe un botón (**Figura 31e**) que permite guardar como imagen PNG la representación de los datos de baja dimensión escogida por el usuario.

## 5.2. CONTROLABILIDAD DE LA INTERFAZ PROPUESTA

Para probar la controlabilidad del sistema se escogieron aleatoriamente cuatro puntos dentro de la superficie del triángulo (**Figura 32)**. Tres de los puntos están cercanos a los vértices y uno aproximadamente en el centro del triángulo. La ventaja de este enfoque radica en que un usuario puede tener un alto número de posibles representaciones de los datos de alta dimensión en un espacio bidimensional, para escoger la representación que más se acomode a sus necesidades. Además, como se ha dicho anteriormente es más fácil para un ser humano analizar información de forma visual, por lo tanto, la selección de los factores de ponderación por medio de colores es coherente con los lineamientos que propone los estudios de la percepción humana en [22]. De modo que, la controlabilidad de la interfaz gráfica se reduce a la interacción del usuario con el modelo de color. Para que el usuario tenga una mejor noción del color seleccionado (método RD) se hizo uso de un rectángulo que indique el color seleccionado y unas barras que especifican el porcentaje de rojo, verde y azul.

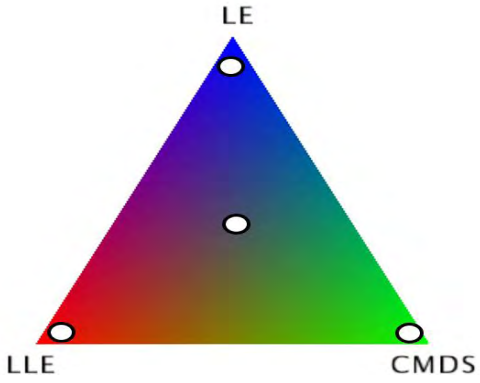


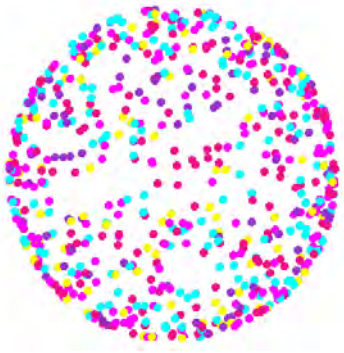

**Figura 32.** Puntos aleatorios escogidos para probar la interactividad y controlabilidad del modelo. Los espacios embebidos resultantes y las respectivas medidas de calidad $R_{NX}(K)$ pueden ser encontradas en **la sección 5.3**. **Fuente:** Esta investigación.

Cabe resaltar que al trabajar con imágenes los cambios que puede sufrir $\alpha_m$ son discretos por lo tanto una resolución de cambio aproximada es definida como, $\alpha_{res} \approx \frac{1}{255}$, por lo tanto, el cambio más pequeño que se puede dar en los factores de ponderación es aproximadamente $\frac{1}{255} = 3{,}92 \times 10^{-3}$.

## 5.3. DATOS EMBEBIDOS RESULTANTES

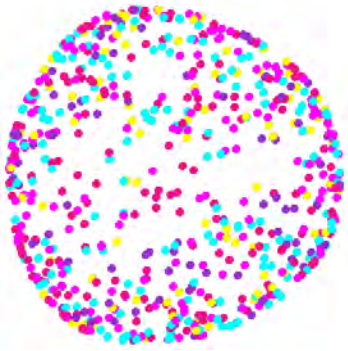
Los experimentos realizados para la obtención de los siguientes resultados pueden ser observados en la **Figura 32**. Cabe resaltar que tres métodos de RD son considerados y son aplicados de acuerdo con los colores seleccionados. El principal objetivo del experimento es observar los espacios embebidos generados a partir de los métodos de RD, además de observar de acuerdo con el criterio $\mathrm{R_{NX}(K)}$ si la mezcla escogida de métodos de RD puede conservar la topología de los datos aún mejor que algunos métodos de RD convencionales. Los resultados se muestran en las **Figuras 33 a 36**.

En los resultados se pueden apreciar los diagramas de dispersión en donde se representan los datos embebidos, además se dan a conocer las curvas de calidad que permiten de alguna manera establecer un criterio sobre el desempeño de la representación en un espacio de baja dimensión, teniendo en cuenta la preservación de la topología de los datos a través del área bajo la curva. De esta manera, si el valor del área bajo la curva es mayor, el rendimiento de los datos embebidos será mejor debido a la conservación de los k-vecinos más cercanos. Un resultado interesante puede observarse en la base de datos conocida como Cascarón esférico en 3-D (**Figura 31**), dado que los espacios embebidos resultantes (**Figura 31a-31d**) no presentan cambios significativos y esto se refleja claramente en la curva de calidad $\mathrm{R_{NX}(K)}$ donde las áreas bajo la curva no difieren mucho, lo que significa que la topología de los datos se preserva de manera similar en los tres métodos de reducción de dimensión y en la mezcla. Por otra parte, en las demás bases de datos si existen diferencias considerables en cuanto a la preservación de la topología de los datos y la forma de los espacios embebidos al escoger bien sea un método de reducción de dimensión en particular o una mezcla en este caso. Un ejemplo claro de las diferentes formas en que una base de datos puede ser deformada por un método de RD en particular puede ser observada en la **Figura 32,** en la base de datos conocida como rollo suizo.

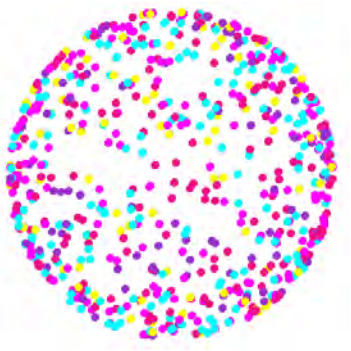


(a) LLE

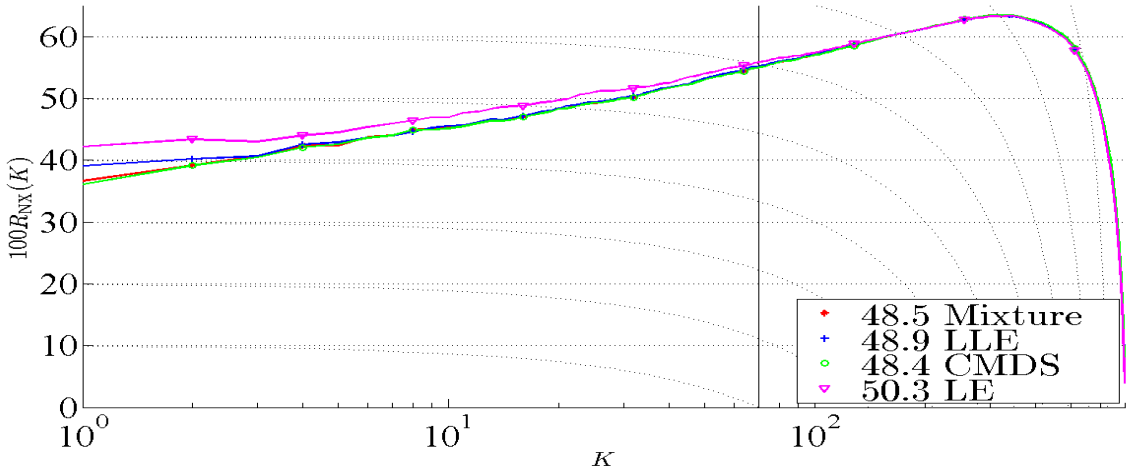
(b) CMDS



(c) LE

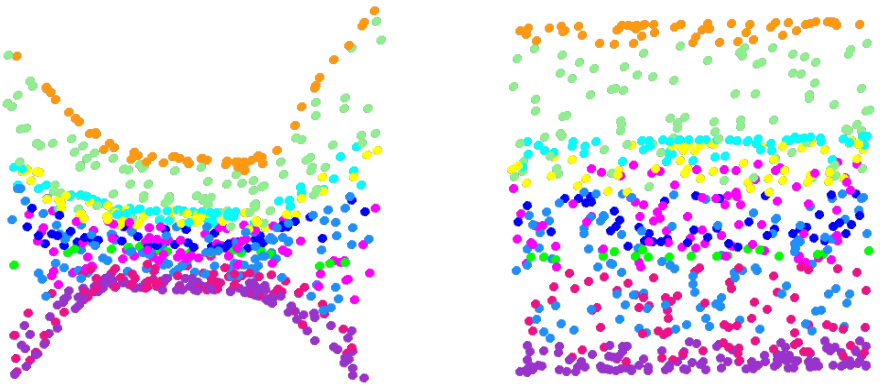


(d) MEZCLA

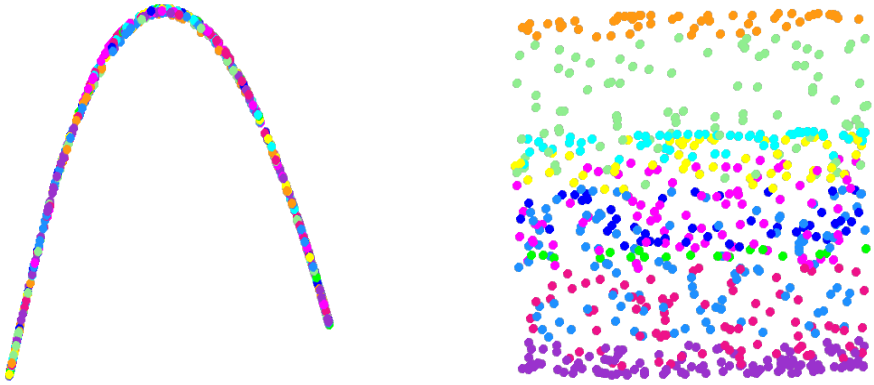



(e) $R_{NX}(K)$

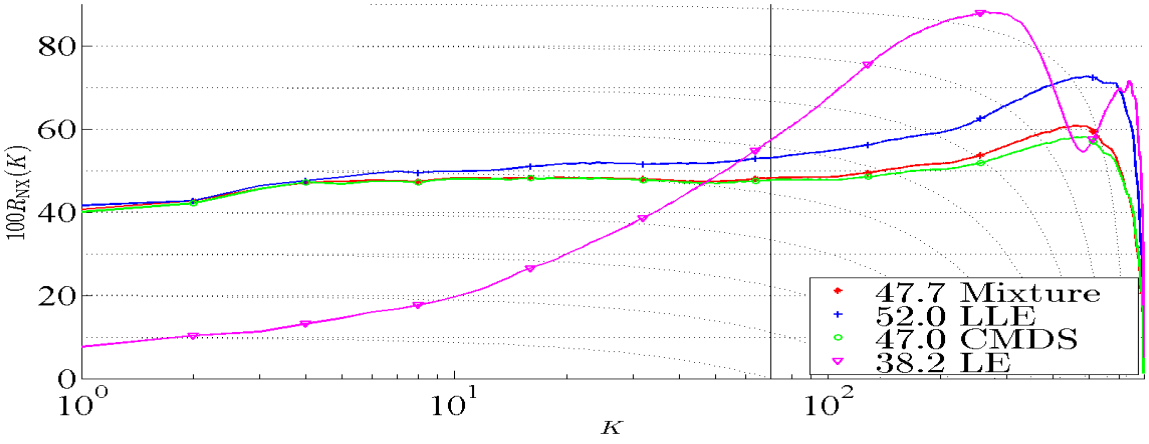
**Figura 33.** Resultados obtenidos a partir de la interfaz gráfica para la base de datos cascarón esférico en 3D. Las figuras $(a)$-$(d)$ indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios dentro del modelo cromático. La figura $(e)$ indica la medida de calidad $R_{NX}(K)$. **Fuente:** Esta investigación.



(a) LLE (b) CMDS



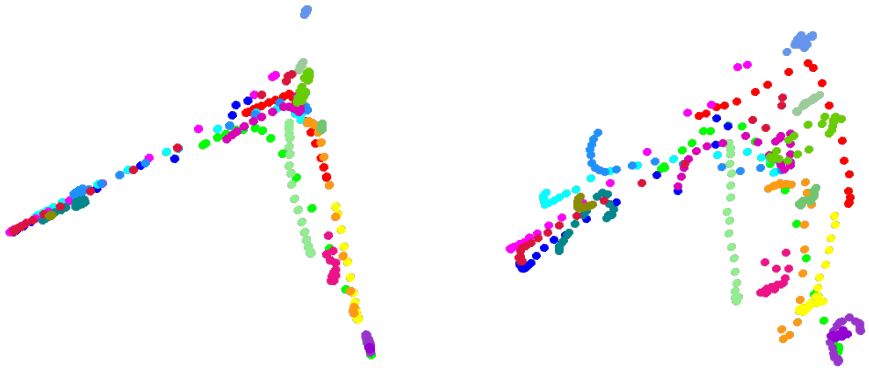
(c) LE (d) MEZCLA



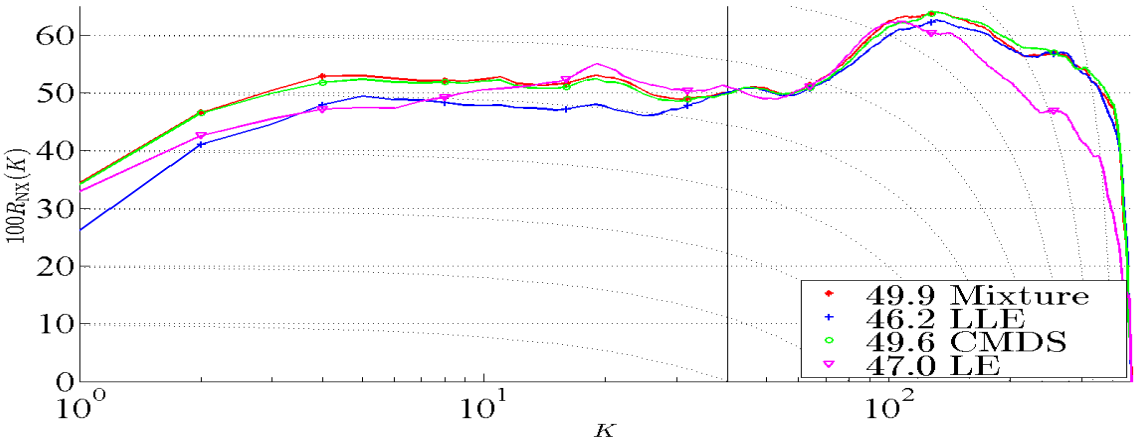

(e) $R_{NX}(K)$

**Figura 34.** Resultados obtenidos a partir de la interfaz gráfica para la base de datos conocida como rollo suizo. Las figuras $(a)$-$(d)$ indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios dentro del modelo cromático. La figura $(e)$ indica la medida de calidad $R_{NX}(K)$. **Fuente:** Esta investigación.

(a) LLE (b) CMDS

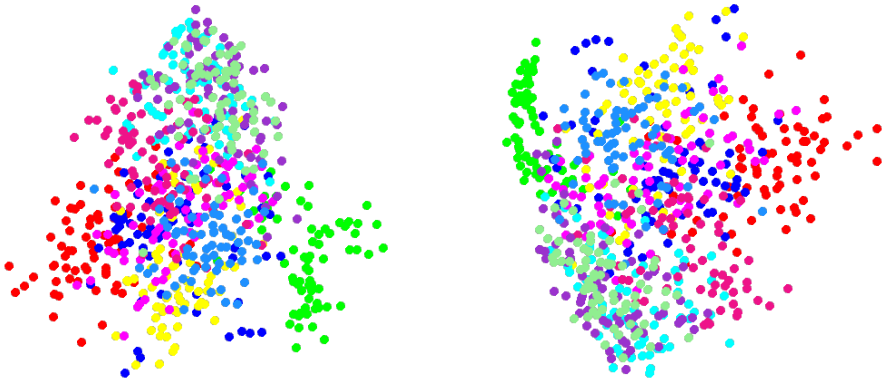


(c) LE (d) MEZCLA

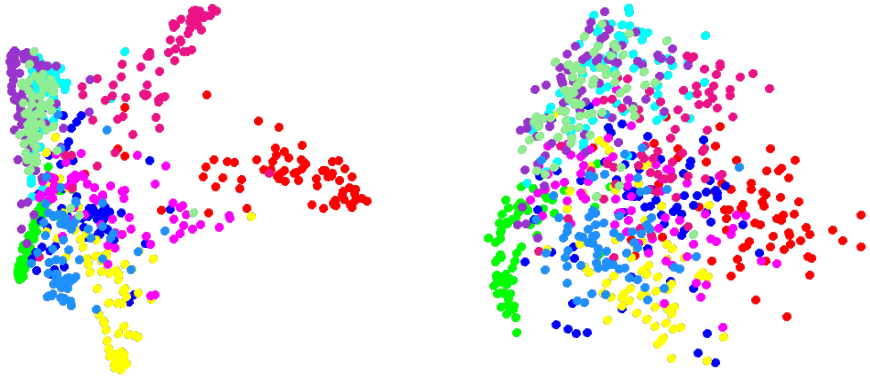



(e) $R_{NX}(K)$

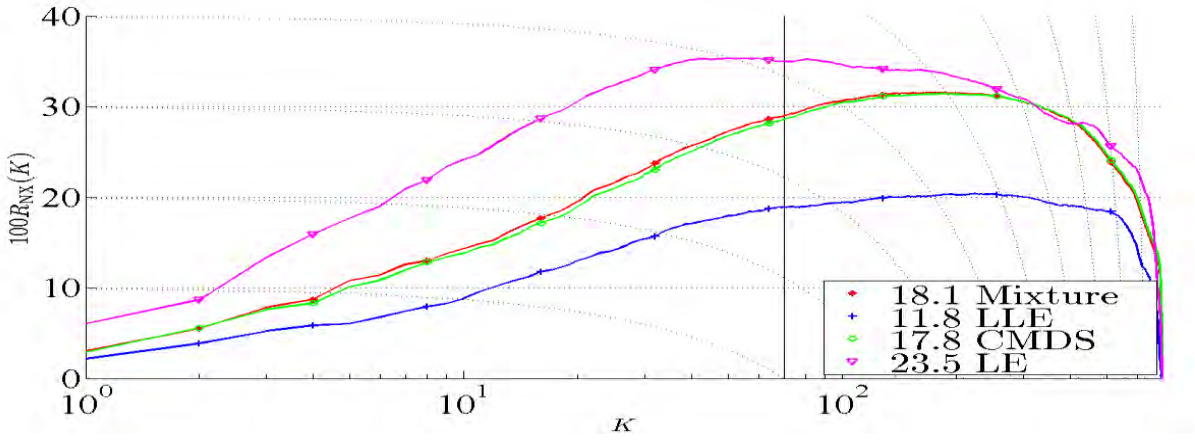
**Figura 35.** Resultados obtenidos a partir de la interfaz gráfica para la base de datos Coil-20. Las figuras $(a)$-$(d)$ indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios dentro del modelo cromático. La figura $(e)$ indica la medida de calidad $R_{NX}(K)$. **Fuente:** Esta investigación.



(a) LLE (b) CMDS



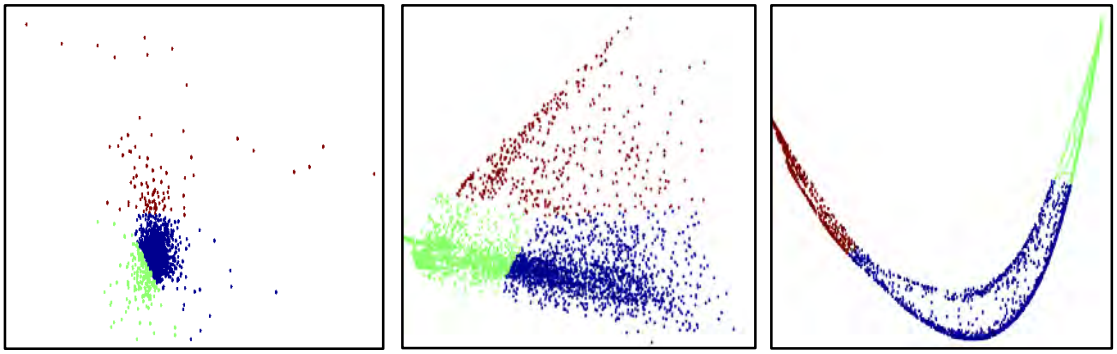
(c) LE (d) MEZCLA



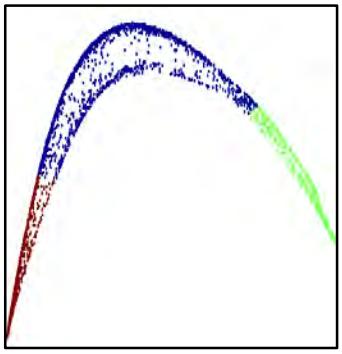

(e) $R_{NX}(K)$

**Figura 36.** Resultados obtenidos a partir de la interfaz gráfica para la base de datos MNIST. Las figuras $(a)$-$(d)$ indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios dentro del modelo cromático. La figura $(d)$ indica la medida de calidad $R_{NX}(K)$. **Fuente:** Esta investigación.

Un ejemplo de aplicación real de la metodología de visualización propuesta es realizada con una base de datos del proyecto **ALTERNAR**[1] generada a partir de imágenes satelitales de **Lansat 8** para una pequeña parte del municipio de Tumaco. La base de datos está conformada por 4483 puntos (muestras) en donde cada punto representa un terreno de treinta por treinta metros con 7 bandas (7 dimensiones o mediciones) que representan los diferentes tipos de radiación que son captados por los sensores del satélite. Las pruebas realizadas son las mismas mostradas en la **Figura 32,** es decir los métodos de RD por separado y una mezcla de los tres en partes iguales. Una vez que se tienen los espacios embebidos se aplica un algoritmo de clasificación no supervisado k-medias [24] con tres centroides para observar en un espacio embebido bidimensional la forma en que los datos son agrupados. Los resultados en el diagrama de dispersión pueden ser visualizados en la **Figura 37**.



$(a)$ LLE $(b)$ CMDS $(c)$ LE

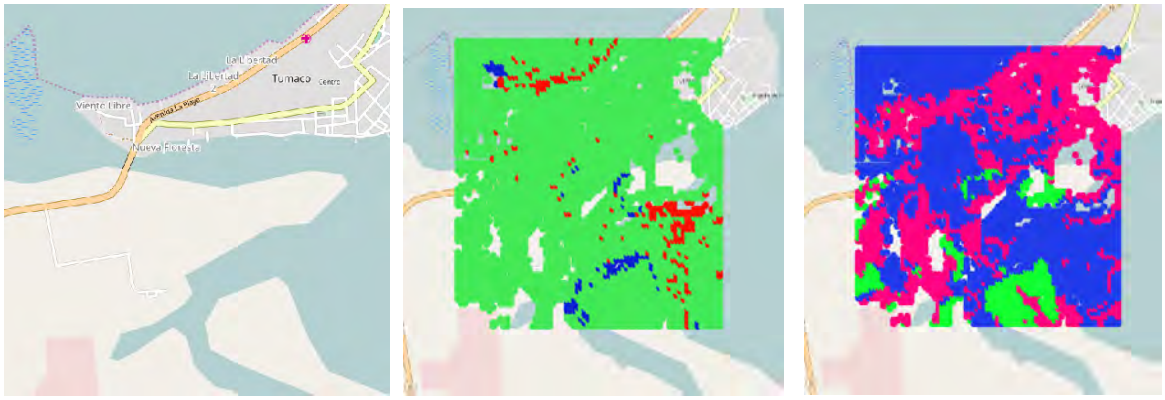


$(d)$ MEZCLA

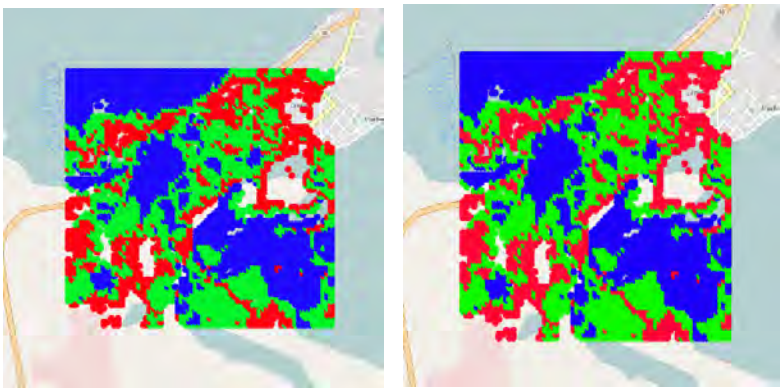
**Figura 37.** Resultados obtenidos para la base de datos que contiene las 7 bandas de una pequeña región en Tumaco. Las figuras (a)-(d) indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios, los vértices del triángulo y un punto central. **Fuente:** Esta investigación.

[1] **http://190.254.4.127:90/alternar**

Luego con las coordenadas y las clases definidas por el algoritmo de clasificación k-medias se pueden ubicar los puntos agrupados en el mapa de la región seleccionada, con el fin de observar los efectos en el agrupamiento de los datos después de aplicar métodos de RD. De este modo, la metodología de visualización propuesta también puede ayudar a realizar tareas de clasificación en donde el usuario pueda explorar el modelo cromático y en este caso observar cual es el espacio embebido que define de una mejor manera tres tipos de terreno como, por ejemplo, agua, ciudad o vegetación. Los resultados de ubicar los puntos en el mapa pueden ser observados en la **Figura 38**. Visualmente el espacio embebido que mejor define el rio y la parte urbana de Tumaco es el método **CMDS**.




$(a)$ Región $(b)$ LLE $(c)$ CMDS



$(d)$ LE $(e)$ MEZCLA

**Figura 38.** Resultados gráficos obtenidos a partir de ubicar las coordenadas de los puntos de la base de datos de Lansat 8 en el mapa. Donde las Figuras $(b)$-$(d)$ representan el agrupamiento debido a un método de RD en específico y la Figura $(d)$ a una mezcla de métodos de RD. **Fuente:** Esta investigación.

# 6. CONCLUSIONES

La metodología propuesta representa una forma alternativa de visualización de datos, donde métodos de reducción de dimensión son aplicados de manera que se pueda apreciar bajo algunos criterios, la naturaleza de los datos de alta dimensión en un espacio bidimensional. Si bien la aplicación de métodos de reducción de dimensión puede ser ambigua para el usuario inexperto, la implementación de un modelo intuitivo basado en el espacio de color RGB se convirtió en una gran herramienta que permite reducir la brecha existente entre los usuarios y las bases de datos permitiendo, un fácil uso de la interfaz y la aplicación de métodos de RD por medio de colores de manera que se tengan en cuenta conceptos básicos de la percepción humana. En este sentido tres diferentes métodos de reducción de dimensión pueden ser seleccionados o combinados a través de colores y no directamente con coeficientes, con el fin, de que la interfaz gráfica sea más cercana al usuario.

Un aspecto importante es que los efectos producidos por los métodos de RD sobre una base de datos varían de acuerdo con la naturaleza de los datos que se desea visualizar, como por ejemplo en las bases de datos cascarón esférico en 3D y rollo suizo los efectos de los métodos de RD y la mezcla tienen repercusiones muy diferentes. En el cascarón esférico en 3D los espacios embebidos son muy similares en cambio en el rollo suizo el espacio bidimensional varía mucho con cada método RD aplicado. Por lo tanto, los espacios de baja dimensión varían ampliamente de acuerdo con la base de datos que se desea analizar.

Con el fin de buscar un costo computacional más bajo se emigró la interfaz gráfica implementada en MATLAB al entorno de desarrollo Processing, sin embargo, los tiempos de ejecución observados fueron similares lo que significa que si se quiere mejorar el tiempo de respuesta de la interfaz gráfica se tiene que afrontar esta problemática desde dos frentes: la optimización de algoritmos para generar las matrices kernel o emigrar la interfaz gráfica a un entorno de programación con un nivel más bajo.

Las aproximaciones kernel son una adecuada y versátil forma de implementar métodos espectrales de reducción de dimensión, sin embargo, las matrices kernel al ser funciones de distancia tendrán una dimensión de $\boldsymbol{N} \times \boldsymbol{N}$ donde $\boldsymbol{N}$ es el número de registros que una base de datos tiene. Esto representa un claro inconveniente al momento de trabajar con bases de datos con un número elevado de registros, debido a que la matriz kernel será muy densa y aplicar el algoritmo generalizado de PCA supondrá un costo computacional muy elevado al momento de calcular los valores y vectores propios de la matriz kernel. Si bien este trabajo se enfocó en abordar el problema de la alta dimensión de los datos, un trabajo futuro seria

encontrar una forma de optimizar la implementación de las matrices kernel, así como encontrar algoritmos que permitan calcular los valores y vectores propios con un costo computacional más bajo.

La necesidad actual de procesar y analizar los grandes volúmenes de información que se generan en actualidad ha permitido que los aportes científicos, que ayuden a mejorar la manera como se procesa y representa la información sean tenidos en cuenta en eventos científicos importantes. Una muestra de ello es la publicación de tres artículos científicos en tres eventos diferentes a partir del desarrollo de este trabajo de grado.

# RECOMENDACIONES

Debido a la naturaleza interdisciplinar de la generación de bases de datos, es importante explorar otras metodologías y modelos de interacción usuario-máquina que permitan crear herramientas de visualización de un uso intuitivo con el fin de que usuarios sin conocimiento previo puedan manipular los datos para obtener representaciones gráficas que más se ajusten a las necesidades del usuario dentro de una tarea en específica.

Es importante seguir explorando métodos de reducción de dimensión que permitan crear espacios embebidos en donde se pierda la menor cantidad de información posible, así como una buena conservación de la topología de los datos. Además de aproximaciones de métodos de reducción de dimensión que permitan obtener un costo computacional más bajo.

Si bien el modelo cromático solo permite la combinación de tres diferentes métodos de reducción de dimensión, la interpolación baricéntrica puede ser extendida para más métodos de reducción de dimensión es decir se pueden tener en cuenta más vértices y tener como resultado un espacio de baja dimensión más amplio.

Muchas veces un espacio embebido bidimensional o tridimensional no es suficiente para representar una base de datos de alta dimensión, por lo tanto, los resultados obtenidos en el diagrama de dispersión pueden no ser los más precisos, por lo que se requiere graficar más dimensiones. En consecuencia, es recomendable buscar técnicas de representación de datos que permitan graficar una base de datos de más de tres dimensiones de una manera más sencilla y diciente para el usuario.

## BIBLIOGRAFÍA


[1] M. Sedlmair y M. Aupetit, «Data-driven Evaluation of Visual Quality Measures,» Computer Graphics Forum, vol. 34, nº 3, pp. 201-210, 2015.

[2] M. Sedlmair, M. Brehmer, S. Ingram y T. Munzner, «Dimensionality reduction in the wild: Gaps and guidance,» Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep., 2012.

[3] M. Sedlmair, T. Munzner y M. Tory, «Empirical Guidance on Scatterplot and Dimension Reduction,» IEEE Transactions on Visualization and Computer Graphics, vol. 19, nº 12, pp. 2634-2643, 2013.

[4] D. H. Peluffo Ordoñez, J. A. Lee y M. Verleysen, « Recent methods for dimensionality reduction: a brief comparative analysis,» European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2014.

[5] D. H. Peluffo-Ordónez, J. A. Lee y M. Verleysen, «Dimensionality reduction methods based on stochastic neighbour embedding. In Advances in Self-Organizing Maps and Learning Vector Quantization,» Springer International Publishing, pp. 65-74, 2014.

[6] D. H. Peluffo Ordóñez, A. E. Castro Ospina, J. C. Alvarado Pérez y E. J. Revelo Fuelagán, «Multiple Kernel Learning for Spectral Dimensionality Reduction,» IberoAmerican Congress on Pattern Recognition (CIARP), pp. 626-634, 2015.

[7] D. Z. L. Sacha, M. L. J. A. Sedlmair, J. Peltonen, D. Weiskopf y D. A. Keim, «Visual interaction with dimensionality reduction: a structured literature analysis,» IEEE Transactions on Visualization and Computer Graphics, 2016.

[8] D. H. Peluffo-Ordóñez, J. A. Lee y M. Verleysen, «Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In Computational Intelligence and Data Mining,» Computational Intelligence and Data Mining (CIDM), pp. 171-177, 2014.

[9] S. T. Roweis y L. K. Saul, «Nonlinear dimensionality reduction by locally linear embedding.,» Science, vol. 290, nº 5500, pp. 2323-2326, 2000.

[10] S. H. Bae, J. Qiu y G. Fox, «High performance multidimensional scaling for large high-dimensional data visualization,» IEEE Transaction of Parallel and Distributed System., 2012.

[11] Y. Aflalo y R. Kimmel, «Spectral multidimensional scaling.,» Proceedings of the National Academy of Sciences, vol. 110, nº 45, pp. 18052-18057, 2013.

[12] M. Belkin y P. Niyogi, «Laplacian eigenmaps for dimensionality reduction and data representation.,» Neural computation, vol. 15, nº 6, pp. 1373-1396, 2003.

[13] J. A. Lee y M. Verleysen, «Quality assessment of dimensionality reduction: Rank-based criteria,» Neurocomputing, vol. 72, nº 7, 2009.

[14] J. Han, J. Pei y M. Kamber, de Data mining: concepts and techniques, Elsevier, 2011.

[15] G. Shmueli, N. R. Patel y P. C. Bruce, Data Mining for Business Analytics: Concepts, Techniques, and Applications in XLMiner, 2016.

[16] M. Scholz, «Approaches to analyse and interpret biological profile data,» Universitat Potsdam, 2006.

[17] W. Dai y P. Hu, «Research on Personalized Behaviors Recommendation System Based on Cloud Computing,» Indonesian Journal of Electrical Engineering and Computer Science, vol. 12, nº 2, pp. 1480-1486, 2013.

[18] González-Torres, A., F. J. García-Peñalvo y R. Therón, «Human–computer interaction in evolutionary visual software analytics,» Computers in Human Behavior, vol. 29, nº 2, pp. 486-495, 2013.

[19] J. R. Harger y P. J. Crossno, «Comparison of open-source visual analytics toolkits,» In IS&T/SPIE Electronic Imaging International , 2012.

[20] M. Bostock y J. Heer, «Protovis: A graphical toolkit for visualization,» IEEE transactions on visualization and computer graphics , pp. 1121-1128, 2009.

[21] M. L. Kersten, S. Idreos, S. Manegold y E. Liarou, «The researcher's guide to the data deluge: Querying a scientific database in just a few seconds.,» PVLDB Challenges and Visions, 2011.

[22] C. Ware, Information visualization: perception for design, Elsevier, 2012.

[23] U. Fayyad, G. P.-Shapiro y P. Smyth, «Data mining to knowledge discovery in databases.,» AI Magazine, vol. 17, nº 3, pp. 37-54, 1996.

[24] C. M. Bishop, Bishop Pattern Recognition and Machine Learning., New York: Springer, 2006.

[25] D. H. Peluffo Ordóñez, C. Alzate, J. A. Suykens y G. Castellanos-Domínguez, «Optimal Data Projection for Kernel Spectral Clustering,» European Symposium on Artificial Neural Networks (ESANN), 2014.

[26] J. Ham, D. D. Lee, S. Mika y B. Schölkopf, «A kernel view of the dimensionality reduction of manifolds.,» Proceedings of the twenty-first international conference on Machine learning (ICML), p. 47, 2004.

[27] J. Gijón Gómez, «Visualización bidimensional de problemas de clasificación en alta dimensión,» PROYECTO FIN DE CARRERA. UNIVERSIDAD CARLOS III DE MADRID, 2013.

[28] T. Nasser y R. S. Tariq, «Big Data Challenges,» J Comput Eng Inf Technol 4: 3., vol. 9307, p. 2, 2015.

[29] M. A. Belabbas y P. J. Wolfe, «On landmark selection and sampling in high-dimensional data analysis.,» Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 367, nº 1906, pp. 4295-4312, 2009.

[30] N. Cristianini y J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, UK: Cambridge university press, 2000.

[31] L. A. Belanche, «Developments in kernel design,» EESANN, 2013.

[32] J. Cook, I. Sutskever, A. Mnih y G. E. Hinton, «Visualizing Similarity Data with a Mixture of Maps,» AISTATS, vol. 7, pp. 67-74, 2007 .

[33] R. C. Gonzalez y R. E. Woods, Digital image processing, (2002).

[34] M. Meyer, A. Barr, H. Lee y M. Desbrun, «Generalized barycentric coordinates on irregular polygons. Journal of graphics tools,» Journal of graphics tools, vol. 7, nº 1, pp. 13-22, 2002.

[35] D. Shiffman, Learning Processing, Elsevier, 2008.

[36] D. H. Peluffo Ordoñez, J. A. Lee, M. Verleysen, J. L. Rodriguez y G. Castellanos-Dominguez, «Unsupervised relevance analysis for feature extraction and selection. A distance-based approach for feature relevance,» International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2015.

[37] S. Bergner, M. Sedlmair, T. Moller, S. N. Abdolyousefi y A. Saad, «Interactive parameter space partitioning for computer simulations,» IEEE transactions on visualization and computer graphics, vol. 19, nº 9, pp. 1499-1512, 2013.

# ANEXOS

Esta sección ha sido destinada a los resultados tangibles logrados con el trabajo realizado en esta tesis. Estos anexos contienen una descripción más ampliada de los resultados mencionados en la sección 5.

## ANEXO 1. LISTA DE ACRÓNIMOS

**DCBD** Descubrimiento de conocimiento en base de datos

**KDD** *Knowledge Discovery in Databases*

**RD** Reducción de dimensión

**DR** *Dimensionality Reduction*

**RGB** Rojo (*red*), verde (*green*) y azul (*blue*)

**3-D** Tres dimensiones

**2-D** Dos dimensiones

**PCA** Análisis de componentes principales (*principal component analysis*)

**LLE** *Locally Linear Embedding*

**CMDS** *Classical Multidimensional Scaling*

**LE** *Laplacian Eigenmaps*

**KPCA** *Kernel principal component analysis*

**STSIVA** *Symposium on Signal Processing, Images and Artificial Vision*

**CIARP** *Iberoamerican Congress on Pattern Recognition*

**LA-CCI** *Latin American Conference on Computational Intelligence*

## ANEXO 2. PSEUDOCODIGO DEL SCRIPT DE PROGRAMACION

1. **Inicio**

2. **Escribir** “Escoger base de datos:”; % 1=Esfera, 2=Rollo, 3=Coil-20, 4=MNIST.
3. **Leer** opt;
4. **Switch** opt;
    a. **Caso 1:** crear esfera;
    b. **Caso 2:** crear rollo suizo;
    c. **Caso 3:** cargar Coil-20;
    d. **Caso 4:** cargar MNIST;
5. **Hacer** aproximaciones Kernel;
6. **Hacer** M = {Kernel 1, Kernel 2, Kernel 3} **%** Matriz embebida con kernels.
7. **Graficar** datos;
8. **Hacer** Generar modelo cromático;
9. **Hacer** asociar Kernel a colores primarios RGB; %Cada color es un método RD.
10. **Escribir** “Seleccione un pixel dentro de la superficie triangular:”;
11. **Leer** R, G, B; % valores de rojo, verde y azul del pixel dentro del modelo.
12. **Leer** R=R/255, G=G/255, B=B/255; % Se normalizan los valores de intensidad.
13. **Hacer** Pesos= [R, G, B]; **%** Se definen los factores de ponderación.
14. **Para** i=1 hasta i=3 **Hacer** Ksum = Ksum + Pesos(j)*M{j};
15. **Hacer** [V, D] = eigs(Ksum); % Calcular vectores y valores propios.
16. **Hacer** D = diag(D); % Se obtienen los valores propios de la diagonal principal.
17. **Hacer** [~, índices] = sort(D); % Se ordena de manera descendente los valores propios.
18. **Hacer** Xsum_o=Xsum(:indices); % Se asocia el mayor valor propio con respectivo vector propio.
19. **Graficar** Xsum_o; % Graficar matriz de baja dimensión solo tomando los dos primeros vectores propios.

## ANEXO 3. IMPLEMENTACIÓN DEL MODELO CROMÁTICO Y LAS MATRICES KERNEL EN MATLAB

```
%%%%%%%%%%%%%%%%%%%%%%%%%%% Data sets %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% 1. Esfera.
% 2. Rollo suizo.
% 3. MNIST
% 4. Coil 20

switch opt

    case 1
        spd = 3;
        nbr =700;
        X = randn(nbr,3);
        X = bsxfun(@rdivide, X, sqrt(sum(X.^2,2)));
        if spd>3
            X = [X,0.5*randn(nbr,spd-3)./sqrt(spd-3)];
        end
        L = 32+64/180*atan2(X(:,1),X(:,2));
        str = ['Sph',num2str(spd)];
        save(['data_',str],'X','L','str');
        colmap = hsv(64);
        figure(1);
        scatter3(X(:,1),X(:,2),X(:,3),60,L,'o','filled');
        Y = X;

    case 2

        s = RandStream('mcg16807','Seed',29);
        RandStream.setGlobalStream(s);
        str = 'swissroll';
        N =700;
        t = (3*pi*(rand(N,1).^0.65)+pi/2);
        height = 100*rand(N,1);
        Y = [t.*cos(t) height  t.*sin(t)];
        L = t;
        figure(1);
        scatter3(Y(:,1),Y(:,2),Y(:,3),50,t,'o','filled')

    case 3

        load 'mnist_train_1.mat'
        X = train_X;
        L = train_labels;
        [~,numcl,~]=unique(L);
        nbr=70;
        L1=[];
        Y=[];
```

```
        for i=1:10
          Lr=L(numcl(i):(numcl(i)+(nbr-1)));
          L1=[L1;Lr];
          Yr=X(numcl(i):(numcl(i)+(nbr-1)),:);
          Y=[Y;Yr];
        end
       L=L1;

    case 4

        load 'coil_1440.mat'
        L = reshape(bsxfun(@plus, (1:20), zeros(72,1)), [1440,1]);
        stp = 1;
        X = X(1:stp:end,:);
        L = L(1:stp:end);
        nr = 20;
        nc = 9;
        [r,c] = meshgrid(1:nr,1:nc);
        mosaic1([c(:),r(:)],X(1:8/stp:1440/stp,:),128,128,9,20);
        [~,numcl,~]=unique(L);
        nbr=20;
        L1=[];
        Y=[];
        for i=1:20
          Lr=L(numcl(i):(numcl(i)+(nbr-1)));
          L1=[L1;Lr];
          Yr=X(numcl(i):(numcl(i)+(nbr-1)),:);
          Y=[Y;Yr];
        end
       L=L1;
end


%%%%%%%%%%%%%% Kernels %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Kernel 1: LLE method

d = 2;
knn = round(size(Y,1)*0.10);
[X_LLE, ~,M,conn_comp] = lle(Y, d, knn);
L   = L(conn_comp);
Y   = Y(conn_comp,:);
nbr       = length(conn_comp);
lamb      = eig(M);
lamb      = max(lamb);
K_LLE     = lamb*eye(nbr) - M;
K         = K_LLE;
K         = 0.5*(K + K');
kS        = sum(K,1)./nbr;
K1        = K - bsxfun(@plus, kS, kS') + sum(kS)/nbr;
K1        = K1/max(max(abs(K1)));
```

```
%% Kernel 2: CMDS method

DX = pairwisedistances(Y);
DXL = DX;
% compute scalar products from squared distances
S0 = DX.^2;
sS = sum(S0,1)./nbr;
S0 = -1/2*(S0 - bsxfun(@plus, sS, sS') + sum(sS)/nbr); % double centering
K_CMDS = -0.5*(eye(nbr) - ones(nbr))*DX.^2*(eye(nbr) - ones(nbr));
kS     = sum(K_CMDS,1)./nbr;
K_CMDS = K_CMDS - bsxfun(@plus, kS, kS') + sum(kS)/nbr;
K      = K_CMDS;
K2     = 0.5*(K + K');
K2      = K2/max(max(abs(K2)));

%% Kernel 3: LE method
rng(1);
% Type of Laplacian and dimension of latent space
nl = 0; d = 2;
[Wp,beta] = x2p(Y',knn); Wp = (Wp+Wp')/2;
tic; [XLE, LL] = lapeig(d,Wp,nl); t1 = toc;
try
       K_LE   = pinvs(LL);
catch
       K_LE   = pinv(LL);
end
K      = K_LE;
K      = 0.5*(K + K');
kS     = sum(K,1)./nbr;
K3     = K - bsxfun(@plus, kS, kS') + sum(kS)/nbr;
K3     = K3/max(max(abs(K3)));

%% Normalization

K1 = K1./repmat(max(K1),size(K1,1),1);
K2 = K2./repmat(max(K2),size(K2,1),1);
K3 = K3./repmat(max(K3),size(K3,1),1);

%%%%%%%%%%%%%%%%%%%%% Modelo %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
x = [ 0 255 127]; % se definen los vértices en el eje x
y = [ 0 0 255];  % se definen los vértices en el eje y
T = 0.5*det([ones(3,1) x' y']); % área total del triangulo
[xg,yg] = meshgrid(linspace(0,255,256)); % se define el plano cartesiano
yg = flipud(yg); % se define el plano cartesiano
z = inpolygon(xg,yg,x,y); % se crea la máscara triangular
figure(1);
imshow(z); % se visualiza la máscara triangular
L1 = 0.5*(x(2)*y(3)-x(3)*y(2)-xg*(y(3)-y(2))+yg*(x(3)-x(2)))/T; % Canal R
L2 = 0.5*(x(3)*y(1)-x(1)*y(3)+xg*(y(3)-y(1))-yg*(x(3)-x(1)))/T; % Canal G
L3 = 0.5*(x(1)*y(2)-x(2)*y(1)-xg*(y(2)-y(1))+yg*(x(2)-x(1)))/T; % Canal B
rgb = cat(3,L1.*z,L2.*z,L3.*z); % superponen los canales
figure(2) % se visualiza la máscara triangular el modelo cromático
imshow(rgb); % se visualiza la máscara triangular el modelo cromático
imwrite(rgb,'RGB.png') % se guarda el modelo cromático
```

## ANEXO 4. CODIGO DEL PROGRAMA PARA ANALISIS VISUAL (Processing)

Para la implementación de la metodología de visualización propuesta en Processing se importan desde MATLAB las matrices kernel y el modelo cromático que han sido guardados como archivos csv y png respectivamente. Otro aspecto importante fue el uso de librerías de java para el cálculo de los vectores y valores propios (***papaya statistics library***[2]) además de una librería con una gran cantidad de gráficos (***gwoptics***[3]) 2-D y 3-D. Para propósitos de este trabajo y en cumplimiento de los objetivos se hace uso del grafico conocido como diagrama de dispersión (***Scatter plot***).

El código fuente de la aplicación en Processing y las librerías necesarias para el correcto funcionamiento, además, del archivo ejecutable para probar la aplicación sin necesidad de tener un programa en específico instalado en el equipo, están disponibles en la página web descrita en el **ANEXO 6.**

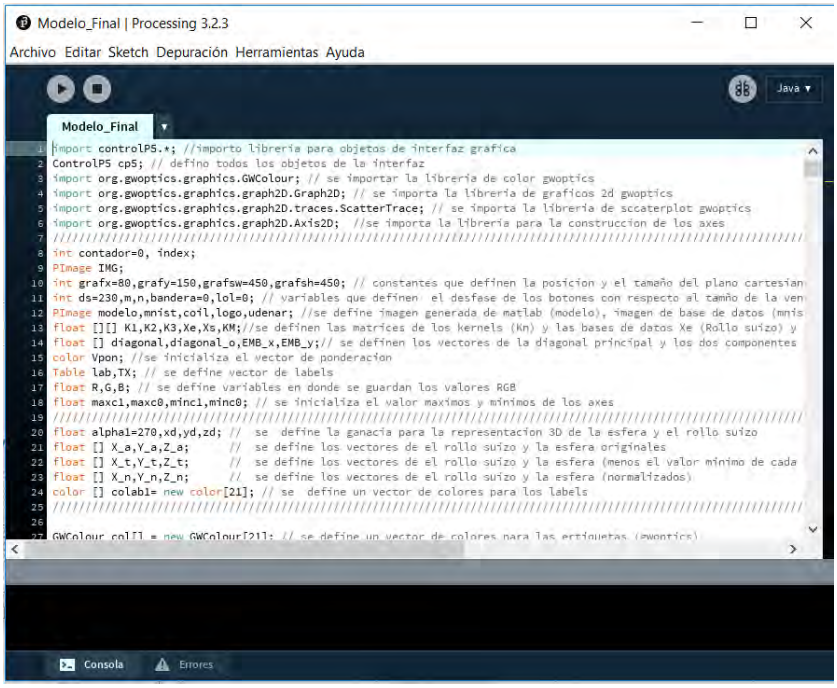



**Figura 39**. Sktech principal de la interfaz gráfica que contiene el modelo cromático en el entorno de desarrollo integrado Processing. Este archivo y las librerías necesarias están disponibles en la página web. **Fuente:** Esta investigación.

[2] http://www.adilapapaya.com/papayastatistics/
[3] http://www.gwoptics.org/processing/gwoptics_p5lib/

# ANEXO 5. ARTICULO: *SYMPOSIUM ON SIGNAL PROCESSING, IMAGES AND ARTIFICIAL VISION* (STSIVA)

## Interactive Visualization Methodology of High-Dimensional Data with a Color-Based Model for Dimensionality Reduction


**Diego F. Peña-Unigarro**, **Jose A. Salazar-Castro**
Universidad de Nariño
Pasto, Colombia
Universidad Nacional de Colombia sede Manizales
Manizales,Colombia
diferpun@gmail.com,
alejo26st@udenar.edu.co

**Diego H. Peluffo-Ordóñez**, **Paul D. Rosero-Montalvo**, **Omar R. Oña-Rocha**, **Andrés A. Isaza**
Universidad Técnica del Norte,
Instituto Tecnologico 17 de Julio
Ibarra,Ecuador
Universidad Surcolombiana,
Universidad Tecnológica de Pereira
Pereira,Colombia
dhpeluffo@utn.edu.ec,
pdrosero@utn.edu.ec,
oronia@utn.edu.ec,
andres.anaya@usco.edu.co

**Juan C. Alvarado-Pérez**, **Roberto Theron**
Universidad de Salamanca
Salamanca, España
Corporación Universitaria Autónoma de Nariño
Pasto, Colombia
jcalvarado@usal.es,theron@usal.es



### Abstract

*Nowadays, a consequence of data overload is that world's technology capacity to collect, communicate, and store large volumes of data is increasing faster than human analysis skills. Such an issue has motivated the development of graphic ways to visually represent and analyze high-dimensional data. Particularly, in this work, we propose a graphical interface that allow the combination of dimensionality reduction (DR) methods using a chromatic model to make data visualization more intelligible for humans. This interface is designed for an easy and interactive use, so that input parameters are given by the user via the selection of RGB values inside a given surface. Proposed interface enables (even non-expert) users to intuitively either select a concrete DR method or carry out a mixture of methods. Experimental results proves the usability of our interface making the selection or configuration of a DR-based visualization an intuitive and interactive task for the user.*


### 1. Introduction

The transformation of high-dimensional data into a lower-dimensional version that preserves as much information as possible from the original data is a research area widely studied [17, 18], given its ability to reduce the computational cost and/or improve the performance of both pattern recognition and information visualization systems [12, 13]. In spite of the existence of tools reaching efficiency indicators in terms of computational performance, exploration and representation of high dimensional data, they lack of properties like interactivity and controllability. Therefore, it is required an expert intervention providing prior knowledge to the system for testing DR techniques as well as interpreting their results being no always readily understood [12,15]. In consequence there is a gap between the users knowledge and the database to be analyzed [16, 18]. The reduction of this gap is the premise that this research is based on.

This paper attempts to jointly take advantage of techniques from the field of DR and concepts from the information visualization aiming to enable the user (not necessarily expert) to directly interact with the database. Doing so, users can get an overview of the data in order to draw conclusions and make decisions [16]. This paper presents an intuitive model that allows the combination of three DR methods providing both interactivity and controllability. Proposed model is based on the RGB color space, where every primary color (red (R), green (G),and blue (B)) represents a particular DR method while the whole range of colors derived from the combination will be reflected in the mixture of DR methods. To do so, conventional DR methods are implementing through kernel approximations [10, 11, 15], which are combined to reach a final kernel matrix. Finally, such a kernel matrix feeds a generalized algorithm of kernel principal component analysis (KPCA) [10]. The benefit of this approach is that user may utilize DR methods over the data, even with no knowledge about the theoretical foundations behind them. The user control the results by just exploring an intuitive, color-based interface. This chromatic model uses the color points within a surface, defining the degree or level at which the DR methods (Kernel matrices)


978-1-5090-3797-1/16/$31.00 ©2016 IEEE

are used, that is, the set of weighting factors. Such surface is a superposition of channels to form the full range of colors and a point on the surface is translated into an RGB value, which defines the mixture of the kernels. This approach allows to evaluate visually the behavior of the low-dimensional data regarding the kernel mixture. The chromatic model proposed in this paper is evaluated using three DR methods, namely: locally linear embedding (LLE) [14], multidimensional classical scaling (CMD) [3] and laplacian eigenmaps (LE) [2]. The experiments are performed over two real databases (images of objects - COIL 20 digits - MNIST) and two artificial databases (spherical shell and Swiss roll) [1]. The DR performance is quantified by a scaled version of the average agreement rate between K-ary neighborhoods explained in [8].

The rest of this paper is organized as follows: Section 2 outlines data visualization via dimensionality reduction. Section 3 describes the Proposed color-based model for the combination of DR methods. Experimental setup and results are presented in Sections 4 and 5, respectively. Finally, some final remarks are drawn in section 6.

## 2. Data visualization via dimensionality reduction

Visualization is the first stage of data analysis where the goal is to make sense of the data before proceeding with others steps like modeling, classification and analysis [18]. Given a large set of measured variables, an obvious idea is to reduce the attributes or features in the measurements by representing them with a smaller set of more condensed variables [16]. Dimensionality reduction allows the extraction of lower dimensional, relevant information from big collections of data aimed at improving the performance of a pattern recognition system or allowing for intelligible data visualization. In other words, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq i\leq N}$, such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq i\leq N}$, being $\boldsymbol{x}_i \in \mathbb{R}^d$, where $d < D$ [12, 13]. In Figure 1 the effect of one DR method is shown.

Classical DR approaches were conceived following an intuitive criterion, such as variance preservation (principal component analysis - PCA) or distance preservation (classical multidimensional scaling - CMDS) [3]. Nowadays, more developed, recent methods are aimed at preserving the data topology. Such a topology is often given by a data-related graph, built as a non-directed and weighted one, in which data points represent the nodes, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. This representation is exploited by both spectral and divergence-based methods. On one hand, for spectral approaches, similarity matrix can represent the weighting factor for pairwise distances as happens in Laplacian eigenmaps [2]. On the other hand, once normalized, it can also represent a probability distribution. The latter is the case of the methods based on divergences such as stochastic neighbor embedding [13].

(a) 3D sphere (b) Embedded data from sphere

Figure 1. Dimensionality reduction effect over an artificial 3d sphere, when The DR method called LLE is applied.

## 3. Proposed model for interactive dimensionality reduction using a color-based approach

This Section describes the proposed model, here so-called, color-based or chromatic model is based on RGB color space and enables an interactive combination of three different spectral unsupervised DR methods, for an (even inexperienced) user, allowing the improvement of the data visualization procedure. A suitable and versatile approximation for spectral DR methods are kernel matrices because they make a linear combination feasible [10, 11, 15]. Our approach works as follows: A normalized image can be defined as a matrix array described by the function $I : \mathbb{N}^3 \rightarrow [0, 1]$, where every pair of numbers $x, y : \mathbb{N}^2$ are known as pixels and each value of $I(x, y, c)$ is associated with the pixel intensity $(x, y)$ of channel $c$ [5]. Decompositions are associated with channels $c$, whose values are between 0 and 1 if they are normalized, where 0 indicates complete absence of that channel (black color) and 1 is related with the maximum intensity (white color) [5]. This model takes advantage of two properties of RGB images: spacial resolution and intensity resolution. On one hand, spacial resolution is defined as the number of pixels that an image contains and can be calculated by the $N_p = m * n$ where $m$ and $n$ are the number of rows and the number of columns, respectively [5]. On the other hand, intensity res-

olution is the intensity values that each pixel can take. For this model a 8-bit intensity resolution is taken, this means that there are $2^8 - 1 = 255$ intensity values since 0 value is considered [5].

In Figure 2, a two-channel image is considered where the spacial resolution is $m = 256$ rows by $n = 100$ columns, with this in mind it can be seen that $m$ is equal to the intensity resolution, then each row can have 256 different intensity values from 0 to 255, nevertheless each channel has a different direction of change, $c_1$ has a decreasing change whereas $c_2$ has a increasing change. If an intensity value is taken from the image, two values of intensity are obtained due that there are two channels. Also if the intensity values are added the result is always equal to 225 (1 if a normalization is made).

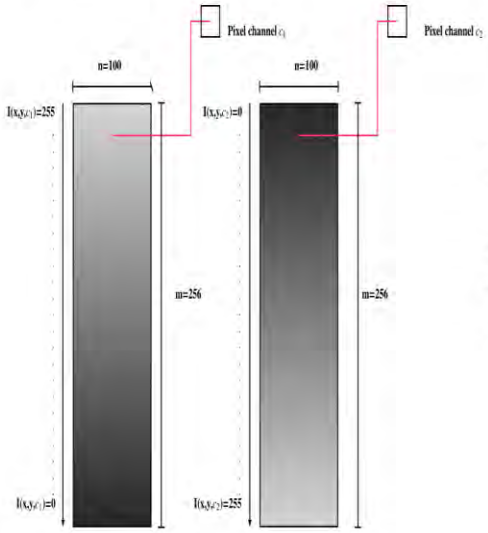



Figure 2. A two channels image.

This paper works with a RGB space, therefore the images have three channels $c_1 = R$, $c_2 = G$, $c_3 = B$ and each channel represents a DR method that in turn has been represented by kernel matrices. Red channel represents the first DR method ($DR_1$), green channel represents the second DR method ($DR_2$) and finally blue channel represents the third DR method ($DR_3$). The proposed interface enables the user to choose multiple combinations of DR methods which will be reflected in the range of colors from the chosen combination. Firstly, a combination of two methods is made with an RGB image which has one of the three channels with intensity value equal to 0 for all pixels, in consequence this image can be considered as a two-channel image and the property explained above can then be applied. For instance, with the combination of two DR methods there are three possible combinations as seen in Figure 3.

Finally, for the combination of three DR methods the

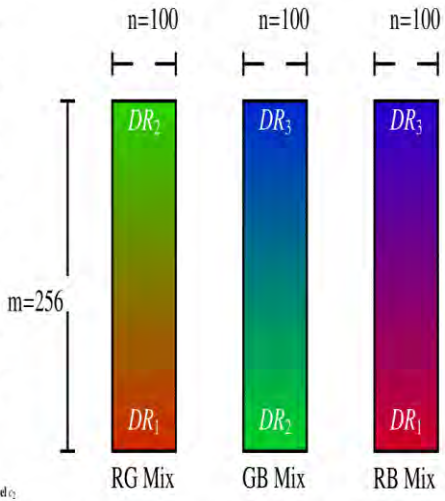
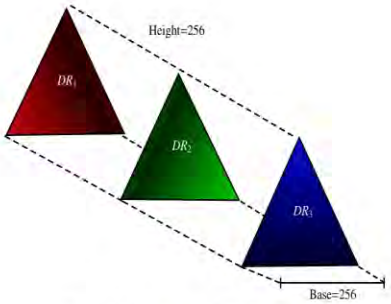



Figure 3. Possible combinations of two DR methods.

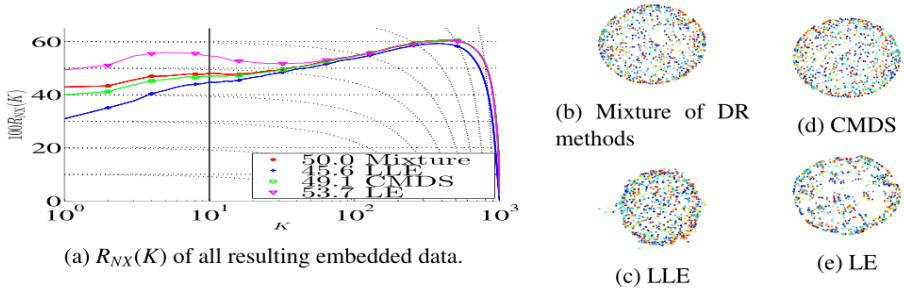
same methodology has to be applied, nevertheless three directions of change have to be founded. A triangle was chosen because its three vertexes can contain the primary colors that represent the three different DR methods. In Figure 4a three directions of change can be seen to build the RGB space. However, the sum of the intensity of the three channels (red,green,blue) must be normalized, this means that it must be equal to 1 for any point in the triangle.




(a) Three channels of the RGB space.


(b) Chromatic model.

Figure 4. In (a) the three Chanel of RGB space are represented and the final model in (b) is the superposition of the three channels




(a) $R_{NX}(K)$ of all resulting embedded data.

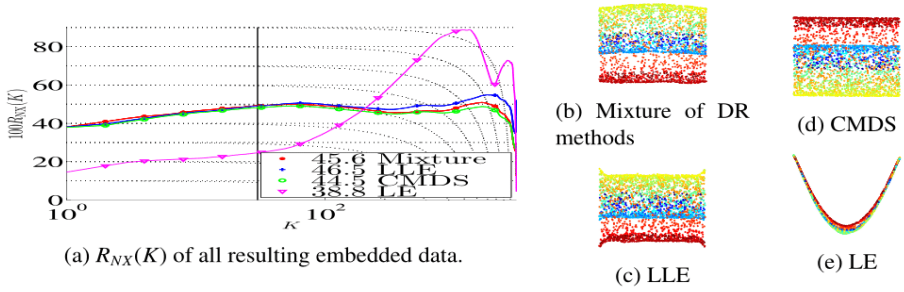
(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 6. Results for the 3D sphere data-set.




(a) $R_{NX}(K)$ of all resulting embedded data.

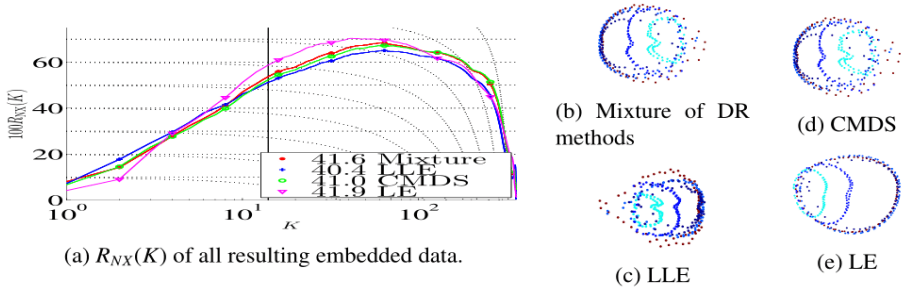
(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 7. Results for the Swiss roll data-set.




(a) $R_{NX}(K)$ of all resulting embedded data.

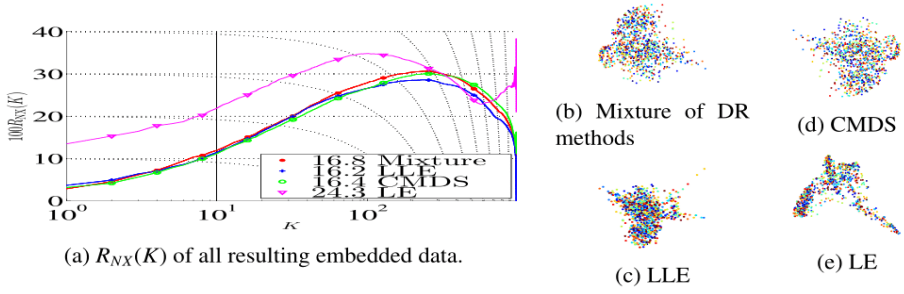
(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 8. Results for COIL data-set.




(a) $R_{NX}(K)$ of all resulting embedded data.

(b) Mixture of DR methods

(c) LLE

(d) CMDS

(e) LE

Figure 9. Results for MNIST dataset.

### 4.4. Experiment description

To assess the performance of the interactive visualization interface, a testings were done by clicking on the colored surface. Doing so a collection of weighting factors are established to consequently carry out the mixture. Here, particularly test the vertexes and a random point inside the surface Figure 10.

Figure 10. The chosen points for the experiment.

## 5. Results and discussion

The general interface's scheme is shown in Figure 11. The RGB color selected within the color surface defines the mixture of DR methods and the embedded data.

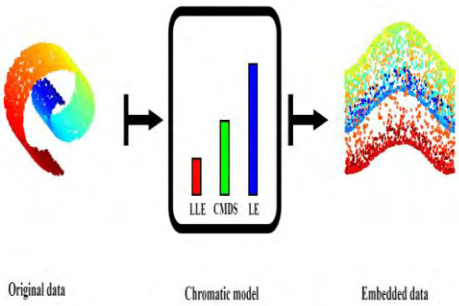



Figure 11. A general scheme of the proposed interface. This is an example for Swiss roll dataset.where the RGB value of the pixel define a new embedding space (right hand).

In section 4 the experiment is explained, three DR methods are considered and they are represented by the chromatic model. The experiment of Figure 10 is carried out to test the DR methods and the mixture, that it was taken approximately in a central point into the chromatic model. The results are shown on Figures 6 to 9. In the results can be appreciated the embedded data and several curves that gives a notion to the user about the performance of the low dimensional space and the preservation of neighbors. If the value of the area under the curve is greater, the performance of the embedded data will be better. An interesting fact can be seen in Figure 6a where the mixture have an area under the curve greater than some concrete DR methods.

## 6. Conclusions and future Work

The proposed chromatic model represents a suitable alternative to reduce the gap between users and data base because DR methods can be selected/mixed through a color-based framework. This approach results appealing since color is one of the first levels of human perception making naturally intuitive its use. Incorporating the chromatic model within an interface, the user can easily explore all the color surface to find the best representation of the input data into a lower-dimensional space. To do so, unsupervised DR methods are approximated by kernels matrices. Consequently, such matrices are linearly combined by means of weighted sum, whose coefficients are provided interactively by the user.

As a future work, more developed and interactive models are to be explored. As well, new kernel approaches from other dimensionality reduction methods that allow the arising of new DR approaches.

## Acknowledgments

This work was supported by the Department of Electronics from Universidad de Nariño and Facultad de Ingeniera en Ciencias Aplicadas from Universidad Tcnica del Norte. Authors acknowledgment the research projects "Diseño e implementación de un prototipo electrónico de bajo costo para terapias de biofeedback en tratamientos de trastornos psicofisiológicos" funded by Fundación CEIBA and Gobernación de Nariño, as well as "Combinación interactiva de mtodos reducción de dimensión, a partir de una interfaz de visualización inteligente" from Corporación Universitaria Autónoma de Nariño.

## References

[1] A. Asuncion and D. Newman. Uci machine learning repository. irvine, ca: University of california, school of information and computer science. *Available online a t http://www. ics. uci. edu/ mlearn/MLRepository. html*, 2007.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] I. Borg. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[4] J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton. Visualizing similarity data with a mixture of maps. In *International Conference on Artificial Intelligence and Statistics*, pages 67–74, 2007.

[5] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital image processing using MATLAB*. Pearson/Prentice Hall,, 2004.

[6] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47. ACM, 2004.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[8] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 2013.

[9] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). *Dept. Comput. Sci., Columbia Univ., New York.[Online] http://www. cs. columbia. edu/CAVE/coil-20. html*, 62, 1996.

[10] D. H. Peluffo-Ordonez, J. Aldo Lee, and M. Verleysen. Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, pages 171–177. IEEE, 2014.

[11] D. H. Peluffo-Ordóñez, A. E. Castro-Ospina, J. C. Alvarado-Pérez, and E. J. Revelo-Fuelagán. Multiple kernel learning for spectral dimensionality reduction. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 626–634. Springer, 2015.

[12] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen. Recent methods for dimensionality reduction: A brief comparative analysis. In *European Symposium on Artificial Neural Networks (ESANN)*. Citeseer, 2014.

[13] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen. Short review of dimensionality reduction methods based on stochastic neighbour embedding. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 65–74. Springer, 2014.

[14] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[15] J. Salazar-Castro, Y. Rosas-Narvaez, A. Pantoja, J. C. Alvarado-Perez, and D. H. Peluffo-Ordonez. Interactive interface for efficient data visualization via a geometric approach. In *Signal Processing, Images and Computer Vision (STSIVA), 2015 20th Symposium on*, pages 1–6. IEEE, 2015.

[16] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. In *Computer Graphics Forum*, volume 34, pages 201–210. Wiley Online Library, 2015.

[17] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner. Dimensionality reduction in the wild: Gaps and guidance. *Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03*, 2012.

[18] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2634–2643, 2013.

## ANEXO 6. ARTICULO: *IBEROAMERICAN CONGRESS ON PATTERN RECOGNITION* (CIARP)

# Interactive data visualization using dimensionality reduction and similarity-based representations

P. Rosero-Montalvo[1,2], P. Diaz[2,3], J. A. Salazar-Castro[4,5], D.F. Peña-Unigarro[5], A. J. Anaya-Isaza[6,7], J. C. Alvarado-Pérez[8,9], R. Therón[8], and D. H. Peluffo-Ordóñez[1]

[1] Universidad Técnica del Norte - Ecuador,
[2]Universidad de las Fuerzas Armadas ESPE - Ecuador,
[3] Universidad Nacional de la Plata - Argentina,
[4] Universidad Nacional sede Manizales - Colombia,
[5] Universidad de Nariño - Colombia,
[6] Universidad Surcolombiana - Colombia,
[7] Universidad Tecnológica de Pereira - Colombia,
[8] Universidad de Salamanca - Spain,
[9] Coorporación Universitaria Autónoma de Nariño, Pasto - Colombia,

**Abstract.** This work presents a new interactive data visualization approach based on mixture of the outcomes of dimensionality reduction (DR) methods. Such a mixture is a weighted sum, whose weighting factors are defined by the user through a visual and intuitive interface. Additionally, the low-dimensional representation space produced by DR methods are graphically depicted using scatter plots powered via an interactive data-driven visualization. To do so, pairwise similarities are calculated and employed to define the graph to be drawn on the scatter plot. Our visualization approach enables the user to interactively combine DR methods while provided information about the structure of original data, making then the selection of a DR scheme more intuitive.

**Keywords:** Data visualization, dimensionality reduction, pairwise similarity.

## 1 Introduction

The aim of dimensionality reduction (DR) is to obtain lower dimensional representations of high-dimensional input data keeping -under a pre-established criterion- the structure of data as well as possible. Reaching this aim, entails both the performance of a pattern recognition system and intelligible data representation can be improved [1]. Traditionally, DR methods are designed by following pre-established optimization criteria and design parameters. But they mostly lack of properties like interactivity and controllability, being important characteristics of the field of Information Visualization (InfoVis) [2]. InfoVis provides interfaces and graphical ways of representing data making the available information more usable and intelligible for the user. However, it turns out that DR outcomes can be enhanced by taking advantages of some properties of InfoVis methods [3, 4].Following this premise, some approaches have proposed [5, 6],

making use of interactivity with equalizer-bar like interfaces or geometric interaction models. In general, such approaches implement interesting interactive models but their final visualization lacks the information about structure of the data from the original input space -at least in an easy to understand and/or visual way-.

In this work, we introduce a new visualization approach using an interactive mixture of data representations resultant from DR methods. After performing the DR methods on the input data, a set of lower-dimensional representation spaces are obtained. Particularly, the mixture is done via a weighted sum. In order to give users a sense of the structure of data, we implement a data-driven visualization in addition to the conventional scatter plot. Such a visualization captures the structure of the input data by using a similarity matrix (as well, affinity matrix from graph theory), which captures the degree of similarity or affinity between every pair of data points. The visualization consists of plotting lines (edges) between data points exhibiting the highest value of similarity. Additionally, to provide more sense of interactivity, user can control the number of edges by a varying parameter -working as a slider bar within an interface-. By design, affinity is selected as a Gaussian one so that the structure of local neighbor points can be taken into account. Particularly, low-dimensional spaces are obtained by the state of the art of methods such as: Classical Multidimensional Scaling (CMDS) [2], Laplacian Eigenmaps (LE) [7], Locally Linear Embedding (LLE) [8], Stochastic Neighbor Embedding (SNE), and t-Student-distributed-SNE (t-SNE) [1, 7]. To perform the mixture, user can set the weighting factors by picking up values from a equalizer-bar-like interface. To test our visualization approach, we use a 3D artificial spherical shell data set. The quality of resultant representation spaces is quantified by a scaled version of the average agreement rate between K-ary neighborhoods [9]. The proposed mixture may represent every single dimensionality reduction approach as well as it helps users to find a suitable representation of input data within a visual and friendly user interface.

The remaining of the paper is organized as follows: In section 2, Data visualization via dimensionality reduction is outlined. Section 3 introduces the proposed interactive data visualization scheme. Experimental setup and results are presented in Sections 4 and 5, respectively. Finally, Section 6 gathers some final remarks as conclusions and future work.

## 2 Data visualization via dimensionality reduction

Perhaps, one of the most intuitive ways of visualizing numerical data is through a 2- or 3-dimensional representation of original data, which can be readily represented using a scatter plot. In consequence, dimensionality reduction arises as an Correspondingly, DR is aiming at reaching a low-dimensional data representation, upon which both the classification task performance is improved in terms of accuracy, as well as the intrinsic nature of data is properly represented [10]. So, when performing a DR method, a more realistic and intelligible visualization for the user is expected [1]. More technically, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq 1\leq N}$ such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq 1\leq N}$ being $\boldsymbol{y}_i \in \mathbb{R}^d$, where $d < D$ [1, 11]. Fig. 1 depicts an instance where

a manifold, so-called 3D spherical shield, is embedded into a 2D representation, which resembles to an unfolded version of the original manifold.

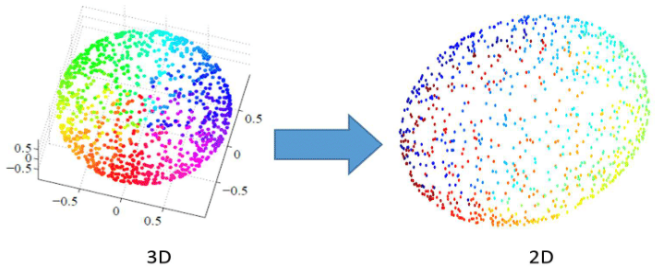



**Fig. 1.** Dimensionality reduction effect over an artificial (3-dimensional) spherical shell manifold. Resultant embedded (2-dimensional) data is an attempt to unfolding the original data.

## 3 Interactive data visualization scheme

The proposed visualization approach, here called DataVisSim, involves three main stages: mixture of DR outcomes, interaction, and visualization, as depicted in the block diagram of Fig. 2. One of the most important contributions of this work is that information on the structure of the input high-dimensional space is added to the visual final representation, by using a pairwise-similarity-based scheme.

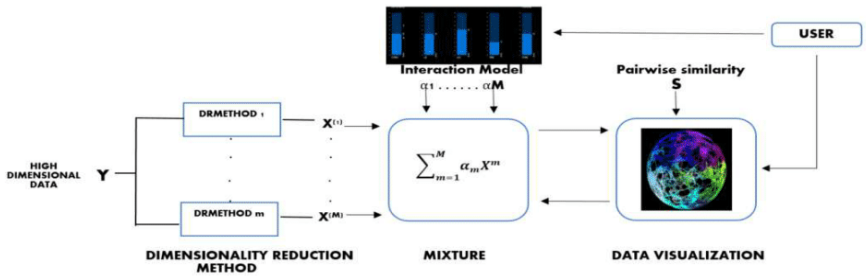



**Fig. 2.** Block diagram of proposed interactive data visualization using dimensionality reduction and similirarity-based representations (DataVisSim). Roughly speaking, it works as follows: first performs a mixture of resultant lower-dimensional representation spaces by taking advantage of conventional implementations of traditional DR methods. The interaction is provided through a interface that enables user to dynamically input the weighting factors for the aforementioned mixture. For visualization, a novel similarity-based approach is used.

### 3.1 Mixture

Let us suppose that the input matrix $\boldsymbol{Y}$ is reduced by using $M$ different DR methods, yielding then a set of lower-dimensional representations: $\{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(M)}\}$. Herein, we propose to perform a weighted sum in the form:

$$\bar{\boldsymbol{X}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{X}^m, \tag{1}$$

where $\{\alpha_1, \cdots, \alpha_M\}$ are the weighting factors. To make the selection of weighting factors intuitive, we use probability values so that $0 \leq \alpha_m \leq 1$ and $\sum_{m1=1}^{M} \alpha_m = 1$, and therefore all matrices $\boldsymbol{X}^{(m)}$ should be normalized to rely within a hypersphere of ratios.

### 3.2 Interaction model

For the sake of interactivity, the values of every $\alpha_m$ , required to calculate $\bar{\boldsymbol{X}}$ according to equation (1), are to be defined by the users using an equalizer-bar available in the interface. Within a friendly-user and intuitive environment, weighting factors can be readily inputted by just picking up values from bars. In order to provide quick views of resultant representation space, as soon as a point is picked up the remaining ones are automatically completed following a uniform density probability function. The same is done in case than more than one value is selected.

### 3.3 Similarity-based visualization

The most used method to visualize 2- or 3-dimensional data is the scatter plot. In this work, we introduce a similarity-based visualization approach with the aim to provide a visual hint about the structure of the high-dimensional input data matrix $\boldsymbol{Y}$ into the scatter plot of its representation in a lower-dimensional space To do so, we use a pairwise similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$, such that $\boldsymbol{S} = [\boldsymbol{s}_{ij}]$. In terms of graph theory, entries $\boldsymbol{s}_{ij}$ defines the similarity or affinity between the i-th and j-th data point from $\boldsymbol{Y}$. Doing so, we can hold the structure of original input space in a topological fashion, specifically in terms of pairwise relationships. For visualization purposes, such a similarity is used to define graphically the relationship between data points by plotting edges. In order to control the amount of edges and make an appealing visual representations, the value of $\boldsymbol{s}_{ij}$ is constrained as $\boldsymbol{s}_{ij} > \boldsymbol{s}_{max}$, being $\boldsymbol{s}_{max}$ a maximum admissible similarity value to be given by the users as well. In other words, our visualization approach consists of building a graph with constrained affinity values.

## 4 Experimental setup

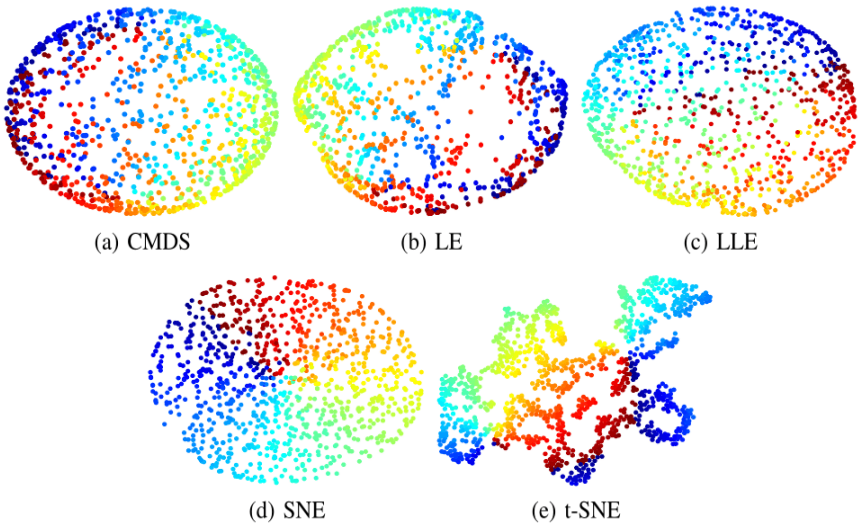
**Database:** In order to visually evaluate the performance of the DataVisSim approach, we use an artificial spherical shell (N = 1500 data points and D = 3), as depicted in Fig. 1.

**Parameter settings and methods:** In order to capture the local structure for visualization, i.e. data points being neighbors, we utilize the Gaussian similarity given by: $\boldsymbol{s}_{ij} = -exp(-0.5\|\boldsymbol{y}_{(i)} - \boldsymbol{y}_{(j)}\|^2/\sigma^2)$.The parameter is a bandwidth value set as 0.1, being the 10 % of the hypersphere ratio (applicable once matrices are normalized as discussed in Section 3.1. To perform the dimensionality reduction we consider $M = 5$ DR methods, namely: CMDS, LE, LLE, SNE, and t-SNE. All of them are intended to obtain spaces in dimension $d = 2$.

**Performance measure:** To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [9] is used, which is ranged within the interval [0, 1]. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N - 1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). The AUC assesses the dimension reduction quality at all scales, with the most appropriate weights.

## 5 Results and discussion

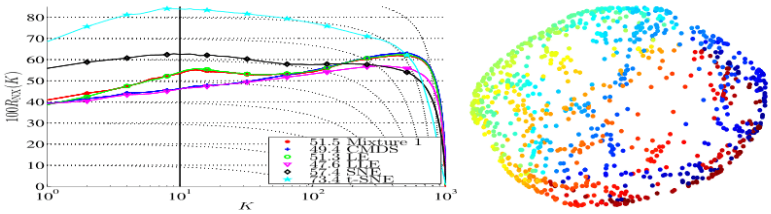
Figure 3 shows the scatter plots for the resultant low-dimensional spaces obtained by the considered dimensionality reduction methods, as well as the performed mixture. Quality curves and corresponding scatter of each mixture are shown in Fig. 4 to 8. As seen, $R_{NX}(K)$ measure allows for assessing both the different mixtures and the RD methods independently. Since the area under its curve represents a quality measure of the low-dimensional space, is in turn a visual and intuitive indicator that helps the user to find the best either a single DR method or the proper mixture.



(a) CMDS (b) LE (c) LLE

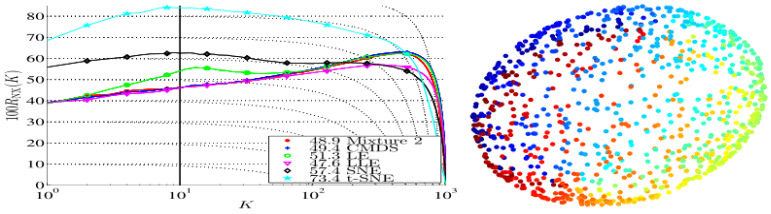
(d) SNE (e) t-SNE

**Fig. 3.** The effects of dimensionality reduction of RD methods considered on the 3d sphere. The results are embedded data represented in a bidimensional space.



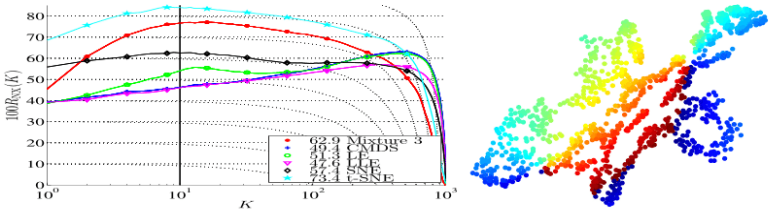

(a) $R_NX$(K) for all considered methods. (b) Mixture 1

**Fig. 4.** a) Performance of the mixture 1 and all methods deemed RD. In (b) the embedded data resulting from mixture 1 are indicated.



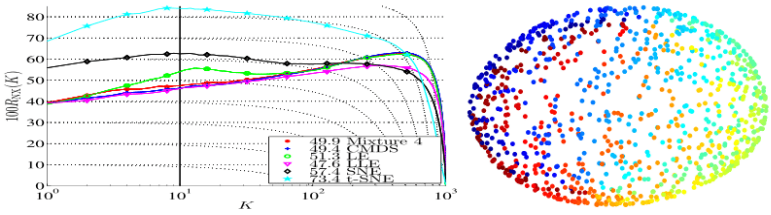

(a) $R_NX$(K) for all considered methods. (b) Mixture 2

**Fig. 5.** a) Performance of the mixture 2 and all methods deemed RD. In (b) the embedded data resulting from mixture 2 are indicated.




(a) $R_NX$(K) for all considered methods. (b) Mixture 3

**Fig. 6.** a) Performance of the mixture 3 and all methods deemed RD. In (b) the embedded data resulting from mixture 3 are indicated.




(a) $R_NX$(K) for all considered methods. (b) Mixture 4

**Fig. 7.** a) Performance of the mixture 4 and all methods deemed RD. In (b) the embedded data resulting from mixture 4 are indicated.

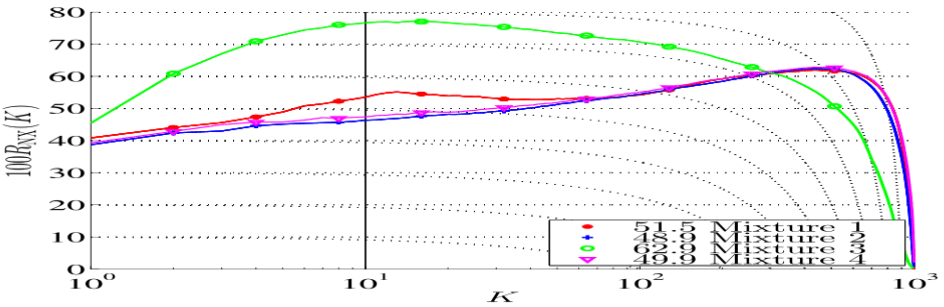



**Fig. 8.** Performance of all selected mixtures.

To test the DataVis approach, we implement an interface on Processing software, which allows to easily code visual arts. Then, it results appealing for creating visual analytics interfaces. Fig. 9 shows a view of the implemented interface. For the sake of easily handling so that (even non-expert) users may interact with DR methods and their feasible combinations in an intuitive manner using equalizer-like bars. This is possible because of resultant data representations are properly set according to the human perception. As well, the interface incorporates a slider bar to dynamically draw the edges between nodes. This is useful for visual analysis given that it allows to relate the structure of high-dimensional data (original data) within the visualization of the low-dimensional representation space. Therefore, it is provided a powerful tool for making decisions of the most suitable representation of the original data, in other words, the most proper DR methods.

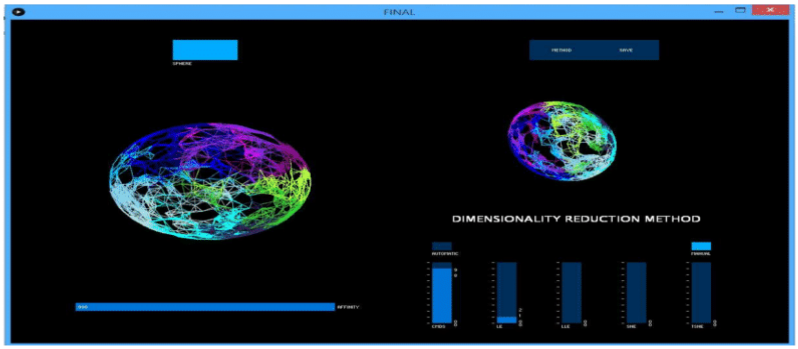



**Fig. 9.** View of the DataVisSim interface implemented on Processing software [1].

[1] https://sites.google.com/site/intelligentsystemsrg/home/gallery.

## 6 Conclusions and future work

This work presents a new interactive data visualization approach based on mixture of the outcomes of dimensionality reduction (DR) methods. The core of this approach consists of plotting lines (edges) between data points exhibiting the highest value using a similarity matrix which measure the degree of similarity or affinity between every pair of data points capturing the structure of the input data. Such visualization of a topology can be represented by a data-driven graph in addition to the conventional scatter plot, to provide more sense of interactivity to the user for selecting and/or combining DR methods while providing information about the structure of original data. Correspondingly, data points represent the nodes and an affinity matrix holds the pairwise edge weights As a future work, other dimensionality reduction methods are to be integrated into data-driven graph, so that a good trade between preservation of data structure and intelligible data visualization can be reached. More mathematical properties will be explored to design data-driven schemes that best approximate the topology data.

## References

1. Peluffo-Ordónez, D.H., Lee, J.A., Verleysen, M.: Short review of dimensionality reduction methods based on stochastic neighbour embedding. In: Advances in Self-Organizing Maps and Learning Vector Quantization. Springer (2014) 65–74
2. Borg, I., Groenen, P.J.: Modern multidimensional scaling: Theory and applications. Springer Science & Business Media (2005)
3. Dai, W., Hu, P.: Research on personalized behaviors recommendation system based on cloud computing. Indonesian Journal of Electrical Engineering and Computer Science **12** (2013) 1480–1486
4. Ward, M.O., Grinstein, G., Keim, D.: Interactive data visualization: foundations, techniques, and applications. CRC Press (2010)
5. Peluffo-Ordónez, D.H., Alvarado-Pérez, J.C., Lee, J.A., Verleysen, M., et al.: Geometrical homotopy for data visualization. In: European Symposium on Artificial Neural Networks (ESANN 2015). Computational Intelligence and Machine Learning. (2015)
6. Díaz, I., Cuadrado, A.A., Pérez, D., García, F.J., Verleysen, M.: Interactive dimensionality reduction for visual analytics. In: Proceedings of the 22th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014), Citeseer (2014) 183–188
7. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation **15** (2003) 1373–1396
8. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290** (2000) 2323–2326
9. Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing **112** (2013) 92–108
10. Bertini, E., Lalanne, D.: Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In: Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration, ACM (2009) 12–20
11. Peluffo-Ordóñez, D.H., Lee, J.A., Verleysen, M.: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, IEEE (2014) 171–177

## ANEXO 7. ARTICULO: *LATIN AMERICAN CONFERENCE ON COMPUTATIONAL INTELLIGENCE* (LA-CCI)

# Dimensionality reduction for interactive data visualization via a Geo-Desic approach


Jose A. Salazar-Castro,
Diego Peña-Unigarro

Universidad de Nariño
Pasto, Colombia
Universidad Nacional de Colombia Sede Manizales
Manizales, Colombia
alejo26st@udenar.edu.co,
diferpun@udenar.edu.co

Diego H. Peluffo-Ordóñez,
Paul D. Rosero-Montalvo,
H. Mauricio Domínguez-Limaico

Universidad Técnica del Norte
Ibarra, Ecuador
Universidad de Nariño
Pasto, Colombia
dhpeluffo@utn.edu.ec,
pdrosero@utn.edu.ec,
hmdominguez@utn.edu.ec

Juan C. Alvarado-Pérez,
Roberto Therón

Corporación Universitaria Autónoma de Nariño
Pasto, Nariño
Universidad de Salamanca
Salamanca, España
jcalvarado@usal.es,
theron@usal.es



***Abstract*—This work presents a dimensionality reduction (DR) framework that enables users to perform either the selection or mixture of DR methods by means of an interactive model, here named Geo-Desic approach. Such a model consists of linear combination of kernel-based representations of DR methods, wherein the corresponding coefficients are related to coordinated latitude and longitude inside of the world map. By incorporating the Geo-Desic approach within an interface, the combination may be made easily and intuitively by users –even non-expert ones– fulfilling their criteria and needs, by just picking up points from the map. Experimental results demonstrates the usability and ability of DR methods representation of proposed approach.**



***Keywords—dimensionality reduction, data visualization, data information, controllability, interaction, intelligible data, interface.***


## I. Introduction

Dimensionality reduction is possible because the Big Data are often simpler than their dimensionality implies, because they contain redundant information, such as correlations or variables with noisy information [1]. Therefore, the number of independent variables which can be described satisfactorily (intrinsic dimension) is less than its nominal dimension. The concept of dimensionality reduction can be approached from different viewpoints. In the field of pattern recognition it is often regarded as a feature extraction process that transforms input data to a more compact and manageable representation. In the visualization of complex data sets, DR is a crucial process that involves transforming data into visual representations that allow an analyst to more easily understand the process and recognize new patterns [2]. This strategy of data mining based on visual data exploration is called Visual Data Mining and Info Vis. Which it aims to develop graphical data representation forms so that information can be more useful and understandable to the user. The calculation of the intrinsic dimension is a central problem of DR, because their knowledge is useful for tuning training. In any case, this dimension may depend on the criteria considered from a priori information and design parameters and optimization preset criteria. However, in the visualization, the goal is not to reduce the data set to its intrinsic dimension, but a viewable (2 or 3), preserving as much information as possible.

This paper presents an attempt to link the field of dimensionality reduction with the information–visualization. DR can be improved by importing some properties of the multimedia visualization as the ability to user interaction [3] which may allow data analysts to parameterize the reduction according to their expectations and have more control over the process of dimension reduction which should make the DR outcomes significantly more understandable and tractable for the user (no-necessarily-expert) to have freedom for selecting the best way for representing data [4, 5]. Briefly put, DR methods are aiming at the extraction of embedded data (lower-dimensional output data) like relevant information from high-dimensional input data. To develop a graphical way to illustrate data with the purpose of getting a more useful and intelligible information to the user in the field of information-visualization (Info Vis). This is, in fact, the premise on which this research is based.

Specifically, we propose a geodesic strategy to set the weighting factors for linearly combining DR methods. This is done from kernel approximations [6, 7] of conventional methods (Classical Multidimensional Scaling - CMDS [3], Laplacian Eigen maps – LE, and Locally Linear Embedding - LLE), which are combined to reach a mixture of kernels. To involve the user in the selection of a method, we use a geographic Earth approach so the geographic coordinates selected inside a world map defines the degree or level that a kernel is used, that is, the set of weighting factors. Such map allows to have as many geographic coordinates as the number of considered kernels. This approach allows to evaluate visually the behavior of the embedding data regarding the kernel mixture.


978-1-5090-5105-2/16/$31.00 ©2016 IEEE

For experiments, we use publicly available databases from the UCI Machine Learning Repository [8] as well as a subset of images from Columbia University Image Library [9]. To assess the performance of the kernel mixture, we consider conventional methods of spectral dimensionality reduction such as CMDS, LE and LLE [10]. The quality of obtained embedded data is quantified by a scaled version of the average agreement rate between $K$-ary neighborhoods [11]. Provided mixture represents every single dimensionality reduction approach as well as it helps users to find a suitable representation of embedded data within a visual and intuitive framework.

An intuitive way of visualizing numerical data is via a 2D or 3D scatter plot, which is a natural and intelligible visualization fashion for human beings. Therefore, it entails that the initial data should be represented into a lower-dimensional space. In this sense, dimensionality reduction takes places, being an important stage within both the pattern recognition and data visualization systems. Correspondingly, DR is aiming at reaching a low-dimensional data representation, upon which both the classification task performance is improved in terms of accuracy, as well as the intrinsic nature of data is properly represented [1]. So, a more realistic and intelligible visualization for the user is obtained [2]. In other words, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq i\leq N}$, such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq i\leq N}$, being $\boldsymbol{x}_i \in \mathbb{R}^d$, where $d < D$. Figure 1 depicts an instance where a manifold, so-called Swiss roll, is embedded into a 2D representation, which resembles to an unfolded version of the original manifold.


Fig. 1. Dimensionality reduction effect over a swiss roll manifold. Resultant embedded data is an attempt to unfolding the original data.

Classical DR approaches aims to preserve variance (principal component analysis - PCA) or distance (classical multidimensional scaling - CMDS) [3]. Nowadays, more developed, recent methods are aiming at preserving the data topology. Such a topology can be represented by a data-driven graph, built as a non-directed and weighted one, in which data points represent the nodes, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. This representation is exploited by both spectral and divergence-based methods. On one hand, for spectral approaches, similarity matrix can represent the weighting factor for pairwise distances as happens in Laplacian Eigen maps (LE) [10]. As well, using a non-symmetric similarity matrix and focusing on data local structure, the Locally Linear Embedding (LLE) method arose [12]. On the other hand, once normalized, similarity matrix can also represent probability distributions, as do the methods based on divergences such as stochastic neighbor embedding [13]

The remaining of the paper is organized as follows: In section 2, a novel geographical method based on a Geo-Desic approach is introduced to aim at performing DR tasks. Section 3 and 4 show experimental setup and results, respectively. Finally, Section 6 rallies some final remarks as conclusions and future work.

## II. Interactive Reduction Using a Model

In this section, a new method for interactive data visualization is introduced. It consists on the linear combination of spectral unsupervised DR methods prompting the use of the corresponding kernel matrices of each method, following the consideration that the spectral unsupervised DR methods can be represented by kernels [14].

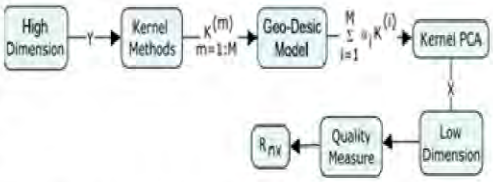



Fig. 2. Block diagram of proposed interactive data visualization using dimensionality reduction which illustrates each stage, step by step, about the interface.

Our method is based on an interactive interface that allows for performing the mixture of kernel matrix in an intuitive fashion. First, the kernel matrices obtained by applying the methods RD in high-dimensional data, are linearly combined applying Geo-Desic model that allows the mixture of these kernels to obtain a new matrix kernel which is introduced into a kernel PCA environment and thus obtain low dimension data. So, users –even non-expert ones – might easily and intuitively select a single method or combine methods fulfilling their needs by just de exploring the world map (*mapamundi*) and picking up points from the surface thereof. Figure 2 shows graphically a diagram of the interface.

### A. DR Methods and Proposed Model

Our method is represented by a geographic approach of the Earth which is called Geo-Desic model that allows us to select points inside of a World Map to develop the mixed activity in an interactive way. Figure 3 shows graphically a possible Geo-Desic model regarding a geographic Earth approach.

In general, any set of methods can be represented by a collection of functions $\{f_1,..,f_M\}$, where $M$ is the number of considered methods, for this work we consider $M = 4$. For data visualization purposes through DR methods, we obtain a resultant kernel matrix $\widehat{\boldsymbol{K}}$ as the mixture of $M$ kernel matrices $\{\boldsymbol{K}^{(1)},\ldots,\boldsymbol{K}^{(M)}\}$ corresponding to the considered DR methods which are the terms combined, so:

$$\widehat{\boldsymbol{K}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{K}^{(m)}, \qquad (1)$$

where $\alpha_m$ is the coefficient or weighting factor corresponding to method $m$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]$ is weighting vector, which is associated with geographic coordinates of points inside the world map.

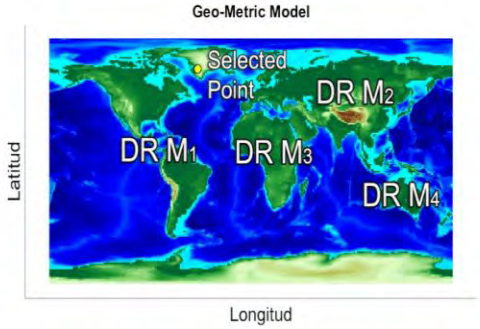



Fig. 3. Geographic-Earth approach to perform the mixture of a set of function (DR methods) $\{f_1,\ldots,f_M\}$ with $M = 4$.

### B. Interactive Coefficients

The corresponding $\alpha_m$ coefficients for each kernel are related to latitude and longitude coordinates inside of the world map and then, they become the weighting factors to get the linear combination of kernel matrices of DR methods. In virtue of the above, the combination might be made easily and intuitively by users –even non-expert ones– fulfilling their criteria and needs, this can be dealt by making an exploration of the map and picking up points from the surface.

In this work, the distance from the center geographic coordinates of every continent (America, Eurasia, Africa and Oceania represent considering methods) to the selected point turns to be the relationship between the coefficients of linear combination and the geographic coordinates, so estimate distances are explained graphically in figure 4. Then, these distances are transformed into the diameters $\{D_1, \ldots, D_M\}$ of spheres with volumes $\{V_1, \ldots, V_M\}$. In this sense, the volume of the $m$-sphere is $V_m = 4\pi r_m^3/3$ with $r_m = D_m/2$.

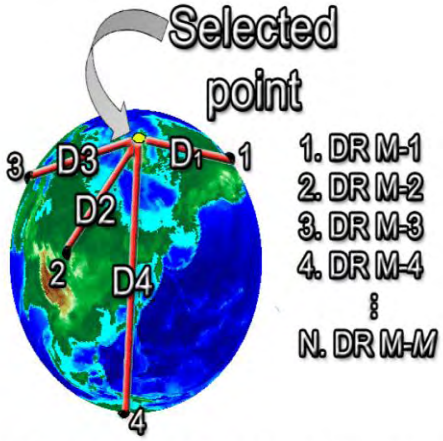
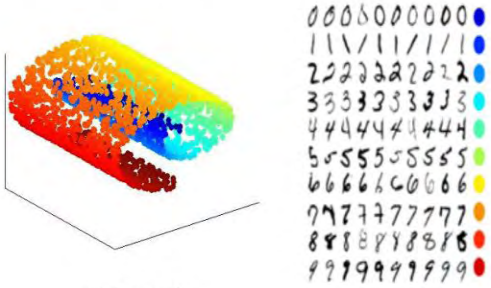


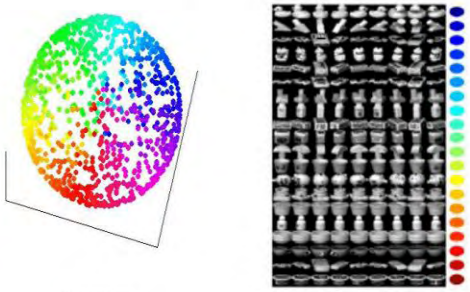

Fig. 4. Graphical explanation of the fashion to estimate weighting factors.

Then, the volumes are normalized to sum to 1 and the value of them becomes a proper estimation of the weighting factors. Additionally, it is evident that the coefficient assigned to the nearest distance is not significantly higher, so it prompting the use of an equalization effect over the weighting factors, this is made using the function $\mathrm{sinc}(\cdot)$. The amounts to say that the values of $\boldsymbol{\alpha}$ are given by:

$$\alpha_m = \mathrm{sinc}\left(1 - \frac{V_m}{\sum_{m=1}^{M} V_m}\right). \qquad (2)$$



(a) *Swiss roll* (b) *MNIST*



(c) *sphere* (d) *COIL-20*

Fig. 5. Each of the four considered datasets.

## III. Experimental Setup

Experiments are carried out over four conventional data sets. The first data set is a toy set here called Swiss roll (N = 3000 data points and D = 3) like shows the Figure 5. The second data set is a randomly selected subset of the MNIST image bank [15], which is formed by 6000 gray-level images of each of the 10 digits (N = 1500 data points –150 instances for all 10 digits– and D = 242). The third data set is an artificial spherical shell (N = 1500 data points and D = 3). The fourth data set is the COIL-20 image bank [13], which contains 72 gray-level images representing 20 different objects (N = 1440 data points 20 objects in 72 poses/angles with D = 1282).

Four kernel approximations for spectral DR methods [6] are considered.

• ***Classical Multidimensional Scaling (CMDS)***: CMDS kernel is the double centered distance matrix D ∈ $\mathbb{R}^{NxN}$ so:

$$\boldsymbol{K}^{(1)} = \boldsymbol{K}_{\text{CMDS}} = -\frac{1}{2}\left(\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N{}^T\right)\boldsymbol{D}\left(\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N{}^T\right), \quad (3)$$

where the $ij$ entry of $\boldsymbol{D}$ is given by $d_{ij} = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2$ and $\|\cdot\|^2$ stands for Euclidean norm.

• ***Locally Linear Embedding (LLE)***: LLE can be approximated from a quadratic form in terms of the matrix $\boldsymbol{W}$ holding linear coefficients that sum to 1 and optimally reconstruct observed data Define a matrix $\boldsymbol{M} \in \mathbb{R}^{NxN}$ as $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{W})(\boldsymbol{I}_N - \boldsymbol{W}^T)$ and $\lambda_{max}$ as the largest eigenvalue of M. Kernel matrix for LLE is in the form:

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LLE} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}. \quad (4)$$

• ***Graph Laplacian Eigenmaps (LE)***: Since kernel PCA is a maximization problem of the covariance of the high dimensional data represented by a kernel, LE can be expressed as the pseudo-inverse of the graph Laplacian L:

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LE} = \boldsymbol{L}^T, \quad (5)$$

where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, $\boldsymbol{S}$ is a similarity matrix and $\boldsymbol{D}$ = Diag $(\boldsymbol{S}\mathbf{1}_N)$ is the degree matrix. All previously mentioned kernels are widely described in [6]. The similarity matrix S is formed in such a way that the relative bandwidth parameter is estimated keeping the entropy over neighbor distribution as roughly $\log(K)$ where $K$ is the given number of neighbors as explained in [16]. The number of neighbors is established as $K = 30$.

• ***Radial Basis Function (RBF)***: RBF kernel is $\boldsymbol{K}^{(4)} = \boldsymbol{K}_{RBF}$ whose $ij$ entries are given by $exp(-0.5||yi - yj||/\sigma^2)$ with σ = 0.1. For all methods, input data is embedded into a 2-dimensional space ($d = 2$).

Accordingly, our approach is performed considering $M = 4$ kernels. The resultant kernel provided $\widehat{\boldsymbol{K}}$ here as well as the individual kernels $\{\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}\}$ are tested by obtaining embedded data from kernel PCA, as explained in [17]. To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [11] is used, which is ranged within the interval [0, 1]. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N - 1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). The AUC assesses the dimension reduction quality at all scales, with the most appropriate weights.

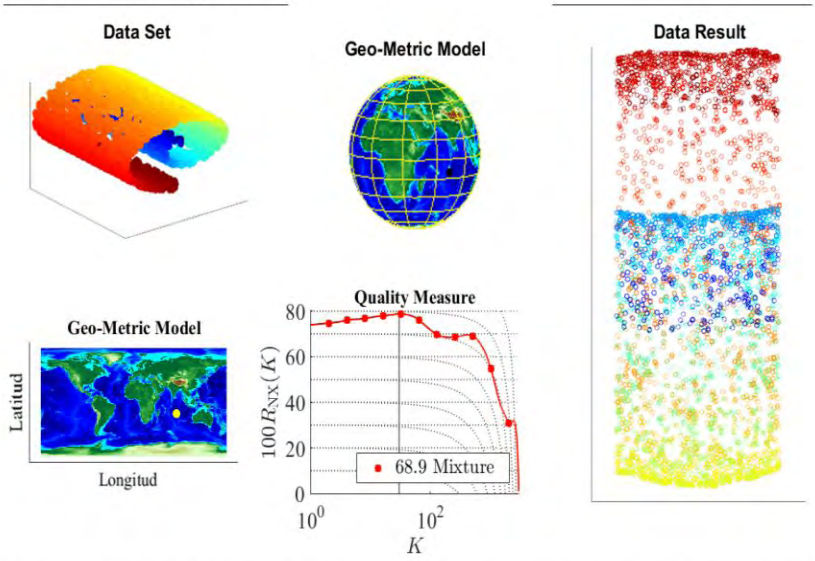



Fig. 6. A view of the proposed interface with an example for Swiss roll dataset. As the geographic coordinates are changed by clicking on a point inside the world map, a new embedding space (data result) is reached with its corresponding $R_{NX}(K)$ curve showed.

## IV. Results And Discussion

In this section we present the experimental results which it aim at testing all the considered datasets regarding the embedded data. A $R_{NX}(K)$ measure is used as a quality indicator to comparison of the different kinds of mixtures, when the point selected is exactly on a method (four possibilities) and when a mixture of kernel matrices is made. Figure 6 shows the interface proposed in which the user load a dataset and get a new result by clicking points inside the map, the result embedding data is shown in the right side and in the middle the $R_{NX}(K)$ curve.

Indeed, the selection of coefficients associated with the continents performs the effect of a single method. In addition, the selection of inner points take into account the effects of each method to calculate the resultant kernel. Therefore, the intuitively selection of coordinates from a world map enable users (even those not expert) to control and interact with the DR outcomes. Overall obtained results are shown in Figures 7 to 10.

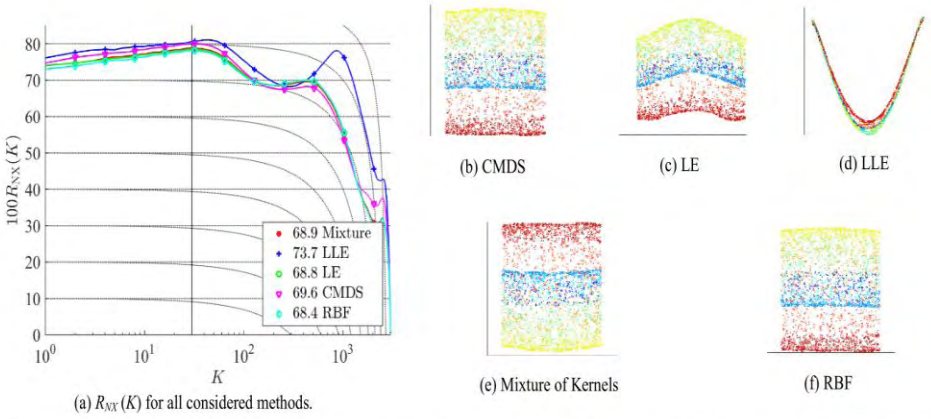



Fig. 7. Results for Swiss Roll dataset. Results are shown regarding the quality measure $R_{NX}(K)$. The curves and their AUC (a) for all considered methods are depicted, as well as the embedding data (b)-(f). Individual embedding data spaces are obtained by selecting the points rightly on each of the fourth continents, the mixture is done with the coefficients associated to a random point.

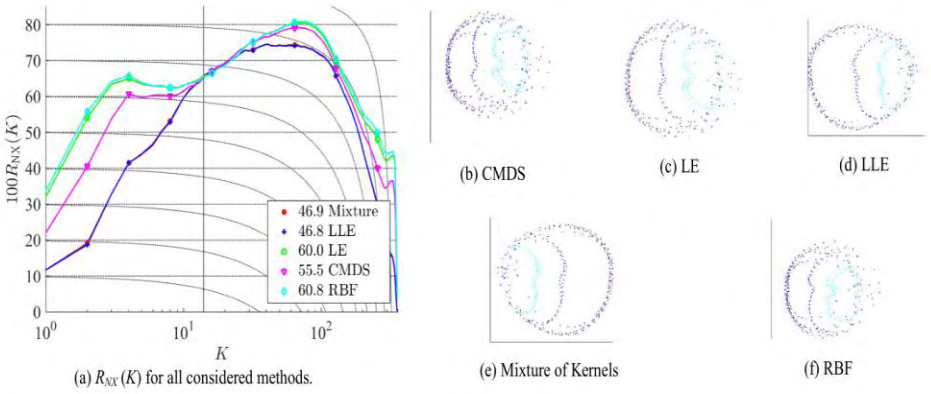



Fig. 8. Results for COIL dataset.

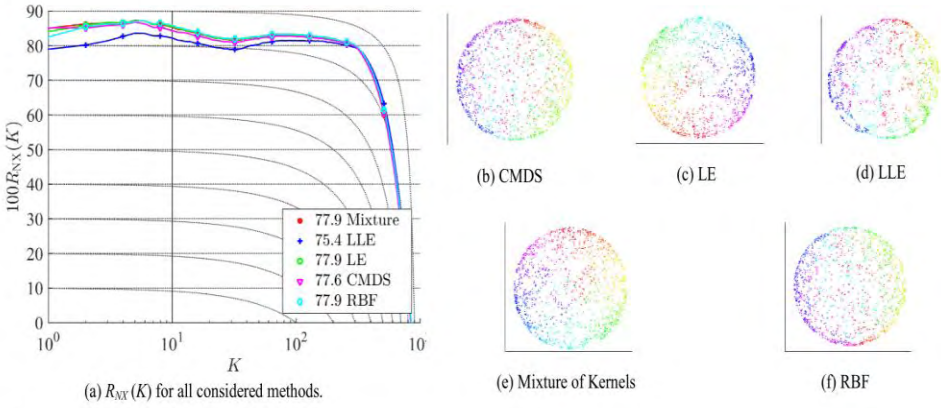
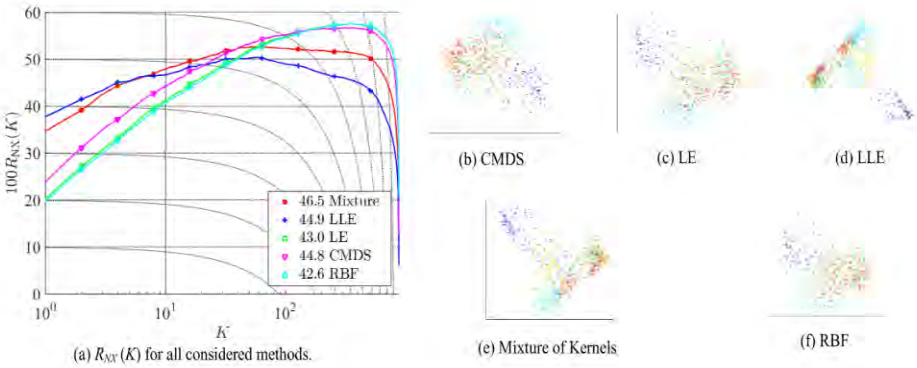



Fig. 9. Results for Sphere dataset.




(a) $R_{NX}(K)$ for all considered methods.

Fig. 10. Results for MNIST dataset.

## V. Conclusion And Future Work

This work presents an interactive interface for data visualization based on mixture of kernel matrices corresponding to the representation of unsupervised spectral DR methods. The geographical coordinates (latitude and longitude) related to the selected points inside of the map allow performing the mixture for the user –even non-expert- in an interactive and controllable way to visualize the embedded data. In this sense, the user might fulfill their specific needs and parameter criteria by picking up points from a world map.

As a future work, other DR methods will be included to mixture, aimed at improve the outcomes of dimensionality reduction. More developed and interactive models can be explored to optimize and speed up the interface and its performance.

## Acknowledgment

This work is supported by *Universidad Nacional de Colombia sede Manizales* as well as "*Grupo de investigación en Ingeniería Eléctrica y Electrónica – GIIEE*" from *Universidad de Nariño*.

Authors acknowledge the research project *"Diseño e implementación de un prototipo electrónico de bajo costo para terapias de biofeedback en tratamientos de trastornos psicofisiológico"* funded by Fundación CEIBA and Gobernación de Nariño, as well as project *"Combinación interactiva de métodos de reducción de dimensión a partir de una interfaz de visualización inteligente"* funded by Cooporación Universitaria Autónoma de Nariño.

## References

[1] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery" Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. ACM, 2009.

[2] D. H. Peluffo-Ordónez, J. A. Lee and M. Verleysen. "Short review of dimensionality reduction methods based on stochastic neighbor embedding". Advances in Self-Organizing Maps and Learning Vector Quantization. Springer International Publishing, 2014. 65-74.

[3] I. Borg and J. Patrick. Modern multidimensional scaling: Theory and applications. Springer Science & Business Media, 2005.

[4] W. Dai, and P. Hu. "Research on Personalized Behaviors Recommendation System Based on Cloud Computing." TELKOMNIKA Indonesian Journal of Electrical Engineering 12.2 (2013): 1480-1486.

[5] M. Ward, G. Grinstein, and D. Keim, Interactive data visualization: foundations, techniques, and applications. AK Peters, Ltd., 2010.

[6] J. Ham, D. D. Lee, S. Mika, and B. Scholkopf. "A kernel view of the dimensionality reduction of manifolds." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004, p-47.

[7] D. H. Peluffo-Ordónez, J. A. Lee, and M. Verleysen, "Generalized kernel framework for unsupervised spectral methods of dimensionality reduction," in Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, Dec 2014, pp. 171–177.

[8] M. Lichman, "UCI Machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml.

[9] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-20)," Dept. Comput. Sci., Columbia Univ., New York. [Online] http://www.cs.Columbia.edu/CAVE/coil-20.html, vol. 62, 1996.

[10] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", Neural computation, vol. 15, no. 6, pp. 1373–1396, 2003.

[11] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing. (2013).

[12] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, pp. 2323–2326. (2000).

[13] S. A. Nene, S. K. Nayar, H. Murase: Columbia object image library (coil-20). Dept. Compute. Sci., Columbia Univ., New York, 62 (1996), http://www.cs.columbia.edu/CAVE/coil-20.htm.l

[14] J. A. Salazar-Castro, Y. C. Rosas-Narváez, A. D. Pantoja, J. C. Alvarado-Pérez, and D. H. Peluffo-Ordóñez, (2015, September). Interactive interface for efficient data visualization via a geometric approach. In Signal Processing, Images and Computer Vision (STSIVA), 2015 20th Symposium on (pp. 1-6). IEEE.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11) (1998) 22782324.

[16] J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton: Visualizing similarity data with a mixture of maps. In: International Conference on Artificial Intelligence and Statistics. (2007) 67–74

[17] D. Peluffo-Ordóñez, J. Lee, and M. Verleysen: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: IEEE Symposium Series on Computational Intelligence. (2014).

## ANEXO 8. PÁGINA WEB

Dentro del desarrollo de este proyecto, se contempla la creación de una página web en Google Sites, donde, se puede encontrar información general acerca de la interfaz desarrollada, así como, el código fuente y el archivo ejecutable, para probar la aplicación en cualquier equipo sin necesidad de Processing. Un manual de usuario y un video tutorial que explica el funcionamiento de la interfaz gráfica son incluidos con el fin de dar un mejor entendimiento acerca del proyecto y fomentar la divulgación de los resultados obtenidos.

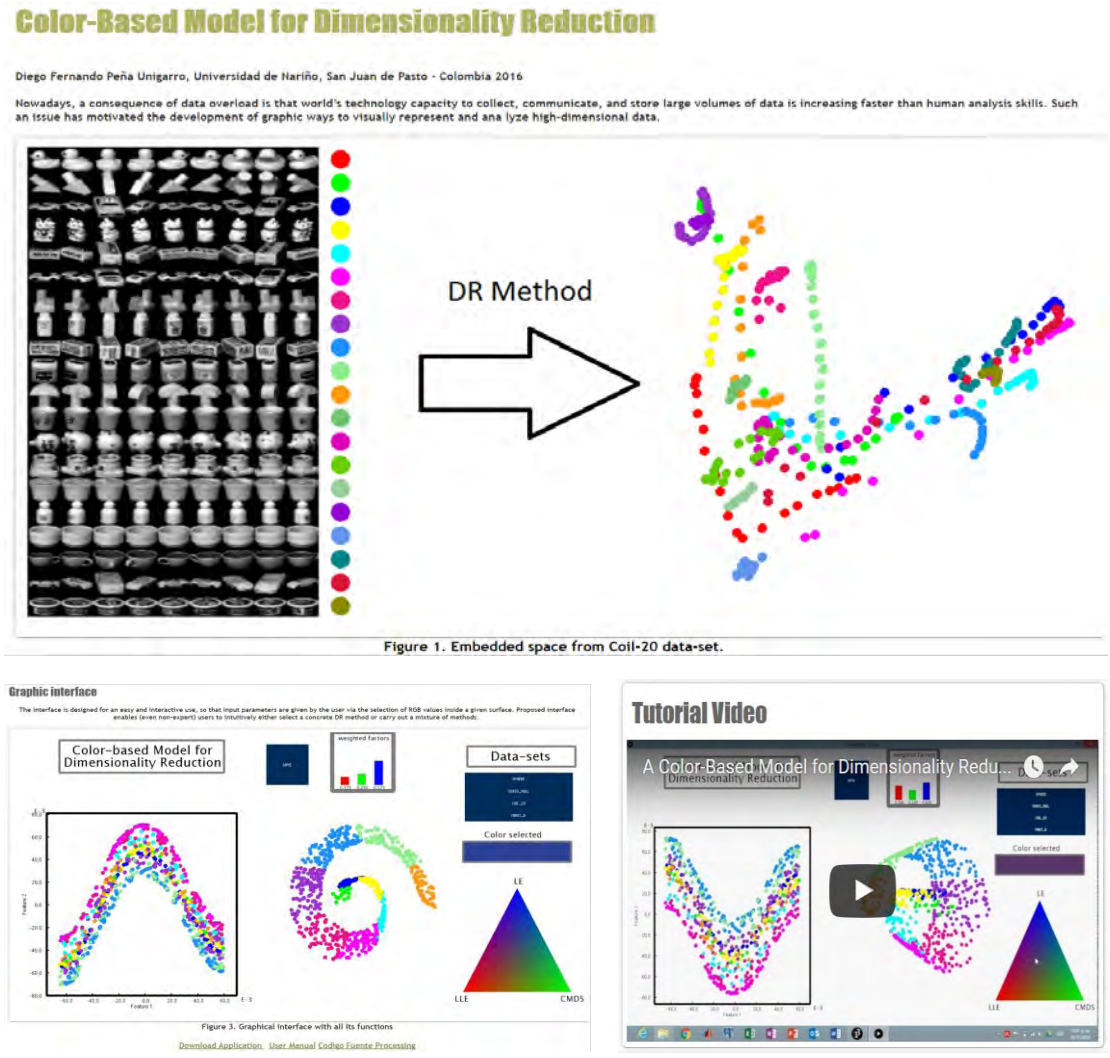



**Figura 40.** Diseño de la página web[4] en donde se puede encontrar una amplia información acerca de la interfaz desarrollada así como el código fuente, tutoriales y manuales. **Fuente:** Esta investigación.

[4] https://sites.google.com/site/degreethesisdiegopeluffo/color_based_model