METODOLOGÍA DE SELECCIÓN DE COMPONENTES PRINCIPALES COMUNES PARA REPRESENTACIÓN Y EXTRACCIÓN DE LAS DERIVAS PRESENTES EN SENSORES DE GAS

EVA SUSANA ALBARRACÍN ESTRADA

Institución Universitaria

INSTITUTO TECNOLÓGICO METROPOLITANO – ITM
MAESTRÍA EN AUTOMATIZACIÓN Y CONTROL INDUSTRIAL
FACULTAD DE INGENIERÍAS
MEDELLÍN, COLOMBIA
MAYO, 2014

METODOLOGÍA DE SELECCIÓN DE COMPONENTES PRINCIPALES COMUNES PARA REPRESENTACIÓN Y EXTRACCIÓN DE LAS DERIVAS PRESENTES EN SENSORES DE GAS

EVA SUSANA ALBARRACÍN ESTRADA

Trabajo de investigación para optar al título de
Magíster en Automatización y Control Industrial

Directores
PhD JORGE ALBERTO JARAMILLO GARZÓN
PhD EDILSON DELGADO TREJOS
M.Sc ANDREY ZIYATDINOV

Línea de investigación
Máquinas Inteligentes y Reconocimiento de Patrones

INSTITUTO TECNOLÓGICO METROPOLITANO – ITM
MAESTRÍA EN AUTOMATIZACIÓN Y CONTROL INDUSTRIAL
FACULTAD DE INGENIERÍAS
MEDELLÍN, COLOMBIA
MAYO, 2014

APROACH SELECTION OF COMMON PRINCIPAL COMPONENTS FOR REPRESENTATION AND EXTRACTION OF DRIFTS EXISTING IN GAS SENSORS

EVA SUSANA ALBARRACÍN ESTRADA

A thesis submitted to the postgraduate program "Masters in Industrial Control and Automation" in partial fulfillment of the requirements for the Master's degree

Supervisors
PhD JORGE ALBERTO JARAMILLO GARZÓN
PhD EDILSON DELGADO TREJOS
M.Sc ANDREY ZIYATDINOV

Research Line
Machine Learning and Pattern Recognition

INSTITUTO TECNOLÓGICO METROPOLITANO
MASTERS IN INDUSTRIAL CONTROL AND AUTOMATION
FACULTY OF ENGINEERING
MEDELLÍN, COLOMBIA
MAYO, 2014

Dedico este gran logro a los pilares de mi existencia:
mi esposo, mis padres y mis hermanos.

AGRADECIMIENTOS

A Jorge Alberto Jaramillo Garzón, profesor del Instituto Tecnológico Metropolitano**,** por su valioso aporte académico como director de esta tesis de maestría y su contribución constante en el desarrollo y finalización de este trabajo.

A Edilson Delgado Trejos, Decano Facultad de Ingenierías del Instituto Tecnológico Metropolitano, quien con su gran experiencia ha orientado esta investigación y además ha sido gestor y partícipe en todo el proceso llevado a cabo para la finalización de este trabajo de investigación.

A Andrey Ziyatdinov, Docente Becario de la Universitat Politècnica de Catalunya – Barcelona (España), experto en el tratamiento de derivas de sensores químicos y gran colaborador en esta investigación.

Al Instituto Tecnológico Metropolitano - (ITM) por ofrecer todas las garantías, espacios y equipos del laboratorio de Máquinas Inteligentes y Reconocimiento de Patrones para la materialización del proyecto que se describe en este libro.

A la Universitat Politècnica de Catalunya - Barcelona (España) y muy especialmente a Alexandre Perera i Lluna PhD, Investigador del Centre de Recerca en Enginyeria Biomedica (CREB), por haber permitido el desarrollo de la pasantía de investigación que se llevó a cabo en la UPC, enmarcada en este trabajo de investigación.

A University of California, San Diego (EE.UU), por permitir el libre acceso a la base de datos titulada: Gas Sensor Array Drift Dataset, usada como principal insumo en la ejecución de este proyecto.

A los docentes de Maestría en Automatización y Control del ITM, por la experiencia y conocimientos transmitidos, así como por su compromiso en la enseñanza de las diferentes asignaturas propias del pensum de este Posgrado.

TABLA DE CONTENIDO

Pág.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

GLOSARIO

- **Entrenamiento:** Procedimiento matemático relacionando con un algoritmo supervisado de reconocimiento de patrones, a través del cual se pretende encontrar un modelo o los parámetros más adecuados para un modelo a partir de un conjunto de datos.
- **Deriva:** es un proceso dinámico causado por cambios químicos en los sensores más específicamente en la capa activa de los mismos.
- **Escalado: A**daptación de los datos a una escala o con respecto a un punto de referencia.
- **Espécimen:** En biología un espécimen es un individuo o parte de un individuo que se toma como muestra, en este trabajo un espécimen es una correspondiente medida de algún compuesto o elemento, tomada con algún sistema de olfato electrónico.
- **Matriz:** Este término se utiliza mucho en este documento cuando se hace referencia a un arreglo de sensores de gases, esto se debe a que los datos obtenidos de los mismos por lo general son almacenados en matrices.
- **Modelo:** En este documento hace referencia a los modelos matemáticos, los cuales buscan representar fenómenos o relaciones entre ellos a través de una formulación matemática.
- **Pre-procesamiento:** Es un proceso que consiste en la manipulación de los datos con el propósito de extraer información poco útil o impura que pueda interferir posteriormente; en algunos casos sirve para ajustar los datos a unos rangos o parámetros que faciliten su manipulación.
- **Sensor:** Dispositivo capaz de detectar magnitudes físicas o químicas y transformarlas en magnitudes eléctricas, también puede ser visto como un dispositivo que convierte una forma de energía en otra.
- **Sintonización:** En este documento se utiliza el término para referirse al ajuste de parámetros de la máquina de vectores de soporte para regresión.
- **Sistema de reconocimiento de olores:** Un dispositivo que incluye el hardware y software necesario para emular el funcionamiento del sentido del olfato, capaz de detectar olores con mucha precisión y exactitud.
- **Validación:** Procedimiento estadístico que permite estimar la capacidad de generalización del modelo ante datos que no estén presentes en el conjunto de entrenamiento.
- **Volátil:** Sustancia que se transforma fácilmente en vapor o gas cuando está expuesta al aire.

ACRÓNIMOS

- **PCA**: Análisis de Componentes Principales.
- **CPCA:** Análisis de Componentes Principales Comunes.
- **CC:** Corrección de Componentes.
- **CC-PCA**: Corrección de Componentes por Análisis de Componentes Principales.
- **CC-CPCA**: Corrección de Componentes por Análisis de Componentes Principales Comunes.
- **k-NN:** Método de clasificación de los *k* vecinos mas próximos.

RESUMEN

Los sistemas de detección y clasificación de olores a menudo se ven afectados por la presencia de derivas, esto ocasiona que los modelos utilizados en los algoritmos para el reconocimiento de patrones tengan cortos periodos de utilidad y por lo tanto tienen la necesidad de una recalibración constante (Arthurson, Eklöv, Lundström, Marterson, Sjöstrom, & Holmberg, 2000). Este fenómeno, además de hacer obsoletos los modelos construidos, degradan la estabilidad del dispositivo en el proceso de reconocer y cuantificar los compuestos volátiles (Ziyatdinov, y otros, 2009).

Este trabajo presenta una metodología novedosa para enfrentar el problema de las derivas existentes en sensores químicos empleados en sistemas de olfato artificial, por medio de la cual se logra mitigar el efecto causado por las mismas al reducir los errores en la clasificación de diferentes compuestos volátiles. Se aplicó la técnica de análisis estadístico multivariado, denominada Análisis de Componentes Principales Comunes (CPCA) combinada con la técnica de corrección de componentes (CC) planteada por (Arthurson, y otros, 2000) y se determinó un criterio de selección del número de componentes principales comunes a ser substraídas de las medidas para mejorar la exactitud en el proceso de detección de olores.

La metodología propuesta en este trabajo tuvo como punto de partida la investigación realizada en (Ziyatdinov, y otros, 2009), donde los autores emplean sólo la primera componente principal común para hacer la corrección de las derivas con el propósito de asumir ésta corrección como lineal. Las componentes de deriva presentes en sistemas de olfato artificial son realmente no lineales, por tanto en este trabajo se incorporó como novedad el remover un mayor número de componentes para mitigar su efecto y a su vez considerando el no capturar información relevante para el sistema de clasificación.

Los resultados de substraer mayor cantidad de componentes principales comunes en la corrección de las derivas demostraron que el remover más de una componente principal común mediante la corrección de componentes, ocasiona incrementos en los porcentajes de acierto en el proceso de clasificación sobre los datos de validación. Se determina el número adecuado de componentes que se deben remover a partir de un indicador de la separabilidad de los conjuntos de datos, calculado a partir del conjunto de datos usado para el entrenamiento.

ABSTRACT

The detection and classification of odors is often affected by the presence of drifts, which causes the models used in the algorithms for pattern recognition have shorter lifetimes and therefore must be frequently recalibrated (Arthurson, Eklöv, Lundström, Marterson, Sjöstrom, & Holmberg, 2000). This phenomenon, in addition to converting in obsolete the built models, makes the stability of the system be degraded in the process to recognize, quantify and identify the volatile compounds (Ziyatdinov, and others, 2009).

This document presents a novel approach to confront the problem of drifts existing in chemical sensors used in artificial olfaction systems, through which it can mitigate the effect caused by drifts, to reduce errors in the classification of different volatile compounds. The proposed methodology is based on the application of techniques multivariate statistical analysis, Commons Principal Component Analysis (CPCA) combined with the components correction technique (CC), developed by (Arthurson, et al, 2000) and was determined a criterion for selecting the number of principal components common to be subtracted from measures to improve the accuracy in the detection odors process.

The proposed approach in this work, that had as start point the investigation of (Ziyatdinov, and others, 2009), where the authors used only the first principal component common to the correction of drifts in order to take this correction as a linear. The drift components present artificial olfaction systems are nonlinear, hence in this work was incorporated as novelty remove a greater number of components to mitigate its effect and in also considering not capture relevant to the classification system information.

Results of subtracting more components in correcting drifts demonstrated that removal of commons principal components by component correction, produces increases in the percentages of success in the classification process on the validation data. The suitable number of components to be removed through an indicator of the separability of the data sets, calculated from the data set used for training is determined.

INTRODUCCIÓN

Un sistema electrónico de reconocimiento de olores es un instrumento que combina arreglos de sensores de gases y técnicas estadísticas de reconocimiento de patrones para la detección, identificación, o cuantificación de compuestos volátiles (Gutierrez-Osuna, 2002). Estos sistemas de reconocimiento de olores se componen, entre otras partes, de una matriz de sensores químicos con sensibilidades solapadas para poder detectar una amplia variedad de aromas. Lo anterior significa, que no son selectivos a un compuesto químico dado, pero sí levemente más sensibles a determinadas familias químicas tales como solventes orgánicos, ácidos grasos, gases sulfurosos, entre otros. De esta forma, la respuesta de dichos sensores consiste en señales características para cada mezcla química, siendo sensibles a una amplia variedad de productos (Durán Acevedo, 2005).

Los sistemas de olfato electrónico resultan útiles en una gran cantidad de aplicaciones desde la detección de escape de gases tóxicos, medición de nivel y factores de contaminación, detección de bombas, narcóticos, análisis y diagnóstico de enfermedades, monitoreo, hasta la conservación y control de calidad de alimentos (Pearce, Schiffman, Nagle, & Gardner, 2003). En particular, presentan grandes ventajas en el sector agroalimentario, entre las que se destacan las siguientes:

- Análisis no destructivo del producto.
- Obtención de resultados en tiempo real (en segundos o minutos).
- Portabilidad, robustez y bajo precio.
- Adaptación a diferentes cantidades y variedades de productos.
- Facilidad de uso por parte de personal no cualificado.

Entre las aplicaciones específicas en la industria de alimentos se destacan, de modo genérico, la determinación de la calidad de las materias primas, la evolución del producto durante la producción, control de procesos de cocción, monitorización de procesos de fermentación, control de rancidez, verificación de ingredientes para zumos, graduación alcohólica de licores, inspección de olores en contenedores, entre otros; como por ejemplo en (Brezmes, López, Llobet, Vilabona, & et. al., 2005), se describe un equipo para monitorizar el grado de conservación de la fruta; también se han realizado investigaciones relacionadas con la cosecha de vinos (Guadarrama, Fernández, Ïñiguez, Souto, & de Saja, 2001), otras para determinar el grado de conservación del pescado (Zhang, Wang, Liu, Xu, & Zhou, 2012), en el control de calidad del café (Rodríguez, Durán, & Reyes, 2010) y además en la clasificación o identificación de aromas complejos (Rodríguez-Gamboa, Albarracín-Estrada, & Delgado-Trejos, 2011).

OBJETO DE ESTUDIO

Una fuerte restricción en la tecnología de la matriz de sensores, además de las limitaciones en la selectividad y la sensitividad, se presenta a partir de las derivas de los sensores (Ziyatdinov, y otros, 2009). La deriva es capaz de degradar la respuesta del sistema y éste efecto es aún mas considerable cuando se realiza el análisis a través de múltiples mediciones hechas en largos periodos de tiempo, debido a que cuanto mas ciclos de uso tienen los sensores la respuesta de los mismos varía por efectos del envejecimiento en la capa de sensado, aún considerando que las mediciones se realicen bajo condiciones controladas.

De acuerdo a lo anterior, el efecto de las derivas en la respuesta del sistema de reconcimento de patrones, ocasiona que los porcentajes de exactitud en la discriminación de olores disminuyan cuando se intente clasificar compuestos volátiles que se hayan introducido al sistema sensor en un periodo de tiempo amplio con respecto a los datos del conjunto que se tomó como referente para el entrenamiento. De esta forma, se establece que las derivas son un proceso dinámico, causado por cambios químicos en los sensores, los cuales dan una señal inestable a lo largo del tiempo. Además, las muestras y el operador através de la contaminación del instrumento pueden también introducir derivas (Arthurson, y otros, 2000). Por tal razón, ésta tesis de maestría se enfoca en lograr mitigar el problema de las derivas en los sistemas de reconocimiento de olores a partir de la corrección de componentes por CPCA.

ACERCAMIENTO AL PROBLEMA

Como lo explica (Gutierrez-Osuna, 2002) mediante la F**igura 1**, el proceso para la detección de volátiles se divide en cuatro estados secuenciales: preprocesado de las señales, reducción de dimensionalidad, predicción y validación. El bloque inicial en la **Figura 1** representa el hardware del sistema de reconocimiento de olores, el cuál típicamente consiste en una matriz o arreglo de sensores de gases, un subsistema de suministro de olores, una etapa de instrumentación electrónica y un computador para la adquisición de datos.

El proceso de análisis de los datos inicia después de haber adquirido y almacenado las señales de los sensores en el computador. En ésta primera parte se incluye la compensación de la deriva del sensor, además de extraer parámetros descriptivos de la respuesta de la matriz de sensores y preparar el vector de características para el procesamiento. Una etapa de reducción de dimensionalidad proyecta este vector de características inicial sobre un nuevo espacio representativo, con el fin de evitar problemas relacionados con gran cantidad de datos, es decir, conjuntos de datos dispersos. El vector de características de baja dimensionalidad resultante, es usado entonces para resolver un problema de predicción dado, típicamente clasificación, regresión o agrupamiento. El bloque final, realiza la selección y ajuste de parámetros y la estimación de las tasas de error verdadero para el modelo de entrenamiento por las diferentes técnicas de validación.

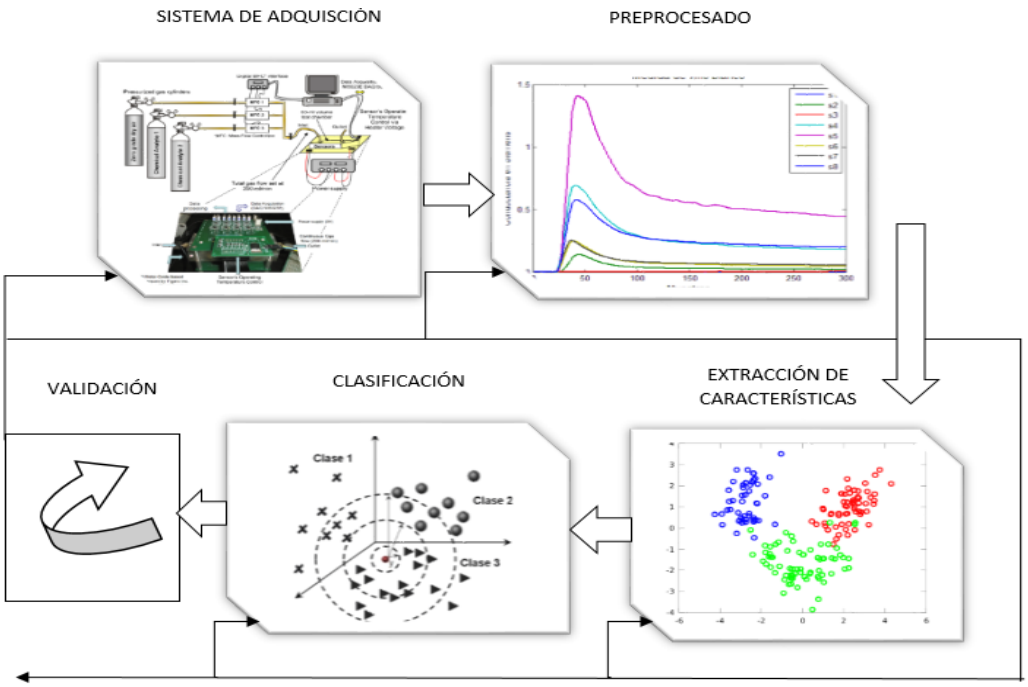


Figura 1. Bloques del sistema de reconocimiento de patrones de un sistema de olfato electrónico. (Gutierrez-Osuna, 2002) , (Vergara, y otros, 2012)

La **Figura 2** muestra las diferentes formas de abordar el problema de las derivas, según la literatura consultada. El primero de ellos corresponde a la corrección de derivas desde el diseño y la construcción de nuevos sensores, el segundo corresponde a la corrección de las derivas en la etapa de clasificación, diseñando e implementando potentes clasificadores que mejoran la exactitud en la respuesta y el poder discriminante de los sistemas de olfato electrónico y el tercer enfoque enfrenta el problema desde el espacio de representación del conjunto de características de los volátiles sensados, es decir, se realiza la mejora de la respuesta de los sensores en la etapa de caracterización para continuar con el ciclo de clasificación y reconocimiento de patrones.

Cada uno de los tres enfoques tiene ventajas y desventajas para la eliminación de las derivas, los cuales se mencionan en la **sección** 1.4, pero debido a las características del tercer enfoque, este trabajo se centra en las técnicas que abordan el problema desde el procesamiento de las señales obtenidas de los sensores de gases, como los son el análisis de componentes principales (PCA) y el análisis de componentes principales comunes (CPCA).

Considerando que para que un sistema de olfato electrónico sea útil es necesario minimizar las derivas (Durán Acevedo, 2005), la importancia de este trabajo radica en mejorar la técnica de corrección de componentes basado en análisis de componentes principales comunes CC-CPCA (Ziyatdinov, y otros, 2009), en donde la novedad radica en hacer un tratamiento no lineal a la deriva determinando el número máximo de componentes a remover, por lo tanto se logró extraer no solo la primera componente principal común, sino

un mayor número de componentes afectando mínimamente la información relacionada con las concentraciones de los gases medidos.

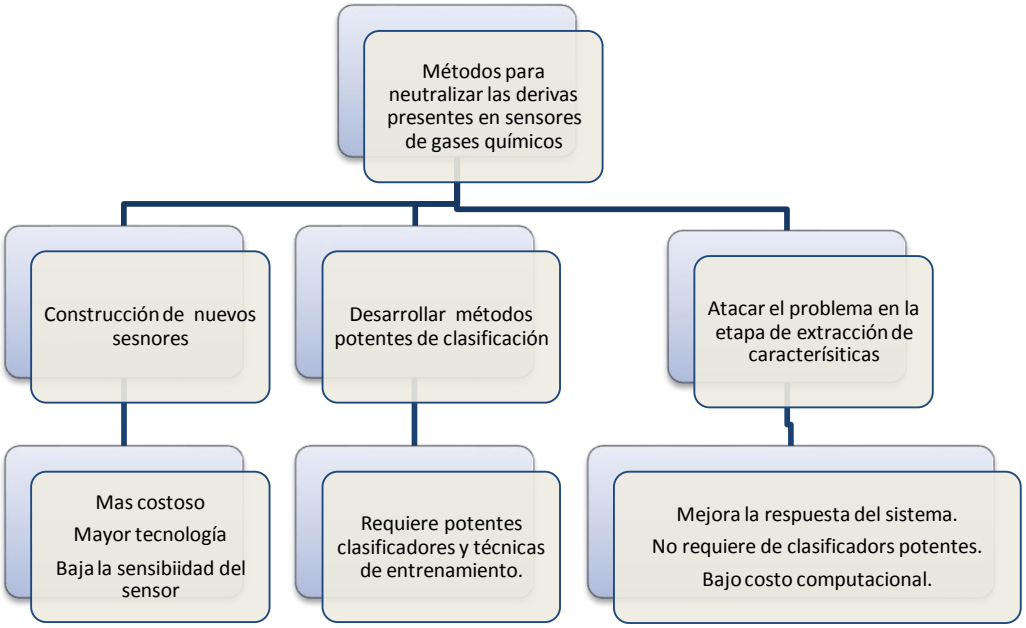


Figura 2. Formas de abordar el problema de neutralizar derivas en sensores químicos para sistemas de olfato electrónicos según la literatura consultada.

Siguiente a esta introducción, se presenta en ésta tesis un capítulo con el marco teórico y el estado del arte, que contienen la temática relacionada con los sistemas de olfato electrónico, las derivas y las diferentes formas de abordar éste problema. Asímismo, se describen las técnicas estadísticas de análisis multivariado que fueron empleadas en la ejecución de este proyecto de maestria, como lo son, el Análisis de Componentes Principales (PCA) y el Análisis de Componentes Principales Comunes (CPCA) y finalmente la fundamentación teórica de la Corrección de Componentes (CC) y su efecto en la neutralización de las derivas. En el segundo capítulo se encuentra el diseño experimental, donde se describen las bases de datos empleadas y la metodología empleada en el procesado de las señales. Un tercer capítulo describe los resultados obtenidos y como parte final se detallan las conclusiones.

La hipótesis de este trabajo, planteó el uso de la técnica de corrección de componentes por análisis de componentes principles comunes (CC-CPCA) y a partir de ella lograr determinar el criterio para hallar cuántas y cuáles componentes principales asociadas directamente con las derivas se deben remover mediante esta técnica. Se estimó que utilizando un orden de componentes superior a uno, se lograba conseguir una mayor exactitud en el sistema de reconocimiento y clasificación, obteniendo mejorar la repetitiidad y la reproducibilidad de los sistemas de clasificación de volátiles.

OBJETIVOS

OBJETIVO GENERAL

Proponer una metodología de selección de las componentes principales que representen las derivas presentes en sensores de gas, empleando corrección por análisis de componentes principales comunes, con el fin de extraer dichas componentes y mejorar el desempeño de un sistema de reconocimiento de olores.

OBJETIVOS ESPECÍFICOS

1. Implementar una estrategia basada en análisis de componentes principales comunes para obtener un espacio de representación efectivo de las respuestas de sensores químicos sometidos a la presencia de derivas.

2. Diseñar una estrategia basada en métodos de selección de características multivariantes y escoger el número de componentes asociadas con las derivas de los sensores.

3. Validar el método propuesto en la clasificación de gases a partir del espacio de representación obtenido, mediante clasificadores de bajo costo computacional.

1. MARCO TEÓRICO Y ESTADO DEL ARTE

Con el propósito de presentar una perspectiva general de los sistemas de detección y reconocimiento de compuestos volátiles, se inicia la temática de este capítulo con la descripción del funcionamiento y las partes que constituyen esta clase de sistemas; a continuación el contenido se enfoca en los sensores de gases, su flujo de trabajo y la forma en que detectan los volátiles para secuencialmente generar las señales eléctricas necesarias en el reconocimiento de olores a través del uso adecuado de un sistema de adquisición de datos, los elementos de software y el sistema de reconocimiento de patrones usado para esta tarea específica. A partir de allí, se ahonda en la temática más importante a tratar en este documento que corresponde a las derivas en los sensores químicos y los efectos que éstas causan en las señales de los sensores, por lo tanto, se presentan en este capítulo las tres perspectivas que existen en cuanto a las diferentes formas de abordar el problema de las derivas, siendo éstas, la construcción de nuevos sensores, el desarrollo de potentes clasificadores y el tratamiento de las derivas en el procesado de las señales.

Enfocando el tratamiento de las derivas por la ruta del procesamiento adecuado de las señales, se describen las técnicas univariadas y multivariadas usadas en la mitigación de derivas en sensores químicos. En las técnicas multivariadas se destacan el análisis de componentes principales (PCA) y el análisis de componentes principales comunes (CPCA), que finalmente fueron las técnicas elegidas en el desarrollo metodológico y experimental de este trabajo de investigación, usando como complemento la **corrección de componentes** que se presenta en la **sección** 1.7.3. Es importante considerar que el éxito en la aplicación de éstas dos técnicas depende del correcto pre-procesado de las señales, por lo tanto se añade un subcapítulo previo a las técnicas multivariadas, dedicado a los temas de remoción de datos anómalos y las técnicas de escalado y normalización, etapas necesarias que se desarrollan previamente a la aplicación de las técnicas mencionadas.

1.1 GENERALIDADES DE LOS SISTEMAS DE OLFATO ARTIFICIAL

Un sistema de olfato artificial imita la estructura de la nariz humana, además ambos sistemas están basados en receptores no específicos (células y sensores), seguidos por un adecuado procesamiento de señales (Quicazán, Díaz M., & Zuluaga D., 2010). En consecuencia, los sistemas electrónicos de detección de olores son básicamente arreglos de sensores químicos conectados a sistemas de procesamiento o de cómputo, en los cuales se aplican técnicas avanzadas de procesamiento digital de señales y reconocimiento estadístico de patrones. Su objetivo fundamental es permitir la cualificación de olores a través de tareas de clasificación, discriminación, predicción e incluso cuantificación de productos, elementos o componentes de acuerdo con sus características organolépticas (Durán & Baldovino, 2009), (Wilson & Baietto, 2009), (Zhou, Homer, Shevade, & Ryan, 2005).

Se pueden distinguir los siguientes módulos en un sistema de reconocimiento de olores: módulo de entrega y direccionamiento de la muestra (olor), el sistema de detección que

puede ser una matriz de sensores o un espectrómetro de masas, el módulo de acondicionamiento y procesamiento de la señal, y el módulo de análisis y reconocimiento de patrones (Moreno, Caballero, Galán, Matía, & Jiménez, 2009).

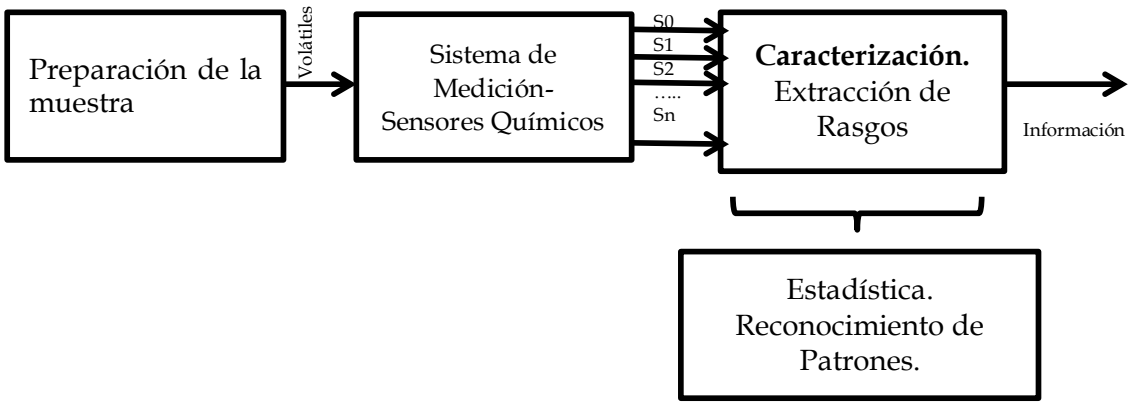


Figura 3. Módulos que conforman un sistema electrónico de reconocimiento de olores.

La **Figura 3** muestra la secuencia de trabajo de un sistema de olfato artificial, cuyos módulos se describen en las secciones subsiguientes.

1.1.1 Preparación de la muestra

Inicialmente la muestra es acondicionada por métodos de extracción de volátiles que permiten el paso del gas a analizar hacia una matriz de sensores. El sistema de muestreo está integrado principalmente por un lugar donde se aloja la muestra (como una cámara de concentración), un sistema de control y un sistema de transporte de flujo (como una bomba de aire, controladores de flujo másico, etc) (Durán Acevedo, 2005).

1.1.2 Sistema de medición

Un sistema de reconocimiento de olores posee como elemento principal una matriz de sensores de gases que es la encargada de trasducir la concentración de volátiles en cambios de su resistencia. Es conveniente que dicha matriz esté localizada en una cámara o compartimiento especial en el que se garanticen las condiciones específicas para su correcto funcionamiento, principalmente se debe asegurar el adecuado aislamiento que impida que se introduzcan contaminantes e igualmente se debe mantener la presión y temperatura adecuadas. Estos parámetros son importantes o críticos en función del tipo de sensor utilizado. (Rodríguez-Gamboa, Albarracín-Estrada, & Delgado-Trejos, 2011)

Otra ventaja de utilizar una cámara de sensores es, facilitar el proceso de medida debido a que los volátiles van a estar en mayor concentración y tienen más contacto con el elemento

activo de los sensores, lo cual permite una mejor y más rápida respuesta de los mismos. También se ha encontrado experimentalmente que cuanto más hermética sea la cámara de sensores, se aprovechan mejor las ventajas mencionadas. En la **Figura 4** se aprecia la fotografía de una cámara de sensores y diferentes clases de sensores químicos.

(a) (b)

Figura 4. Sistema de sensado. **(a)** Cámara de concentración (Rodríguez-Gamboa, Albarracín-Estrada, & Delgado-Trejos, 2011); **(b)** sensores químicos usados en la detección de volátiles (Figaro Company, s.f.).

Este subsistema por lo general presenta derivas en los sensores, por lo que se requiere calibración de los mismos para evitar problemas de repetitividad en las medidas debido a la saturación que se puede presentar. Es importante mencionar que las matrices de sensores de gases usualmente utilizan sensores del mismo tipo pero de diferentes referencias (Ejemplo: TGS822, TGS821, TGS813, etc.), con el fin de obtener un mayor solapamiento entre las señales buscando facilitar las tareas de clasificación y detección de olores (Durán Acevedo, 2005).

Adicional a este módulo se tiene el sistema de transporte de volátiles, que condiciona el funcionamiento y permite que se realicen los procesos de medición y purga de los sensores. Éste básicamente es un sistema que se encarga de transportar hacia la cámara de sensores los volátiles desprendidos por la muestra o elemento que se va a analizar. En algunas ocasiones se inyecta la muestra del olor en la cámara de sensores de forma manual, con los consiguientes problemas de error y lentitud que ello implica; en otras ocasiones un sistema automático se encarga de transportar las moléculas olorosas o volátiles, extrayéndolos de la zona en la que se encuentra la muestra a través de la inyección de algún tipo de gas o aire, hasta llevarlos a la cámara de sensores (Rodríguez, Durán, & Reyes, 2010).

Además, los sistemas de olfato electrónico en su mayoría cuentan con algún tipo de mecanismo de limpieza de la cámara de sensores de forma que las medidas sucesivas se hagan partiendo de las mismas condiciones iníciales y se garantice la repetibilidad de los resultados.

1.1.3 Sistema de Procesamiento

El sistema de procesamiento en la mayoría de los casos está compuesto por un computador con el software adecuado para encausar los datos obtenidos de los sensores. A estos datos se les aplican técnicas de pre-procesamiento para extraer los parámetros estáticos de las medidas y reducir la cantidad de información a analizar, después se aplican técnicas de análisis multivariado como Análisis de Componentes Principales (PCA) y de reconocimiento de patrones como Redes Neuronales Artificiales (RNA), Máquinas de Vectores de Soporte (SVM), entre otras, para realizar tareas tales como: clasificación, discriminación, predicción, cuantificación de muestras de acuerdo a sus características organolépticas (Wilson & Baietto, 2009); (Berna, 2010).

1.2 SENSORES DE GASES

En general, los sensores de gases son dispositivos que constan de dos partes principales, la primera es un elemento activo cuyas propiedades físicas o químicas cambian en presencia del analito que se desea detectar y la segunda parte es un elemento transductor que convierte los cambios de las propiedades del elemento activo en una señal eléctrica. Estos sensores generalmente tienen una membrana selectiva que impide el paso de partículas o material indeseable, actuando como un primer filtro de ruido. En la **Figura 5** se puede observar un esquema simplificado de un dispositivo de este tipo, en el cual se pueden apreciar las principales partes de un sensor de gas y la naturaleza de las entradas y salidas (Tian, Yang, & Dong, 2005).

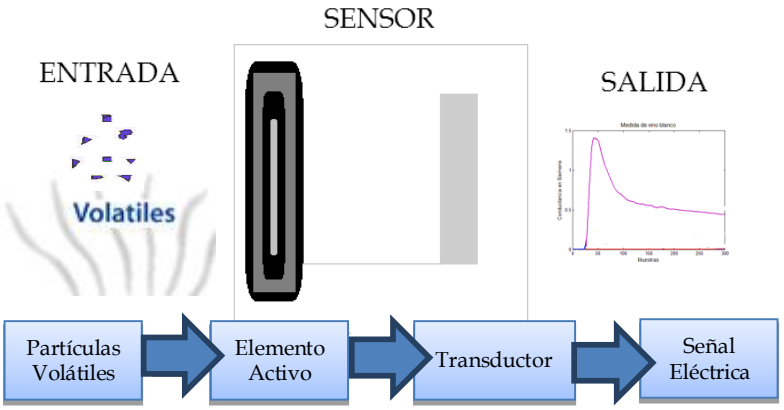


Figura 5. Esquema simplificado de un sensor químico para detección de compuestos volátiles.

Existen diferentes tipos de sensores de gases para emplear en los sistemas de reconocimiento de olores, los más utilizados son: MOX (*Metal Oxide Semiconductor*), QCM (*Quartz Crystal Microbalance*), SAW (*Surface Acoustic Waves*), MOSFET (*Metal Oxide*

Semiconductor Field Effect Transistor), CP (*Conducting Polymers*), FO (*Fiber Optics*). En esta tesis de maestría se trabajó específicamente con datos generados por los sensores MOX, construidos con materiales semiconductores como el óxido de estaño (SnO_2), Óxido de Zinc (ZnO), Oxido de Titanio (TiO_2), entre otros. El principio de funcionamiento de ésta clase de sensores se basa en el cambio de la conductividad de un material sensible cuando éste reacciona con los gases presentes en su entorno.

Tras ser adquiridas y almacenadas, las señales de los sensores son tratadas por métodos de extracción de parámetros y pre-procesado de datos. La técnica de extracción de parámetros es fundamental, especialmente al utilizar sensores de óxido de estaño. Estos basan su funcionamiento en el cambio de conductividad que experimenta el material o capa activa del sensor ante la presencia de gases reductores y/o oxidantes. El cambio de conductividad experimenta transitorios que llevan a la capa activa del sensor desde una situación de reposo a una conductancia que depende del tipo de volátil y su concentración (Durán Acevedo, 2005).

El tipo de señal obtenido a partir de los sensores de gases es mostrado en la **Figura 6**, donde se representa la respuesta del sensor ante una medida de amoniaco. Como se mencionó anteriormente, en los sistemas de olfato artificial se utilizan matrices o arreglos de sensores de gases con las cuales se obtienen señales como las mostradas en la **Figura 7**.

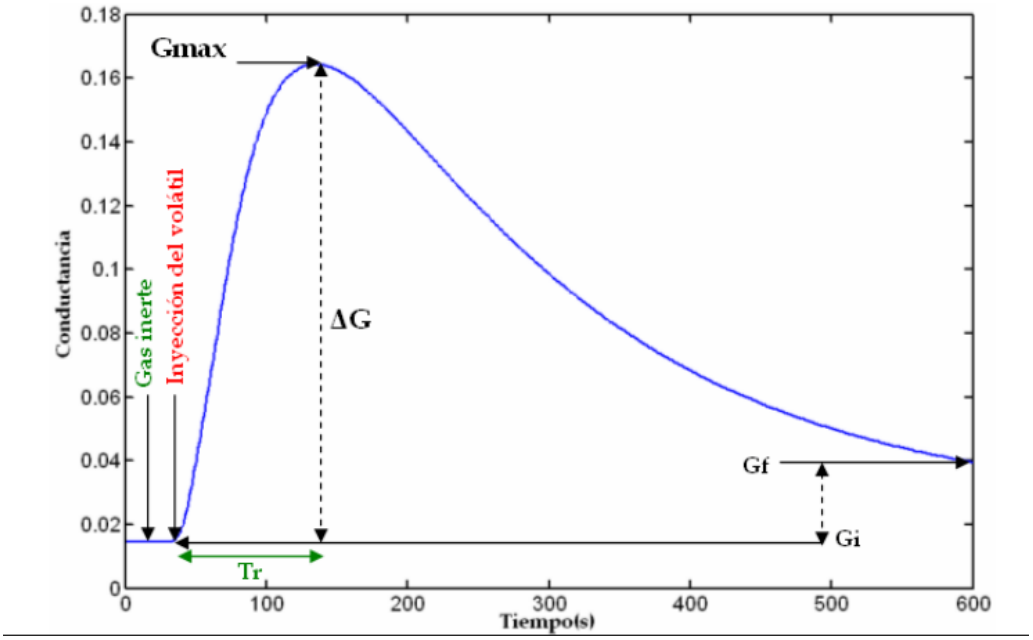


Figura 6. Señal de un sensor de gas para una medida de amoniaco (Durán Acevedo, 2005).

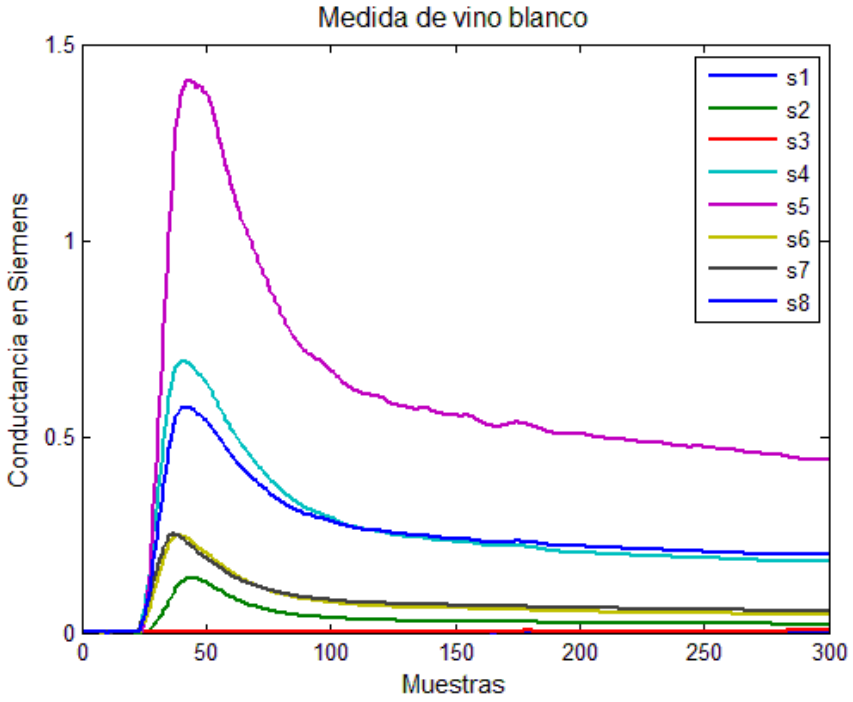


Figura 7. Señal de una matriz de 8 sensores de gases en una medida de vino blanco (Base de datos A-NOSE) (Rodriguez Gamboa, 2013)

1.3 DERIVAS EN SENSORES DE GASES

La deriva en dispositivos con arreglos de sensores químicos, tiene un efecto bastante complejo e inevitable que se genera por diferentes fuentes, entre ellas se destacan el envejecimiento del sensor y el envenenamiento del dispositivo que se refleja directamente a través de un cambio en la capa química para la detección de volátiles (reorganización del material sensor y contaminación). También se encuentra implícita la operación experimental que incluye los efectos térmicos y de memoria de los sensores, los cambios en el entorno y la aparición de otras señales causadas por el ruido del sistema (Ziyatdinov, y otros, 2009).

1.3.1 El concepto de deriva

La respuesta de un sensor de gas contiene no solo su señal verdadera, sino también algunas perturbaciones, estas a su vez, se componen de múltiples frecuencias y todas ellas afectan la señal. La parte que corresponde a alta frecuencia es llamada ruido y la que se compone de baja frecuencia es a menudo conocida como **deriva**, la cual puede ser vista como un cambio gradual a través del tiempo en la respuesta del sensor bajo condiciones constantes. La deriva es un proceso dinámico causado por cambios químicos en los sensores más específicamente en la capa activa de los mismos (Arthurson, y otros, 2000).

1.3.2 Problemas ocasionados por las derivas

La limitación más seria de los actuales sistemas de olfato artificial son las derivas inherentes a los sensores de gases, las cuales ocasionan en el tiempo una lenta variación aleatoria de la respuesta del sensor, cuando es expuesto a algunos gases bajo condiciones idénticas (Gutierrez-Osuna, Pattern analysis for machine olfaction: a review, 2002).

Una consecuencia de la influencia de la deriva, la cual puede afectar la parte de la línea base del sensor cuando es aditiva y la sensitividad cuando es multiplicativa, es el hecho de lograr que el aprendizaje previo de los patrones de las señales entregadas por los sensores se vuelva obsoleto a través del tiempo y en consecuencia, los sistemas pierden la habilidad para identificar los olores ya reconocidos. El medio más efectivo para la compensación de la deriva es la recalibración periódica con un gas de referencia que es químicamente estable y altamente correlacionado con los analitos objetos de análisis, en términos del funcionamiento del sensor. De esta forma, la respuesta de la matriz de sensores para el gas de calibración, puede restarse directamente de la respuesta de los analitos, así se deduce un modelo de deriva temporal para cada sensor individual o para la matriz de sensores (Gutierrez-Osuna, 2002).

Para entender los efectos ocasionados por las derivas, se puede observar la **Figura 8,** en donde se grafican las componentes principales del análisis PCA de los datos. Se observa claramente en la imagen izquierda, como la deriva ocasiona que los datos resultantes de la matriz de sensores presentan cambios visualizados como datos con mayor dispersión y menor separación entre clases. La imagen del lado derecho representa las componentes principales de los datos a los cuales se les ha aplicado corrección de la deriva, por medio de la técnica de corrección de componentes que se aplicó en este trabajo.

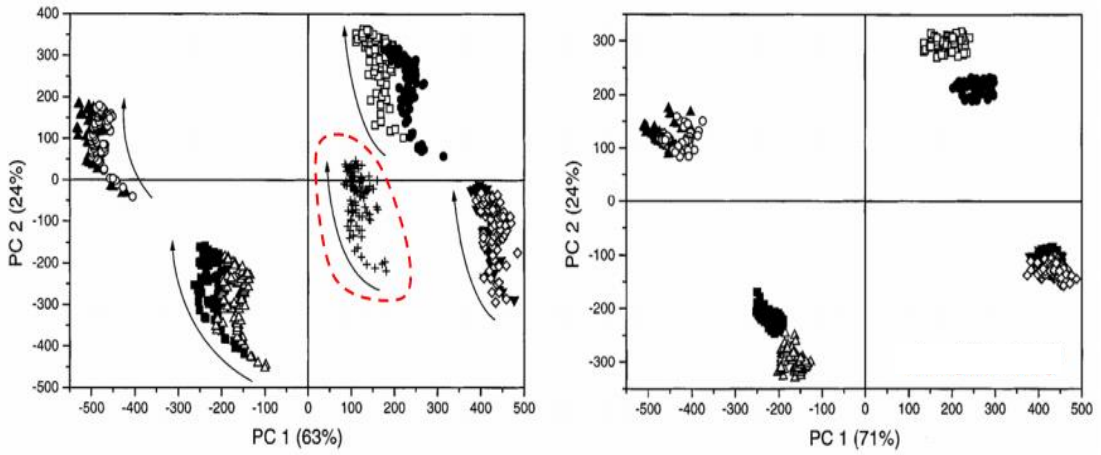


Figura 8. Proyección de las dos primeras Componentes Principales de la respuesta de un conjunto de sensores ante la presencia de diferentes mezclas de gases. A la izquierda antes y a la derecha después de la compensada la deriva (Gutierrez-Osuna, 2002).

Este mismo efecto puede observarse si se analiza la señal de respuesta de los sensores en el eje del tiempo, tal como se aprecia en la F**igura 9.** Esta imagen fue tomada y adaptada del trabajo presentado por (Ziyatdinov, y otros, 2009), para describir la fuerte influencia de las derivas en las señales entregadas por los sensores, la figura muestra las señales de estado estable de un sensor en un periodo de tiempo de 7 meses cuando el sensor es sometido a tres clases de gases con diferentes concentraciones.

En la **Figura 9,** cada bloque de medición separado por las líneas verticales continuas, corresponde a un periodo de siete meses. Las señales en color negro, rojo y verde corresponden a las mediciones del gas A (amoniaco) con tres concentraciones diferenciadas por los tres colores. Las señales en azul, cian y magenta corresponden a las mediciones de tres concentraciones diferentes del gas B (ácido propanoico) y finalmente las señales en amarillo y gris corresponden a dos concentraciones diferentes del gas C (n-buthanol). En cada uno de los ocho bloques graficados, las líneas punteadas indican la separación entre las muestras empleadas para el entrenamiento y las restantes, es decir, las que están después de la línea punteada en cada bloque fueron usadas para la validación. Es evidente el cambio de las señales a lo largo del tiempo, lo cual finalmente se traduce en la degradación del sistema empleado para la clasificación.

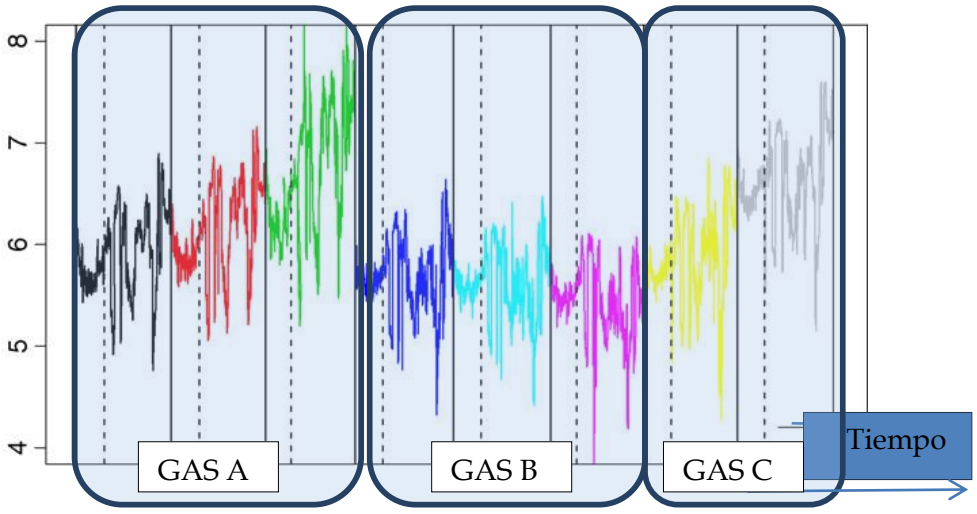


Figura 9. Análisis del comportamiento en el tiempo de las señales de respuesta de un sensor sometido a la presencia de tres gases con diferentes concentraciones, influenciado por las derivas (Ziyatdinov, y otros, 2009)

Los picos de las señales indican presencia de derivas a corto plazo causados por algunos cambios temporales, como lo son por ejemplo el calentamiento de los sensores; por otra parte las derivas a largo plazo se pueden observar en los cambios en la línea base de señales similares para todas las clases.

Otra observación importante sobre esta gráfica corresponde a las señales de respuesta del sensor ante la presencia del ácido propanoico, estas tienen un comportamiento mas estable a través del tiempo y es menos propenso al efecto de las derivas. Sin embargo, es claramente observable como la respuesta del sensor cambia a través del tiempo, en este caso el periodo de tiempo analizado fue de tan solo siete meses, tiempo suficiente para requerir una recalibración de los sensores.

El comportamiento de los datos, tal como se refleja en el análisis anterior ocasiona que el sistema de clasificación producto del reconocimiento de patrones se torne obsoleto después de cierto tiempo.

1.4 FORMAS DE ABORDAR EL PROBLEMA DE LAS DERIVAS

En la literatura se encuentran principalmente tres enfoques diferentes para mitigar los efectos de las derivas. El primero de ellos busca el mejoramiento de la parte física, diseñando y construyendo nuevos sensores. El segundo apunta al desarrollo de potentes clasificadores, lo que necesariamente implica mayor complejidad en los algoritmos de reconocimiento de patrones y mayor costo computacional. Finalmente, el tercer enfoque se orienta a la corrección de la deriva desde la etapa de procesado, aplicando técnicas de análisis estadístico multivariado para lograr una representación efectiva de las derivas y de ésta manera poder eliminarlas del sistema. A continuación se relacionan algunos de los trabajos direccionados en cada uno de los tres enfoques aquí planteados.

1.4.1 Construcción de nuevos sensores para mejorar el problema de las derivas

Enfocados en la construcción de sensores más robustos se encuentra que (Song, Wang, Zhang, & Cheng, 2011) desarrollan un nuevo sensor fabricado con Fe_2O_3 basado en sistema de olfato electrónico inalámbrico usando mínimos cuadrados con regresión de soporte vectorial, este tipo de sensores son insensibles a la humedad y por lo tanto son óptimos para trabajar en el espacio libre, en este articulo se usan esta clase de sensores para la detección de CH_4 y H_2 o sus mezclas. Sin embargo, sólo se especifica insensibilidad a la humedad y no se menciona otra clase de derivas, como por ejemplo la temperatura para lo que realizan un acondicionamiento de señal. (Kwan, Boussaid, & Bermak, 2011) crean un único chip CMOS de reconocimiento de gases para reemplazar el arreglo de sensores de gases de metal-óxido. En este trabajo se construye un chip al que le realizan un método de compensación de las derivas y optimizan el proceso de recalibración de los sensores, obteniendo porcentajes de detección en una tasa del 94,9%, cuando el chip es expuesto a propano, etanol y monóxido de carbono.

Aunque se evidencia la mejora introducida, al enfrentar el problema desde este enfoque se incrementan los costos de los sensores y la facilidad con que estos pueden conseguirse en el mercado, pues los de mayor uso en la actualidad son sensores químicos de óxido de estaño que aún tienen problemas de derivas. Más aún, este tipo de estudios se categoriza para

aplicaciones especiales, donde la instrumentación, los elementos y laboratorios empleados son de uso restringido. Además, la capacidad de detección de este tipo de sensores es tan solo de unas partes por millón (ppm), mientras que los sensores químicos comerciales de óxido de estaño presentan sensibilidades de hasta partes por billón (ppb).

1.4.2 Corrección de las derivas en la etapa de clasificación

El segundo enfoque planteado consiste en abordar el problema de las derivas en la etapa de clasificación. (Paniagua, Llobet, Brezmes, Vilanova, & et. al., 2003), desarrollan una técnica para la neutralización de derivas en arreglos de sensores de gases de metal-oxido semiconductor MOS. Demuestran la técnica para tres compuestos volátiles con aciertos desde el 67,5% hasta el 100%, usando un arreglo de sensores. La técnica empleada comprende un aprendizaje en línea usando FUZZY ARTMAP y mencionan que se podría lograr una recalibración automática de los sensores. (Fu, Li, Qin, & Freeman, 2007) desarrollan un método de reconocimiento de patrones para narices electrónicas basado en una red neuronal caótica llamada KIII, para identificar seis compuestos volátiles orgánicos en el arroz. Con esta red neuronal investigan la capacidad de neutralización de los tiempos de concentración y la disminución de las derivas en sensores TGS. Encuentran que la red neuronal KIII tiene un buen desempeño para discriminar los volátiles del arroz, con bajos tiempos de concentración y realizando medidas un mes más tarde.

Otro trabajo direccionado hacia la corrección de derivas en la etapa de clasificación es el de (Cho, Kim, Jin Na, & Jeon, 2008), en donde diseñan un sistema de olfato electrónico inalámbrico para cuantificar las concentraciones de amoniaco y sulfaldehído de hidrógeno o sus mezclas. Utilizan lógica fuzzy como software de clasificación y emplean un factor de corrección de las derivas que le da una buena estabilidad en la respuesta del sensor. Plantean que en futuros trabajos se puede investigar en términos más avanzados sobre la estabilidad en sensores y sistemas de redes de sensores de olfato inalámbricos y del efecto de la humedad en procesos de clasificación y estimación.

Asímismo, (Vergara, y otros, 2012) realizan la mitigación de la deriva de sensores de gases usando conjuntos de clasificadores. En ésta metodología, una gran cantidad de datos para seis diferentes compuestos orgánicos volátiles, fue recopilada en un período de tres años bajo condiciones estrictamente controladas utilizando una matriz de 16 sensores de óxido de metal. En este caso específico se optó por máquinas de soporte vectorial como clasificador base en el conjunto. Los experimentos indican claramente la presencia de la deriva de los sensores durante el período de tres años, lo que degrada el rendimiento de los clasificadores. Sin embargo, el método de ensamble que utiliza una combinación ponderada de los clasificadores entrenados en diferentes puntos del tiempo es capaz de clasificar aún con presencia de la deriva del sensor.

En los trabajos mencionados, el aprendizaje del clasificador corrige o reduce gradualmente las derivas de sensores comunes de sistemas de olfato electrónico, como lo son los sensores TGS. La corrección de las derivas en la etapa final del reconocimiento de patrones, es decir,

en el clasificador, requiere de técnicas computacionales potentes para disminuir las derivas que vienen inmersas en los datos. Aunque este enfoque mitiga el efecto de las derivas en las respuestas del sistema de olfato electrónico, tiende a aumentar el costo computacional en la implementación de las soluciones planteadas al incorporar clasificadores altamente demandantes o incluso la utilización de múltiples clasificadores. Además, la metodología en general permite poca capacidad de adaptación ya que se encuentra directamente ligada con un clasificador específico.

1.4.3 Corrección de las derivas en la etapa de procesamiento

El tercer enfoque aquí planteado corresponde a la neutralización de las derivas desde la caracterización y el espacio de representación de los volátiles sensados. Se encuentra en esta línea el mayor número de trabajos realizados, entre los que cabe mencionar los siguientes:

En (Kermit & Tomic, 2003), emplean el análisis de componentes independientes (ICA) aplicados en los datos de mediciones de sensores de gases y lo comparan con el análisis de componentes principales (PCA). Se comprueba que ICA tiene mejor desempeño que PCA porque el primero es capaz de manejar la deriva del sensor combinado con la mejora de la discriminación, la reducción de dimensionalidad, y la representación más adecuada de los datos en comparación con PCA.

En (Kashwan & Bhuyan, 2005), se construye un robusto sistema de olfato electrónico con compensación de temperatura y humedad para discriminación de sabores de té y especias. Determinan los coeficientes de las derivas causadas por temperatura y humedad tomando las variaciones en las muestras, estos coeficientes los usan para eliminar las derivas en la respuesta de la nariz electrónica durante la captura y el procesamiento de los datos. Utilizan una red neuronal artificial y aplican PCA para la discriminación y clasificación de los diferentes sabores de té y sus especias. En los resultados expresan que la compensación de las derivas logra incrementar los porcentajes de clasificación en un 4 a 5% aproximadamente.

En (Perera, Papamichail, Barsan, Weimar, & et.al., 2006) realizan una técnica de detección utilizando un análisis dinámico recursivo de componentes principales (RDPCA) en una matriz de sensores de gases bajo condiciones de deriva. Este novedoso método es adaptativo y logra corregir los efectos de las derivas de los sensores, que en comparación con el método PCA en presencia de una nueva varianza los resultados de la descomposición PCA cambian drásticamente. Con el método propuesto en este artículo se utiliza información de respuestas en tiempo pasado de la matriz de sensores para minimizar las derivas en la salida del sistema. En los resultados logran demostrar la eficiencia del novedoso método propuesto.

En (Huang & Leung, 2009), utilizan el método de Papoulis-Gerchberg para desarrollar un algoritmo iterativo de reconstrucción que minimice los errores causados por las derivas de

los sensores. Concluyen que este método tiene buen desempeño y que no se requiere un gas de referencia y además es relativamente fácil de implementar.

En (Padilla, y otros, 2010) se realiza una compensación de derivas de los datos de un arreglo de sensores de gases utilizando corrección de señales ortogonales (OSC), este método es comparado con el método de corrección de componentes (CC). Se concluye que este método tiene buen desempeño en periodos cortos de tiempo.

En (Ziyatdinov, y otros, 2009) se desarrolla un nuevo método de compensación de derivas utilizando el análisis de componentes principales comunes (CPCA) que se compara con otras metodologías que utilizan PCA. El método alcanza porcentajes de clasificación no inferiores al 80%. Sin embargo, los autores emplean un solo componente principal para representar las derivas, lo cual supone un comportamiento lineal, descartando posibles relaciones más complejas.

(Lonwongtragool, Wongchoosuk, & Kerdcharoen, 2011), desarrollan una nariz electrónica portátil para análisis de la calidad en bebidas alcohólicas, se realizó compensación a las derivas de los sensores a partir de una señal de referencia. Emplean PCA como método de caracterización y logran diferenciar entre vinos que se empacan al vacío y vinos que se empacan en un ambiente con nitrógeno para ayudar al fabricante a diseñar el proceso adecuado de embotellado de vinos y el logro de una alta calidad de productos vinícolas.

Este enfoque resulta ser el más apropiado en comparación con los anteriormente descritos, ya que los métodos de extracción de características empleados para abordar la neutralización de las derivas mejoran la respuesta del sistema de olfato electrónico y lo convierte en un instrumento de uso más general ante multiplicidad de analitos, al mismo tiempo que mantienen costos bajos ya que utilizan sensores comerciales. La técnica CPCA ofrece especial interés debido a su capacidad de representar las respuestas no lineales propias de los sensores de gases. Sin embargo, como lo señala (Ziyatdinov, y otros, 2009), es necesario optimizar la dimensión del espacio obtenido, utilizando un orden de componentes mayor con el fin de mejorar la exactitud del sistema de reconocimiento.

En este punto se hace necesario profundizar sobre este tema, por ello en las **secciones 1.5** y **1.7** se presentan las técnicas empleadas en la corrección de las derivas desde el procesamiento de las señales, considerando los dos enfoques principales que corresponden a las técnicas de análisis univariado y las técnicas de análisis estadísitico multivariado. El análisis univariado consiste en el análisis de una sola variable, en este caso el objeto de análisis sería solo un sensor, mientras que en el análisis multivariado se involucra más de una variable dependiente y varias variables independientes, es decir se analiza la información sobre todos los sensores de la matriz de medición. A continuación, se presenta la fundamentación teórica de estas técnicas para dar mayor claridad a sus ventajas y desventajas.

1.5 TÉCNICAS UNIVARIADAS USADAS EN LA CORRECCIÓN DE DERIVAS

En las técnicas univariadas se mencionan el análisis en frecuencia, la manipulación de la línea base, las medidas diferenciales (calibración) y la corrección multiplicativa (calibración).

1.5.1 Filtrado de la Señal – Análisis en Frecuencia.

Considerando que las derivas, el ruido y la información de los volátiles ocurren en diferentes escalas de frecuencias, es decir las derivas aparecen a bajas frecuencias y el ruido ocurre a altas frecuencias, es procedente aplicar el filtrado como técnica de pre-procesado de las señales adquiridas. El ruido puede ser inherente a la adquisición y el proceso de digitalización, en consecuencia, se hace necesario emplear algún tipo de filtro que disminuya la componente de ruido aleatorio de las señales. En este caso se analizan los siguientes filtros: el de media móvil en el tiempo y el filtro *Butterworth*.

1.5.2 Filtro de media móvil en el tiempo.

Es un filtro tipo ventana que se desplaza a través de los datos. La expresión matemática de un filtro de media móvil viene dada por:

$$y(i) = \frac{1}{M}\sum_{j=0}^{M-1} s[i+j] \qquad (1)$$

Donde $s[i]$ es la señal de entrada, $y[i]$ es la correspondiente señal de salida y M es el número de puntos escogidos para promediar (tamaño de la ventana), denominado factor de anchura del filtro. El procedimiento corresponde a calcular los promedios de los M puntos de los datos de entrada e ir desplazándose por los demás datos de entrada repitiendo el procedimiento (Guiñón, Ortega, García-Antón, & Pérez-Herranz, 2007).

1.5.3 Filtro Butterworth.

Los filtros *Butterworth* provienen, como muchos otros elementos del análisis de series temporales cuando existe la necesidad de procesar señales. Los filtros *Butterworth* permiten tanto la estimación de tendencias a largo plazo como la extracción directa de una señal cíclica (Bógalo & Quilis, 2003).

Los filtros *Butterworth* de paso bajo son operadores ARMA (Modelos Auto-regresivos de Media Móvil) cuya función de ganancia obedece a la siguiente expresión:

$$|G(\omega)|^2 = \frac{1}{1+\left[\frac{tan(\omega/2)}{tan(\omega_C/2)}\right]^{2d}} \qquad (2)$$

Donde w es la frecuencia expresada en radianes y está entre $0 \leq \omega \leq \pi$. Esta función está controlada por dos parámetros: la frecuencia de corte (w_c) y el grado del filtro d (Bógalo & Quilis, 2003).

1.5.4 Manipulación de la línea base

Uno de los métodos más simples propuestos para la compensación de las derivas, el cual es ampliamente usado como un método de pre-procesado, consiste en la transformación de las señales de los sensores individuales basados en el valor inicial de la respuesta transitoria llamada línea base, por esto recibe el nombre de manipulación de línea base.

En los sensores de gases se pueden aplicar técnicas de pre-procesamiento para manipular la línea base, produciendo una transformación de la respuesta del sensor relativa a esta línea, con el objeto de mitigar o compensar los efectos de las derivas, como se presenta en el trabajo de (Gonzalez-Jimenez, Monroy, & Blanco, 2011), donde los autores reportan que al inicio de cada experimento las diferencias en la línea de base de los sensores fueron corregidas mediante la adición de una compensación a cada sensor. Para ello un factor de multiplicación fue calculado para cada sensor, con el fin de asegurar la obtención de valores de línea base idénticos. Debido al comportamiento no lineal de los sensores se seleccionó una ganancia media calculada a partir de tres diferentes concentraciones. Cabe destacar que bajo el supuesto de los experimentos de corto plazo, la deriva de referencia por humedad, temperatura o incluso intoxicación, es insignificante, por lo cual no se tuvieron en cuenta para este estudio. Los métodos comúnmente empleados para este propósito son: diferencial, relativo y fraccional. Estas tres transformaciones básicas son explicadas en (Di Carlo & Falasconi, 2012) y son comunes en la práctica:

A. Diferencial

El método diferencial consiste en restar la línea base al conjunto de datos y puede ser utilizado para eliminar la deriva aditiva de la respuesta del sensor (Bahraminejad, Basri, Isa, & Hambali, 2011)

$$\hat{s}(t) = y(t) - y(0) = [x(t) + \delta] - [x(0) + \delta] = x(t) - x(0) \qquad (3)$$

donde:

$\hat{s}$ es la respuesta corregida
y es la respuesta de la medida
x es la respuesta ideal del sensor sin deriva
δ es la contribución de la deriva, la cual es asumida como constante y uniforme

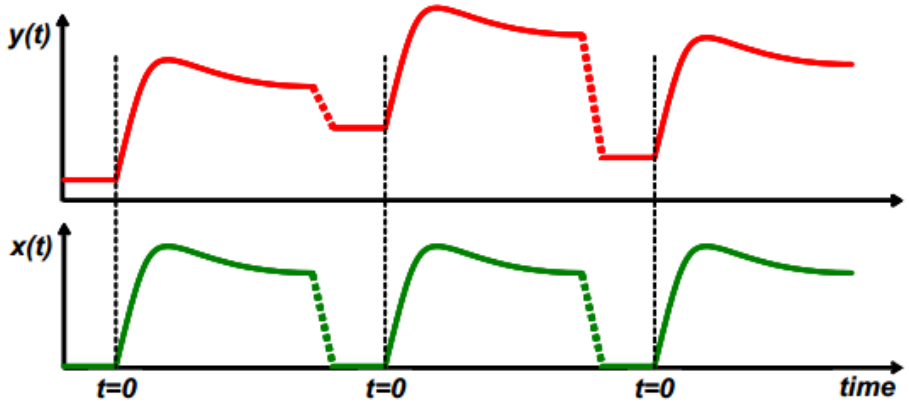


Figura 10. Corrección aditiva de la deriva o deriva de referencia (línea-base), (Gutierrez-Osuna, 2003)

En la **Figura 10** se observa el efecto de la corrección de derivas aditivas aplicando el método diferencial, se destaca que este método se aplica a través de una constante re-calibración. En color rojo se observa la presencia de derivas aditivas en la señal y en color verde se presenta la señal corregida.

B. Relativa

Este método divide por la línea base y podría corregir los efectos de las derivas multiplicativas (constantes y uniformes).

$$\hat{s} = \frac{y(t)}{y(0)} = \frac{x(t)+\delta x(t)}{x(0)+\delta x(0)} = \frac{x(t)(1+\delta)}{x(0)(1+\delta)} = \frac{x(t)}{x(0)} \qquad (4)$$

El efecto de las derivas multiplicativas sobre las señales de los sensores se observa en la F**igura 11,** también se observa el efecto de la corrección aplicando la técnica anterior. La línea roja corresponde a las señales bajo efectos de derivas multiplicativas y la señal en color verde corresponde a la señal corregida.

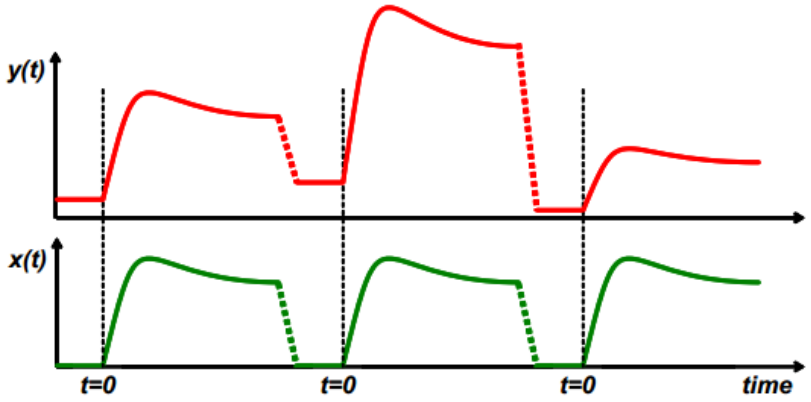


Figura 11. Efecto de las derivas multiplicativas y corrección por medio de la técnica diferencial (Gutierrez-Osuna, 2003)

.

C. Fraccional

Esta técnica es una combinación de las dos anteriores que se aplica para derivas multiplicativas y que tiene la ventaja de proveer medidas de menor dimensionalidad y respuestas de los sensores normalizadas.

$$\hat{s} = \frac{y(t)-y(0)}{y(0)} = \frac{x(t)(1+\delta)-x(0)(1+\delta)}{x(0)(1+\delta)} = \frac{x(t)-x(0)}{x(0)} = \frac{x(t)-x(0)}{x(0)} \qquad (5)$$

Las dos primeras transformaciones son demasiado específicas porque en aplicaciones reales la deriva no es ni aditiva ni multiplicativa, por lo que no son capaces de corregir el efecto de la deriva, mientras que son utilizadas para normalizar la respuesta del sensor. La última transformación no funciona si la deriva aditiva está presente. Por el contrario se puede amplificar el ruido en las mediciones debido a que el término de la deriva que es típicamente pequeño se mantiene en el denominador, degradando así la calidad de la muestra. De ésta manera, estas técnicas proporcionan una pobre corrección de la deriva.

1.5.5 Técnicas de Calibración

La calibración constante, mediante una programación sistemática es posible pero hace que el problema de corrección de derivas se torne más costoso y aún más complejo de tratar. En la F**igura 12** se exhibe un esquema de la recalibración del equipo usando un gas de referencia, es obvio que esta técnica no resulta ser la técnica más adecuada en términos de eficiencia y economía.

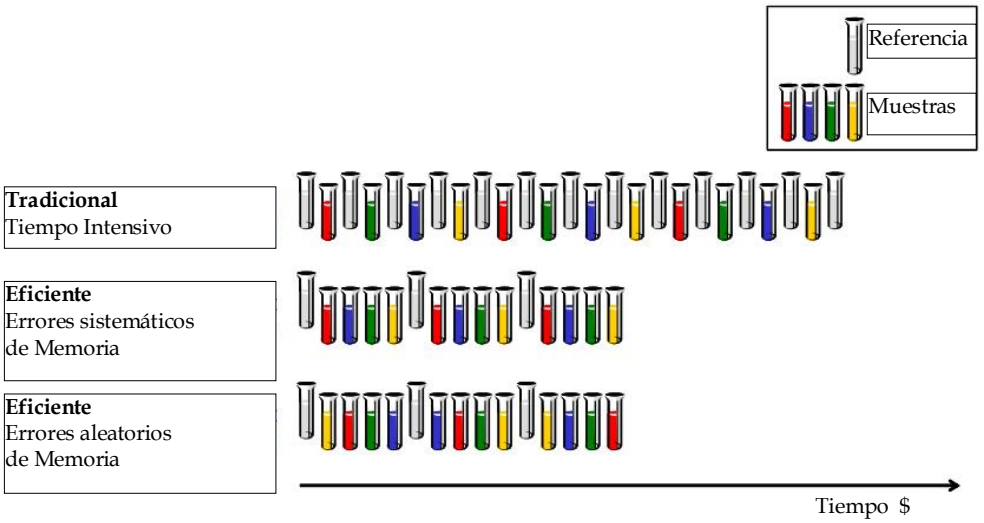


Figura 12. Técnicas de re-calibración para controlar la aparición de derivas (Gutierrez-Osuna, Signal processing methods for drift compensation. 2nd NOSE II Workshop, 2003).

1.6 TÉCNICAS DE PREPROCESADO USADAS EN LAS TÉCNICAS MULTIVARIADAS

Como parte previa a la **sección 1.7,** se expone a continuación el preprocesamiento que se aplica comúnmente a los datos, que corresponde a la etapa previa requerida en la implementación de las técnicas multivariadas empleadas en la corrección de derivas, destacándose la remoción de datos anómalos (*outliers*) y el escalado y normalización de la información que se va a procesar, a fin de preparar el vector de características para el futuro procesamiento de la información (Gutierrez-Osuna, 2002).

En consecuencia de lo anterior, antes de exponer las técnicas de análisis multivariado, se deben estudiar y conocer las etapas de preprocesamiento necesarias antes de aplicar las técnicas multivariadas aquí mencionadas. La remoción de datos anómalos y el escalado y la normalización contribuyen a la obtención de buenos resultados al aplicar PCA y CPCA, por lo tanto se exponen de forma breve a continuación.

1.6.1 Remoción de datos anómalos

El tipo de señales que manejan los sensores de gases permiten la pronta detección de datos anómalos realizando una inspección visual de la representación gráfica de los datos a procesar. Sin embargo, debido a la gran cantidad de medidas empleadas en la base de datos, se hace uso de la función `pcout` del paquete `mvoutlier` en R, para la identificación de datos anómalos en este caso específico de alta dimensionalidad.

1.6.2 Escalado y normalización

Los métodos de escalado y normalización más comunes son los siguientes: (I) Media centrada: la media se resta de cada variable, (II) Auto-escalado: cada variable primero se centra y luego se divide por su desviación estándar, (III) Normalización: las variables se dividen por la raíz cuadrada de la suma de los cuadrados de las variables; (IV) Suma de fila constante: cada variable se divide por la suma de todas las variables en cada muestra; (V) Variable de normalización: las variables se normalizan con respecto a una sola variable, (VI) Transformación de rango: el valor mínimo de una variable se establece en 0, el valor máximo a 1, y todos los valores intermedios se encuentran a lo largo de un rango lineal entre 0 y 1 (Berrueta, Alonso-Salces, & Héberger , 2007).

1.7 TÉCNICAS DE ANÁLISIS ESTADÍSTICO MULTIVARIADO EN LA CORRECCIÓN DE DERIVAS

Posterior al requerido pre-procesado de los datos con las técnicas presentadas en la **sección 1.6,** se desarrolla en este capítulo la presentación de las técnicas de análisis multivariado más desatacadas en el enfoque escogido en esta investigación para la mitigación de las derivas. Las técnicas PCA y CPCA, cada una de ellas combinada con la técnica de corrección de componentes de la cual se hablará en la **sección 1.7.3** son destacadas en el desarrollo metodológico aquí empleado, por lo tanto esta sección se dedica a exponer la fundamentación teórica de cada una de estas técnicas.

Dado que trabajar con conjuntos numerosos de muestras ocasiona patrones de alta dimensionalidad, un bloque explícito o implícito de extracción de características es necesario. El objetivo de la extracción de características es obtener características informativas de transformaciones matemáticas en los tiempos de respuesta multivariados del arreglo de sensores químicos usados para el reconocimiento de volátiles. Estas transformaciones reducen la dimensionalidad del espacio de entrada y tienen la finalidad de mantener la información relacionada con el objetivo del problema (Marco & Gutierrez-Galvez, 2012).

Dos de las técnicas que contribuyen a la reducción de la dimensionalidad son usadas en este trabajo, sin embargo su utilidad no consiste en la reducción de dimensionalidad de la que se habla en el párrafo anterior, sino en la extracción de las características o componentes principales y componentes principales comunes, de acuerdo a si se trata de análisis PCA o del análisis CPCA respectivamente; estas dos técnicas se combinan con la corrección de componentes para lograr extraer las componentes principales de la deriva y removerlas de los datos con el propósito de mejorar las señales de respuesta obtenidas de los sensores y mitigar las derivas. En consecuencia de lo anterior, en las secciones siguientes se detallan estas técnicas y en la sección final se incluye la corrección de componentes, técnica necesaria para lograr la corrección de derivas por medio del análisis mltivariado aquí empleado.

1.7.1 Análisis de Componentes Principales

Estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y posteriormente fueron estudiadas por Hotelling en los años 30 del siglo XX. Sin embargo, hasta la aparición de los equipos de cómputo no se empezaron a popularizar. Para estudiar las relaciones que se presentan entre **p** variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorrelacionadas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales. (Marín Diazaraque, 2013).

Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra. De modo ideal, se buscan **m < p** variables que sean combinaciones lineales de las **p** originales y que sean no correlacionadas, recogiendo la mayor parte de la información o variabilidad de los datos.

Si las variables originales son no correlacionadas desde el inicio, entonces no tiene sentido realizar un análisis de componentes principales. El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes.

En el análisis de componentes principales se consideran una serie de variables (x_1, x_2, ..., x_p) sobre un grupo de objetos o individuos y se trata de calcular, a partir de ellas, un nuevo conjunto de variables y_1, y_2, ..., y_p, no correlacionadas entre sí, y cuyas varianzas vayan decreciendo progresivamente.

El primer componente se calcula eligiendo a_1 de modo que y_1 tenga la mayor varianza posible. El segundo componente principal se calcula obteniendo a_2 de modo que la variable obtenida, y_2 no esté correlacionada con y_1. Del mismo modo se eligen y_1, y_2, $\cdots$, y_p, no correlacionadas entre sí, de manera que las variables aleatorias obtenidas vayan teniendo cada vez menor varianza.

Todos los componentes y (en total p), se pueden expresar como el producto de una matriz formada por los autovectores, multiplicada por el vector x que contiene las variables originales $x_1, \ldots, x_p$.

$$y = Ax \qquad (6)$$

Donde

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}, x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad (7)$$

como

$$Var\ (y_1) = \lambda_1$$
$$Var\ (y_2) = \lambda_2$$
$$\cdots$$
$$Var\ (y_p) = \lambda_p$$

la matriz de covarianzas de y será

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_p \end{pmatrix} \qquad (8)$$

porque $y_1, y_2, \cdots, y_p$ se han construido como variables no correlacionadas, se tiene que

$$\Lambda = Var(Y) = A' Var(X) A = A' \Sigma A \qquad (9)$$

o bien

$$\Sigma = A \Lambda A' \qquad (10)$$

dado que A es una matiz ortogonal (porque $a'_i . a_i = 1$ para todas sus columnas) por lo que $AA' = I$.

En la práctica, al tener en principio $\boldsymbol{p}$ variables, se obtiene un número mucho menor de componentes que almacena un porcentaje amplio de la variabilidad total $\sum_{i=1}^{p} Var(x_1)$. En general, no se suele tomar más de tres componentes principales, a ser posible, para poder representarlos luego en las gráficas (Marín Diazaraque, 2013).

1.7.2 Análisis de Componentes Principales Comunes CPCA

Este análisis puede ser visto como una generalización de PCA para k grupos de clases. Bajo la hipótesis H_c de los componentes principales comunes, existe una matriz ortogonal V, tal que las covarianzas de las matrices Σ_i son las diagonales en el espacio de datos definido por V (Ziyatdinov, y otros, 2009).

$$H_c : L_C = V^T . \Sigma_i . V, \quad i = 1,2, \cdots, k \qquad (11))$$

El método CPCA fue propuesto por Flury en 1984, quien derivó la teoría de la máxima verosimilitud estimada de V y L_i. *CPCA* es también referido al método de Diagonalización Conjunta (DC) y es realizado como diagonalización simultánea de matrices Σ_i.

La solución exacta de CPCA existe si todas las matrices Σ_i conmutan. Cuando esto no ocurre, la aproximación del problema de DC se puede resolver mediante la optimización de diferentes criterios de diagonalidad. En este trabajo, se lleva a cabo la diagonalización ortogonal sobre la base de criterios mínimos cuadrados ponderados por medio de las rotaciones de Givens (Ziyatdinov, y otros, 2009).

La comparación entre PCA y CPCA puede dar una idea de la aplicación de diagonalización conjunta, que se explicará en la sección siguiente, para el problema de compensación de deriva. Por un lado, el PCA encuentra la dirección de máxima varianza en la información de separación de clases. El componente principal puede ser útil para la interpretación sólo en el caso trivial de una clase, que es el gas de referencia. Por otro lado, CPCA analiza la relación entre la clase y sus componentes principales incluida la dirección de la varianza común para todas las clases.

En otras palabras, el enfoque CPCA tiene una base matemática confiable para encontrar la varianza de la deriva de conformidad con la definición de la deriva a largo plazo. El enfoque PCA, el cual toma una clase de referencia funciona bien, sólo si el gas de referencia cumple con los requisitos para ser físicamente representativo. De lo contrario, los componentes principales pueden capturar no sólo la varianza deriva, sino también la varianza en los datos valiosos para el análisis de datos en la clasificación, como componentes orientados de concentración.

Una propiedad importante de CPCA es que la transformación de la matriz V es ortogonal (una formulación no - ortogonal del CPCA también existe). Por lo tanto, esto permite el uso de varios componentes principales de P extraídos como columnas de la matriz V, de la misma manera que para el PCA (Ziyatdinov, y otros, 2009).

<u>Diagonalización Conjunta</u>

La técnica de diagonalización conjunta es utilizada para diagonalizar simultáneamente una serie de $\boldsymbol{K}$ matrices simétricas cuadradas $\boldsymbol{C} = \{\boldsymbol{C_k}\}_{k=1}^{K}$, donde $\boldsymbol{C_k} \in \boldsymbol{\mathcal{R}^{NxN}}$. El esquema estándar de diagonalización conjunta encuentra una matriz común $\boldsymbol{W} \in \boldsymbol{\mathcal{R}^{NxN}}$ tal que $\boldsymbol{WC_kW^T}$ son matrices diagonales para todos los *k*. Hay varias aplicaciones en las cuales se debe emplear la técnica de diagonalización conjunta, tales como la separación a ciegas de señales (BSS), también en el análisis del patrón espacial común (CSPA) y en el análisis de componentes principales comunes (CPCA), está última de interés para este trabajo. Sin embargo, es conocido que para más de dos matrices puede no ser posible obtener la diagonalización conjunta exacta y se debe recurrir a aproximaciones de la diagonalización conjunta para tales aplicaciones. Además, en algunas aplicaciones las matrices $\boldsymbol{C_k}$ incluso pueden no ser simétricas. Para solucionar estos inconvenientes se han desarrollado técnicas que buscan generalizar el estándar de la diagonalización conjunta, en donde a cambio de buscar directamente la matriz $\boldsymbol{W}$, se busca inferir una matriz goblal $\boldsymbol{B} \in \boldsymbol{\mathcal{R}^{NxN}}$, la cual no debe ser necesariamente cuadrada y una serie de K matrices diagonales $\Lambda = \{\Lambda_k\}_{k=1}^{K}$ donde Λ_k corresponde a los valores propios para cada k. (Zhong & Girolami, 2012)

En este trabajo se utilizó la función `djd` del paquete `JADE` en el software estadístico `R` para obtener la matriz de diagonalización conjunta de donde se extraen las componentes principales comunes usadas en la corrección de derivas.

1.7.3 Corrección de Componentes (CC)

La Corrección de Componentes (CC) originalmente propuesta por (Arthurson, y otros, 2000) ha sido el método más popular en la comunidad de olfato artificial. De hecho, este puede ser considerado como el punto de referencia para los métodos multivariados de corrección de derivas (Marco & Gutierrez-Galvez, 2012).

El principio básico de la Corrección de Componentes fue diseñado a partir del método de Corrección Ortogonal de Componentes (OSC). La corrección de componentes asume que la deriva tiene una dirección preferida en el espacio de medidas y no está aleatoriamente distribuida. Lo anterior es cierto por el efecto de envejecimiento que causan muchas de las derivas. Los sensores también han sido expuestos a algún gas, contenido de agua, cambios en la temperatura, entre otros. Lo cual confirma la suposición acerca de que los cambios en las respuestas ocasionadas por las derivas se mantienen principalmente en una dirección.

Si la respuesta de los sensores a cierto gas de referencia tiene deriva significante, las primeras componentes en un análisis PCA de solo este gas describirá la dirección de la deriva. Los coeficientes de direcciones de estas derivas están en el primer vector **p** de carga (loading). El espacio tri-dimensional es usado como ilustración, con la conjetura de que el espacio de alta dimensionalidad tiene propiedades similares. Al remover esta dirección de deriva de la matriz de medidas, información irrelevante es removida; por lo tanto, la estabilidad es incrementada y el tiempo de vida de los modelos de regresión se prolonga. Se supone que la relación entre sensores similares es lineal y que esta relación es la misma para el gas de referencia y para el gas de prueba.

Para las medidas de referencia un vector propio **p** es calculado por análisis PCA. La dirección en el espacio de referencia capturada por el primer vector de `loadings` puede ser atribuida a la deriva calculada en el espacio de referencia. Cuando la matriz de sensores es sometida a algún gas, esta debería dar siempre la misma respuesta al mismo gas con excepción de algún ruido aleatorio que se pueda presentar. La dirección de la deriva es asumida también como la misma para las muestras. Proyectando las muestras $\boldsymbol{X}$ en el primer `loading` **p** dado, se obtienen los los valores **t** para las muestras:

$$\boldsymbol{t} = \boldsymbol{Xp} \qquad (12)$$

La corrección de derivas es expresada como la resta de la expresión bilineal $\boldsymbol{tp}^T$ de los datos originales.

$$\boldsymbol{X}_{corregida} = \boldsymbol{X} - \boldsymbol{t}\boldsymbol{p}^{T} \qquad (13)$$

Es posible continuar con la siguiente componente y luego la otra y tantas como se desee. Una componente podría usualmente ser suficiente si los efectos de la deriva se consideran causados solo por el envejecimiento del sensor; sin embargo, si la deriva es no lineal, por ejemplo, es causada tanto por efectos de envejecimiento como por cambios en la capa química de los sensores, esto podría motivar a substraer más de una componente de $\boldsymbol{X}$.

Es importante darse cuenta que la dirección de deriva que es removida del conjunto de datos es una aproximación lineal de la dirección de la deriva. Esto significa que todas las otras direcciones se mantienen y las variaciones importantes de la varianza que separan los diferentes clúster y concentraciones se mantienen en el conjunto de datos de $\boldsymbol{X}_{corregida}$, a menos que la información se encuentre en la misma dirección de la deriva, lo cual raramente ocurre.

Tres factores importantes que influyen en el resultado de la corrección de componentes son el escalado, la normalización y la remoción de datos anómalos. Dado que los parámetros de los modelos de componentes principales dependen del escalado o normalización, estos se deben aplicar tanto a las muestras del conjunto de referencia y a las muestras de prueba. Otro factor que influye son los datos anómalos, el `loading` en PCA para la calibración de muestras es seriamente influenciado por los anómalos, dado que la dirección podría ser inclinada por la presencia de este tipo de datos. En consecuencia, es importante remover los anómalos antes de calcular los `loading`. Un vector de `loading` incorrecto reduce el desempeño de la corrección de componentes (Arthurson, y otros, 2000).

2. DISEÑO EXPERIMENTAL

2.1 BASES DE DATOS

Se utilizaron dos bases de datos, una de ellas con datos sintéticos denominada `chemosensors` y la segunda corresponde a una base de datos con medidas experimentales de la Universidad California. En las secciones siguientes se detallan las características de las bases de datos empleadas.

2.1.1 Datos sintéticos de Chemosensors

Se compone de un arreglo de sensores virtuales que se generan mediante un paquete en lenguaje `R`, llamado `chemosensors`, el cual es de libre acceso. El flujo de trabajo para generar los datos sintéticos a partir de `chemosensors` consiste en crear el escenario de trabajo, luego se establecen los parámetros del arreglo de sensores y finalmente se generan la matriz de características y el vector de clases o etiquetas. Esta base de datos toma como referencia las medidas experimentales de UNIMAN perteneciente a la Universidad de Manchester en el Reino Unido, que contiene 3925 muestras tomadas en un periodo de 10 meses con 17 sensores y usando tres analitos, amoniaco, ácido propanoico y n-buthanol a diferentes niveles de concentración. (Ziyatdinov & Perera-Lluna, 2014)

Esta base de datos sintéticos fue creada por el Centro de Investigación en Ingeniería Biomédica de la Universidad Politécnica de Cataluña (España), es útil para la evaluación comparativa de reconocimiento estadístico de patrones en el olfato artificial, pertenece al proyecto Neurochem y contiene sensores químicos virtuales. El código es libre y se puede acceder en línea a través de http://neurochem.sisbio.recerca.upc.edu/, además contiene 1020 sensores con 21900 muestras, se utilizan los datos sintéticos dada la importancia de poder hacer el análisis de las derivas conociendo previamente las componentes de deriva añadidas a los datos con el propósito de realizar el análisis exploratorio de la afectación causada por las derivas en los sistemas de reconocimiento de olores.

La **Tabla 1** presenta la lista de parámetros que se emplean en `chemosensors;` en este trabajo se resaltan el uso de el parámetro `dsd` y `ndcomp` para analizar el efecto de la deriva en los datos.

Parámetro	Valor por defecto	Rango de Valores	Descripción corta
num	1:2	1,2,…….17	Tipo de sensor
nsensors	2	1,2,…..	Número de sensores
gnames	3	1,2,3	Número de gases
concUnits	porcentaje		Unidades de concentración
alpha	2,25	> 0	No linealidad del sensor
beta	2	≥0	Diversidad de sensores
csd	0,1	≥0	Concentración de ruido
ssd	0,1	≥0	Ruido del sensor
dsd	0,1	≥0	Ruido de deriva
ndcomp	1	1,2,3	Número de componentes de deriva
ndvar	0,86	(0,1]	Importancia de los componentes de deriva
tunit	1	1,2…..	Longitud del pulso del gas

Tabla 1. Parámetros necesarios para generar datos sintéticos en `chemosensors`.

2.1.2 Base de datos de la Universidad de California

La segunda base de datos utilizada en este trabajo de investigación es una base de datos con medidas experimentales que utilizan sensores químicos para la medición de los volátiles. En la **Tabla 2** se muestra una descripción general de la base de datos mencionada.

NOMBRE	ACCESO	TIPO	TAMAÑO	NÚMERO SENSORES	UBICACIÓN/ CONTACTO
Gas Sensor Array DataSet	Libre	Real	10 archivos (Batch) con un total de 13910 mediciones	16	University of California San Diego. San Diego, California, USA (University of California, 2012). Alexander Vergara (vergara@ ucsd.edu) Ramon Huerta (rhuerta@ ucsd.edu)

Tabla 2. Descripción general de la bases de datos de la Universidad de California.

Esta base de datos fue publicada por la Universidad de California (San Diego, EE.UU) y se usó en este trabajo de investigación por las características que presenta, tales como el tiempo en el que fueron recolectadas las mediciones, que abarcan un periodo de tres años y donde las condiciones experimentales fueron controladas como se detalla en (Vergara, y otros, 2012). Lo anterior, es una condición importante para analizar las derivas, por cuanto se garantiza que los problemas de ruidos y afectaciones dadas por efectos de la temperatura o contaminación del aire que transporta los volátiles se han reducido de tal manera que las

falsedades presentadas en los datos se centren principalmente en el problema de las derivas y de los mismos sensores químicos.

La base de datos de la Universidad de California contiene 13910 registros de seis diferentes gases o analitos (amoniaco, acetaldehído, acetona, etileno, etanol, tolueno) con diferentes concentraciones, recolectados durante 36 meses usando una matriz de 16 sensores. La información en la base de datos contiene solo la clase de gas, sin embargo, no está explícita la concentración del gas, por cuanto el uso de esta base de datos se enfoca en tareas únicamente de clasificación y no se realizan tareas de regresión por no contar con la cuantificación de los gases. Así, se clasifican seis diferentes gases independiente de su concentración y nombrados, por simplicidad en este trabajo, con las letras de la A a la F, cada una de ellas haciendo referencia a los gases mencionados anteriormente en éste párrafo.

La distribución exacta del número de medidas por mes es mostrada en la **Tabla 3,** nótese que algunos de los meses no contienen medidas y además, que la base de datos está organizada en 10 lotes o batch, distribuidos con un número no equitativo medidas por gas pero distribuidos lo mas uniformemente posible para tareas de clasificación, tal como lo referencia (Vergara, y otros, 2012).

Lote (Batch)	Mes	Número de muestras por gas						Total Mes	Total Batch
		Gas A	Gas B	Gas C	Gas D	Gas E	Gas F		
1	1	76	0	0	88	84	0	248	445
	2	7	30	70	10	6	74	197	
2	3	0	0	7	140	70	0	217	1244
	4	0	4	0	170	82	5	261	
	8	0	0	0	20	0	0	20	
	9	0	0	0	4	11	0	15	
	10	100	105	525	0	1	0	731	
3	11	0	0	0	146	360	0	506	1586
	12	0	192	0	334	0	0	526	
	13	216	48	275	10	5	0	554	
4	14	0	18	0	43	52	0	113	161
	15	12	12	12	0	12	0	48	
5	16	20	46	63	40	28	0	197	197
6	17	0	0	0	20	0	0	20	2300
	18	0	0	0	3	0	0	3	
	19	110	29	140	100	264	9	652	
	20	0	0	466	451	250	458	1625	
7	21	360	744	630	662	649	568	3613	3613
8	22	25	15	123	0	0	0	163	294
	23	15	18	20	30	30	18	131	
9	24	0	25	28	0	0	1	54	470
	30	100	50	50	55	61	100	416	
10	36	600	600	600	600	600	600	3600	3600

Tabla 3. Detalles de la base de datos suministrada por la Universidad de California.

Se destaca en la **Tabla 3**, que el último lote de mediciones contiene **3600** registros de los seis analitos, estas medidas se recogieron a propósito cinco meses después de no usar los sensores. Esta diferencia de 5 meses es muy importante para este estudio, no sólo porque nos permite validar el método sugerido en el conjunto de mediciones recolectados cinco meses más tarde, sino también porque es durante este período de tiempo en que los sensores fueron sometidos a grave contaminación tales como interferencias externas que fácilmente y de forma irreversible pueden quedar unidas a la capa de detección debido a la falta de la temperatura de funcionamiento. El equipo empleado y los procedimientos que se utilizaron para adquirir esta base de datos se encuentran detallados en (Vergara, y otros, 2012).

Los autores de la base de datos de San Diego proporcionan los resultados obtenidos después de realizado el proceso de extracción de características y procesado de los datos, por tanto cabe aclarar, que la información contenida en esta base de datos, no corresponde a la serie de tiempos del proceso de medida en cada sensor, sino sencillamente corresponde a un conjunto de 8 características que fueron seleccionadas a partir de la respuesta en el tiempo entregada por cada uno de los 16 sensores y para cada muestra de gas analizada en los tiempos que corresponden a inyección del gas y fase de limpieza o recuperación.

De acuerdo a lo anterior**,** San Diego contiene por cada medida en cada uno de los sensores, 2 características de estado estable, una de ellas es considerada el "estándar de oro" para la extracción de características de sensores químicos; esta es definida como la diferencia del máximo cambio de resistencia y la línea base (ecuación 14), la segunda es la versión normalizada de la anterior, expresada por el radio de la máxima resistencia y los valores de la línea base (ecuación 15).

$$\Delta R = max_k r[k] - min_k r[k] \tag{14}$$

$$\|\Delta R\| = \frac{max_k r[k] - min_k r[k]}{min_k r[k]} \tag{15}$$

Donde, $\boldsymbol{r}[\boldsymbol{k}]$ es el perfil temporal de la resistencia del sensor, $\boldsymbol{k}$ es el tiempo discreto indexado en el intervalo de grabación de $[\boldsymbol{0}, \boldsymbol{T}]$ cuando el vapor químico está presente en la cámara de sensores.

Por otra parte, del estado transitorio toman las fases de adsorción y desorción del gas en la respuesta del sensor, extrayendo un conjunto de seis características adicionales que reflejan la dinámica del sensor en la parte de transición creciente - decreciente de la respuesta del sensor durante todo el procedimiento de medición en condiciones controladas. Utilizan la media móvil exponencial $\boldsymbol{(ema_\alpha)}$**,** para estimar el máximo o mínimo valor de la porción de incremento o de decaimiento de la respuesta del sensor. La media móvil exponencial $\boldsymbol{(ema_\alpha)}$**,** es calculada por,

$$y[k] = (1-\alpha)y[k-1] + \alpha(r[k] - r[k-1]) \qquad (16)$$

Donde, $\boldsymbol{k = 1, 2, \ldots, T}$, $\boldsymbol{y[0]}$ es una condición inicial. Asimismo, se toman 3 diferentes valores de $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha = 0.1, \alpha = 0.01, \alpha = 0.001}$), completando de esta forma las 6 características, 3 de la fase transitoria de ascenso o inyección del gas y 3 de la fase de descenso o recuperación del sensor, sumada a las ya mencionadas 2 características de estado estable, para un total de 8 características por cada serie de tiempos, tal como se detalla en la **Tabla 4** y se observa más claramente en la **Figura 13**.

Características de estado estable	**Características de estado transitorio**	
	Porción de ascenso	Porción de descenso
$\boldsymbol{\Delta R}$	$max_k\ ema_{\alpha=0.001}\ (r[k])$	$min_k\ ema_{\alpha=0.001}\ (r[k])$
	$max_k\ ema_{\alpha=0.01}\ (r[k])$	$min_k\ ema_{\alpha=0.01}\ (r[k])$
$\boldsymbol{\|\Delta R\|}$	$max_k\ ema_{\alpha=0.1}\ (r[k])$	$min_k\ ema_{\alpha=0.1}\ (r[k])$

Tabla 4. Características que proporciona la base de datos San Diego para cada una de las medidas tomadas por cada uno de los 16 sensores (Vergara, y otros, 2012).

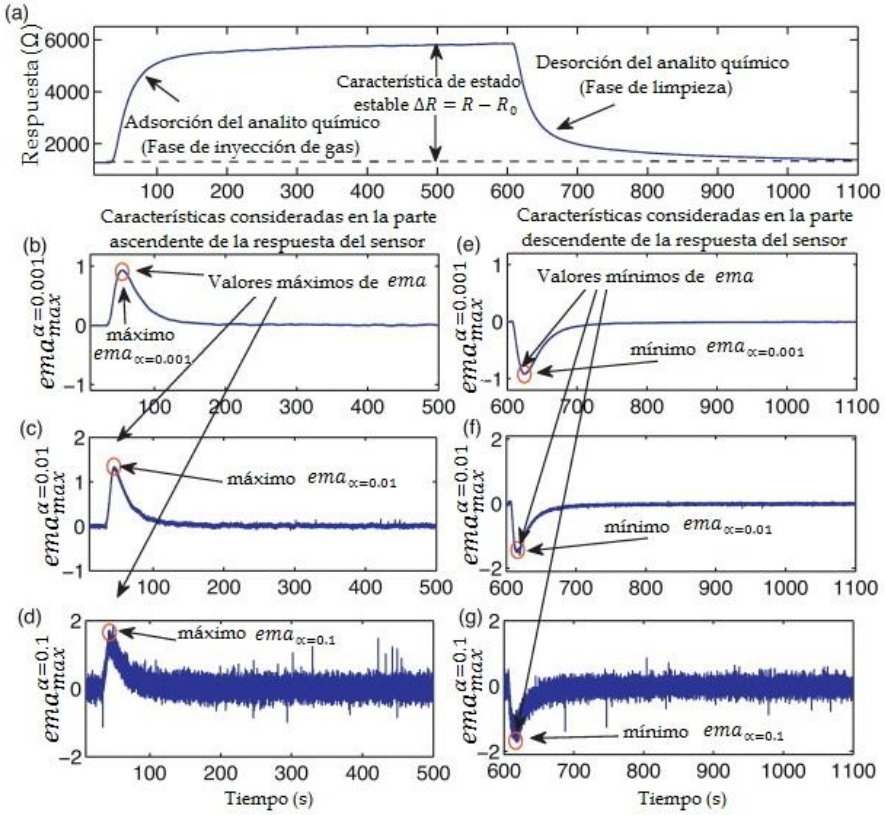


Figura 13. Respuesta típica en el tiempo de un sensor químico. (a) Respuesta del sensor a 30 ppmv de Aceltadehido. (b) – (d) Media móvil exponencial de la porción de risado máximo o inyección del gas. (e) – (g) Media móvil exponencial de la porción de decaimiento mínimo o fase de limpieza (Vergara, y otros, 2012).

Finalmente, se tiene una matriz de características de 128 columnas por 13910 filas, en este caso las filas corresponden a las medidas en orden cronológico realizadas con la matriz de 16 sensores y las columnas reflejan para cada uno de los 16 sensores las 8 características extraídas por los autores de la base de datos de San Diego.

Para efectos del análisis de las derivas, en este trabajo se realiza el tratamiento de los datos aprovechando la estructura de lotes que tienen planteada los autores. Siguiendo esta metodología, se realizó el entrenamiento y la validación de los datos por lotes, tomando como conjunto de entrenamiento el lote 1 y validando con los lotes subsiguientes (ver **Figura 14)**, luego se entrena con los lotes 1 y 2 y se valida uno a uno con los restantes; estos dos experimentos son importantes con el fin de determinar la porción de datos que deben ser usados para el entrenamiento o referencia que tengan un componente de deriva adecuado para ser asumido en el modelo de corrección de la deriva.

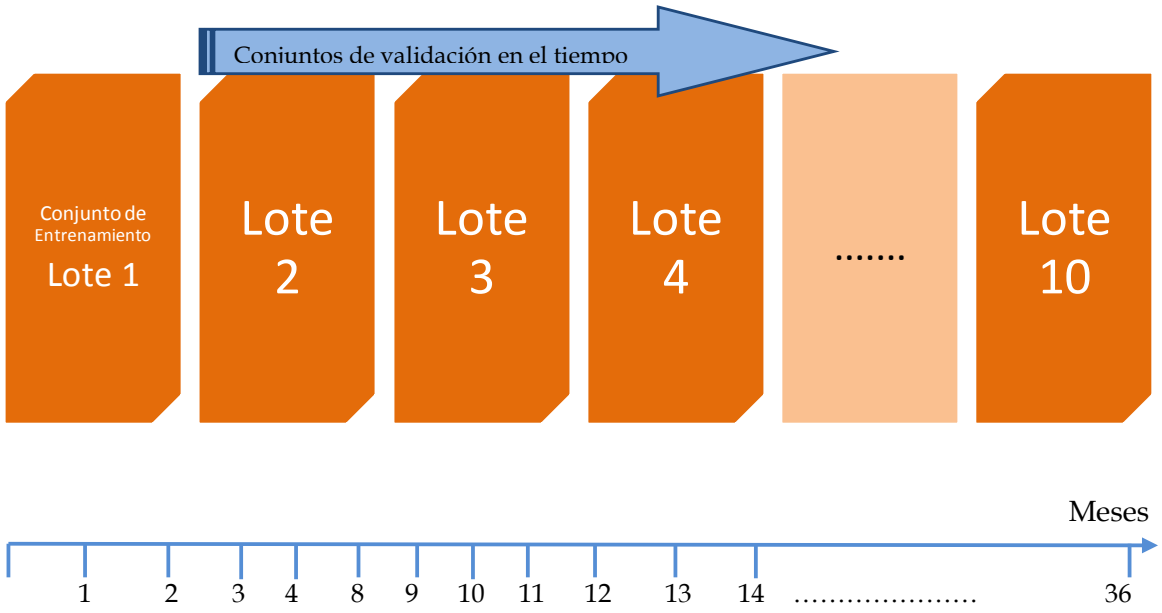


Figura 14. Organización de la base proporcionada por la Universidad de California por lotes, para el caso en el que se entrena con el lote 1 y se valida con los lotes subsiguientes.

2.2 METODOLOGÍA PROPUESTA

En este trabajo de investigación se utilizaron una serie de técnicas y procedimientos que permitieron alcanzar los objetivos propuestos. Se desarrollaron 3 etapas en concordancia con los objetivos específicos, las cuales se describen a continuación. (Ver **Figura 15**)

2.2.1 Caracterización de los datos

Para desarrollar el primer objetivo, se utilizaron los datos generados con `chemosensors` y la base de datos proporcionada por la Universidad de California; la primera de ellas corresponde a datos sintéticos y la segunda fue recolectada en un periodo de 36 meses a partir de las mediciones de seis analitos a diferentes concentraciones. Esta última contiene una gran cantidad de medidas en un extenso periodo de tiempo y además, se garantizan condiciones controladas en el proceso de medición (Vergara, y otros, 2012).

Figura 15. Etapas generales de la metodología.

La población objeto de estudio en esta etapa la constituyen las bases de datos disponibles ya mencionadas en los párrafos anteriores y por otra parte, las técnicas de análisis estadístico multivariante denominadas Análisis de Componentes Principales (PCA) y Análisis de Componentes Principales Comunes (CPCA). Se tomó como referente la revisión de otros trabajos propuestos con técnicas de análisis multivariado (Marco & Gutierrez-Galvez, 2012), (Ziyatdinov, y otros, 2009), así como el estudio y análisis de las bases de datos con las que se cuenta. Las variables a tener en cuenta fueron algunos parámetros de medición (tiempo, número de muestras, condiciones controladas, número de clases, entre otros), determinados a partir de información disponible; asimismo, se tuvieron en cuenta las variables de tipo estadístico resultantes de la utilización de las técnicas multivariantes, tales como la varianza, la covarianza y los resultantes vectores y valores propios. Los corolarios de esta etapa hacen parte del pre-procesamiento de la información y se recopilan en las variables y funciones utilizadas en el software de procesamiento, empleando como herramienta estadística la programación y el uso de funciones y librerías de `R`.

2.2.2 Extracción de Componentes de Deriva

En el desarrollo del segundo objetivo, se usó la técnica de análisis multivariado CPCA con corrección de componentes para lograr la mitigación de las derivas y a partir de ella determinar el espacio de representación adecuado en el modelo de corrección de derivas; para ello se realizó un test estadístico basado en la remoción de las componentes que acumulen los mayores porcentajes de varianza y así determinar el mayor número de componentes que se pueden extraer sin afectar la información importante, con el fin de mejorar la respuesta en la clasificación de compuestos volátiles. Se utilizó como herramienta el **criterio de separabilidad** que calcula la relación entre la dispersión de los datos entre-clases y la dispersión intra-clases.

La población objeto de estudio en este caso corresponde al método de selección de características CC-CPCA (Corrección de Componentes por medio de Análisis de Componentes Principales Comunes), se utilizaron herramientas y aplicaciones de libre acceso, disponibilidad y licenciamiento en el software estadístico R, tales como `mvoutlier`, `robCompositions`, entre otros. Esta parte del trabajo se apoyó en la revisión exhaustiva del estado del arte respecto a las técnicas y métodos más apropiados para realizar la mitigación de las derivas a partir de técnicas de procesado. Además, se tuvo en cuenta la remoción de datos anómalos y el centrado y escalado de los datos.

Las variables que se tuvieron en cuenta fueron la confiabilidad (precisión en las respuestas), la repetitividad (precisión bajo un conjunto de condiciones) y exactitud (valor cercano al real). Se recopilaron los resultados de esta etapa usando tablas de registro de experimentos y gráficas comparativas de las respuestas del comportamiento de los datos al extraer las componentes principales que representan las derivas en los datos, con respecto a los datos originales sin tratamiento de derivas.

2.2.3 Validación

Para lograr el tercer objetivo se empleó el método de los k-vecinos (k-NN) y se hizo uso de la librería de R denominada `caret package` para aplicar las funciones de sintonización del parámetro `k` en el entrenamiento de los datos y la obtención del modelo que se aplicará a los datos de validación. La validación de resultados se lleva a cabo con el fin de determinar la exactitud del modelo propuesto frente al modelo de (Ziyatdinov, Marco, Chaudry, Persaud, Caminal, & Perera, 2010).

Se determinaron los porcentajes de aciertos obtenidos en el proceso de clasificación de diferentes analitos para concluir acerca de la exactitud y repetibilidad del sistema de olfato artificial; se usaron muestras de volátiles en el orden cronológico preestablecido en la recolección de los datos, para determinar el comportamiento del sistema afectado por la deriva en el tiempo. Asimismo, se compararon los resultados obtenidos frente al CC-CPCA (Ziyatdinov, y otros, 2009), en el que se empleó solo una componente principal y así se

concluye acerca de las ventajas de emplear un mayor número de componentes en este problema de corrección de derivas.

Se recopilaron los resultados de esta etapa usando tablas de registro de experimentos, así como graficas donde se comparan los porcentajes de clasificación para cada uno de los conjuntos de datos a lo largo del tiempo. Adicionalmente, se realizó el registro de datos estadísticos de las variables analizadas.

2.3 CONSTRUCCIÓN DEL MÓDULO DE TRABAJO EN R.

La herramienta de programación empleada es la herramienta de análisis estadístico R, por medio de la cual, se diseñó un paquete de software para el tratamiento de la información, que se ha denominado driftout. Los elementos y componentes de este paquete, así como sus librerías se detallan en el **APENDICE A.**

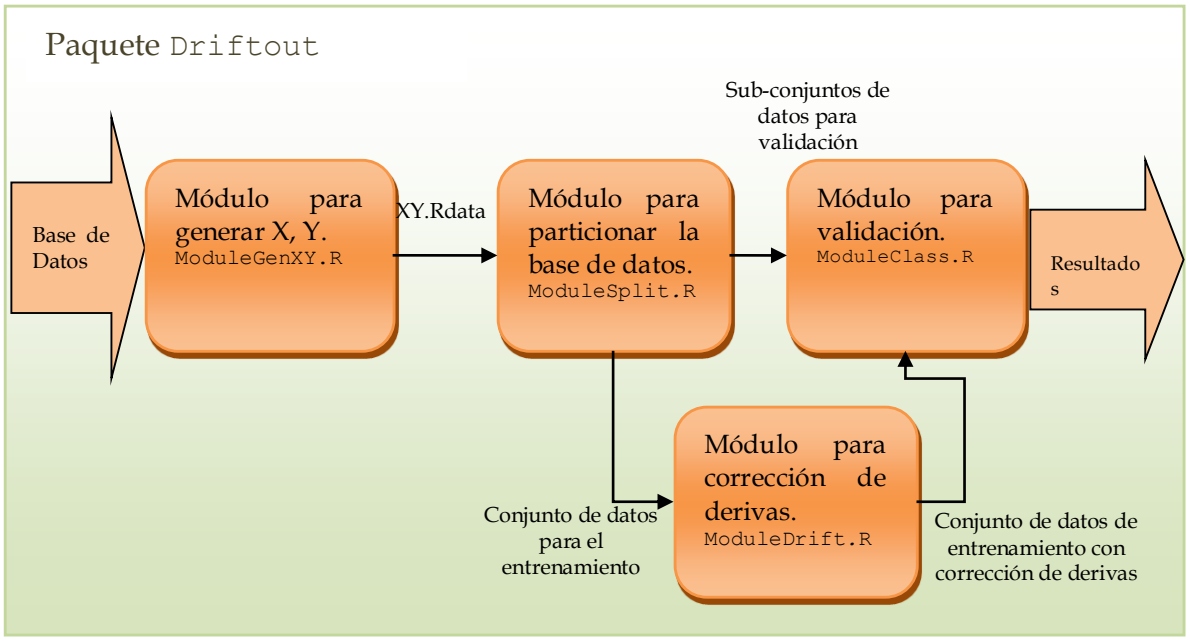


Figura 16. Secuencia de trabajo del paquete `driftout`.

Como se observa en la **Figura 16,** `driftout` se compone de los siguientes módulos: `moduleGenXY.R`, `moduleSplit.R`, `moduleDrift.R` y `moduleClass.R`. Para mayor claridad sobre el proceso llevado a cabo en cada módulo, se explica cada uno de ellos en las secciones siguientes.

2.3.1 Módulo para generar X, Y (`moduleGenXY.R`)

El primer paso en la secuencia de trabajo es tener los datos organizados en una matriz de características y un vector de clases o etiquetas, en el caso de la base de datos de la

Universidad de California, se hizo un tratamiento previo para lograr tener la información apropiadamente organizada al momento de procesarla, este proceso se describe a continuación.

La base de datos proporcionada por la Universidad de California, será llamada de ahora en adelante en este documento como "San Diego"; para la preparación de los datos de San Diego fue necesario organizar los lotes de la base de datos en una matriz de características y un vector de factores o clases. La matriz de características **X** de cada lote, contiene la información registrada desde los 16 sensores y, el vector columna de factores o clases denominado **Y**, especifica la clase de gas en cada medición. Como se mencionó en la **sección 2.1.2,** los gases se etiquetaron con las letras del alfabeto A, B, C, D, E y F, las cuales representan los seis analitos medidos sin tener en cuenta la concentración. En consecuencia, el primer paso correspondió a la lectura de la base de datos que se encuentra organizada en 10 lotes, es decir, contiene 10 ficheros de extensión `.dat`, organizados tal como se explicó en la **Tabla 2.** En `R`, la función empleada para leer este tipo de ficheros es la función `read.matrix.csr` del paquete `e1071`, usada para leer y escribir formato de datos dispersos o tipo `sparse`.

Como se observa en la **Figura 17,** la salida de `ModuleGenXY.R`, entrega para *San Diego* el fichero llamado `SanDiegoBatchXY.RData` que contiene la base de datos fragmentada por lotes; el Lote 1 contiene una matriz **X** con igual número de registros del lote 1 original con 128 columnas que corresponden al número de características extraídas de la serie de tiempos de la respuesta de cada sensor ubicado en la cámara de medidas y su correspondiente vector de clases; el mismo tratamiento se realizó en los lotes restantes.

Se hace salvedad que en la lectura y preprocesado de los datos se encuentra que 120 registros del lote 10 generan error en su tratamiento por contener datos que no especifican la clase, por lo tanto estos datos fueron descartados en el procesado de la información. Los autores de la base de datos de San Diego, han actualizado su sitio web (University of California, 2012), corrigiendo los datos corruptos, tal como lo citan: "`Batch10.dat` fue actualizado el 10/14/2013 para corregir algunos valores dañados en las últimas 120 filas del archivo." Para efectos de este estudio, se trabajó suprimiendo los registros dañados de la base de datos, dado que la corrección por parte de los autores fue hecha posteriormente al desarrollo del diseño experimental de este trabajo, por lo tanto, a continuación en este documento se hablará de un total de 13790 registros de San Diego, los 120 restantes fueron suprimidos.

Se aclara que aunque en este estudio se realizó el análisis de San Diego por lotes aprovechando la estructura preestablecida por sus autores, el método de ventana deslizante se debe utilizar en otras bases de datos que no dispongan de esta segmentación previamente establecida, tal como se hizo con los datos generados en `chemosensors`.

Para otras bases de datos se genera una matriz completa de características y un vector de clases, tal es el caso de la base de datos `chemosensors`. Los datos organizados de esta última forma fueron usados para obtener las medidas que conforman el grupo de

entrenamiento y efectuar una fragmentación equitativa en el subconjunto de validación, proceso que se realiza en el módulo llamado `moduleSplit.R`.

Se aplica una técnica para la remoción de datos anómalos mediante el paquete *rmoutliers* en la base de datos de *San Diego*. Este tratamiento no se aplica a *chemosensors* porque se observa que ésta no lo requiere.

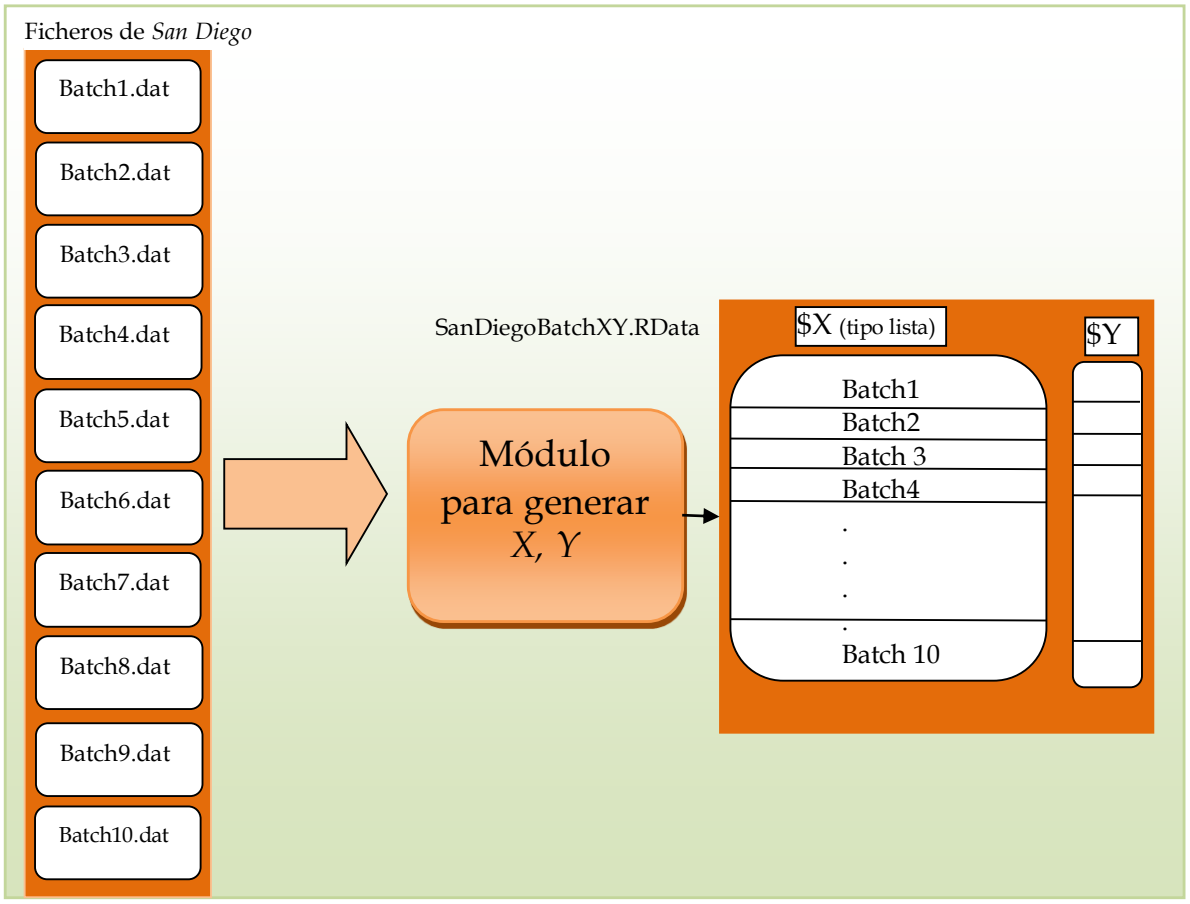


Figura 17. Secuencia de trabajo del módulo `ModuleGenXY.R`, tomando a San Diego como entrada.

2.3.2 Módulo para fraccionar los datos (`moduleSplit.R`).

El módulo llamado `moduleSplit.R`, se encarga de particionar los datos que provienen del módulo anterior, considerando que los datos vienen en un bloque o matriz completa de características y un vector de clases, tal como es el caso de `chemosensors`. Los datos se dividen en un subconjunto de entrenamiento y un subconjunto de validación, éste último a su vez se particiona en grupos iguales de **n** datos cada uno.

En la **Figura 18**, se visualiza el método de la ventana deslizante que corresponde a realizar la validación por subcojuntos de datos que se alejen cada vez más del conjunto que se toma para el entrenamiento, de esta manera se puede concluir qué tanto efecto tienen las derivas a medida que se utilizan mediciones que estén cada vez más lejanas en el tiempo del grupo que se ha tomado como referente para entrenar el clasificador. Esta metodología es tomada de (Ziyatdinov, y otros, 2009) y corresponde al esquema de validación especialmente ideado para probar algoritmos de deriva, propuesto por (Gutierrez-Osuna, 2000). Es necesario además, generar experimentos y diferentes pruebas que consisten en variar la cantidad de medidas para el conjunto de entrenamiento y lograr determinar cuál de ellos es el conjunto de datos más apropiado que represente componentes significativas de deriva, por lo tanto, en esta misma figura se observan las diferentes particiones de subconjuntos de entrenamiento y de **n** conjuntos de validación, separadas por experimentos.

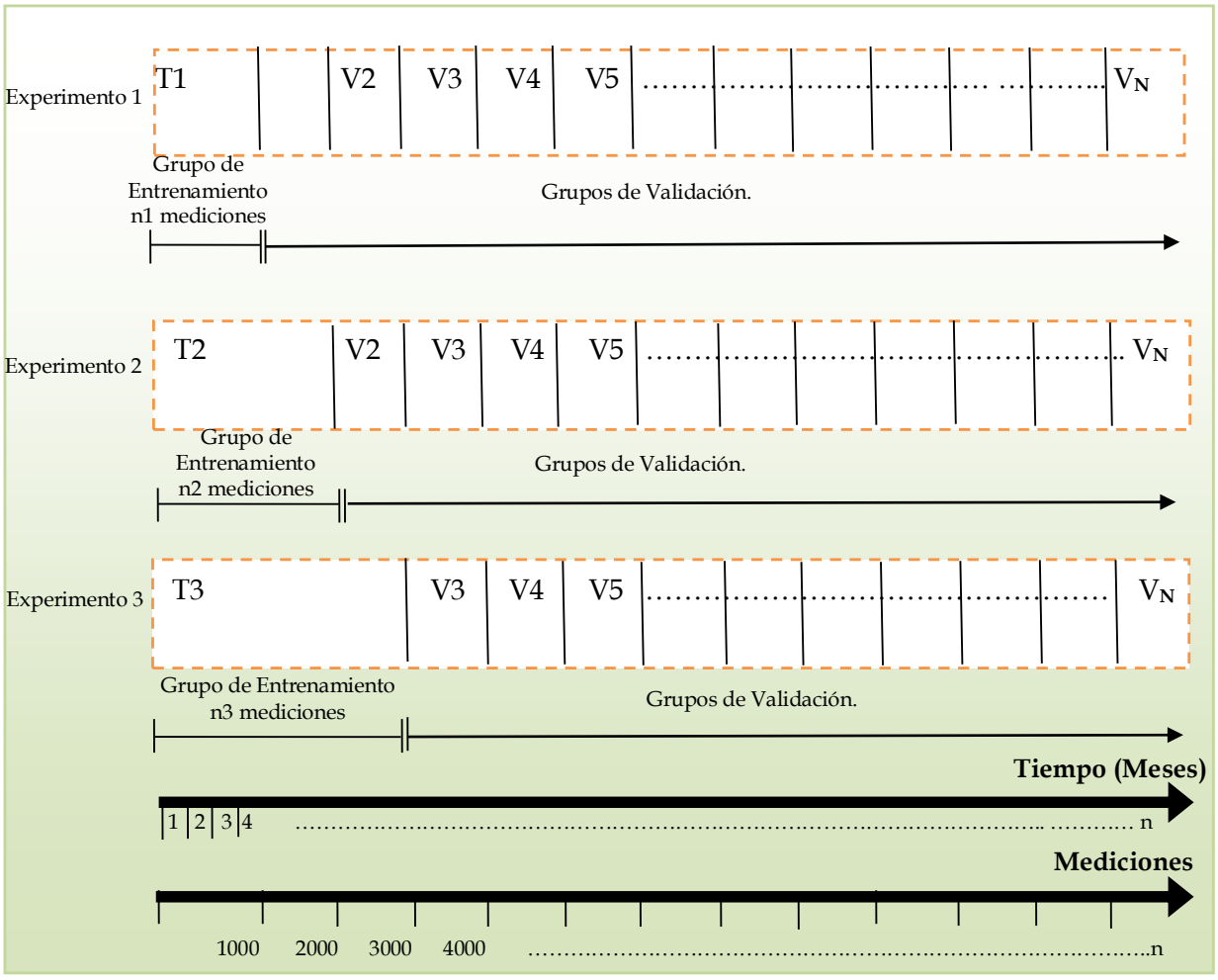


Figura 18. Partición de los datos, aplicando la ventana deslizante en el tiempo.

La ventana deslizante en los grupos de validación permite realizar una comparación de la influencia de las derivas en el tiempo y asimismo, poder determinar qué cantidad de mediciones en el entrenamiento se ajusta más a los resultados esperados, es decir, tiene incorporado componentes significativos de deriva. Luego de dividir los datos se almacenan en ficheros de tipo `RData`, uno para cada partición o experimento.

En el caso de los datos organizados por lotes, se selecciona el grupo de entrenamiento probando con el primer lote y validando con los *lotes* subsiguientes; en la siguiente prueba el entrenamiento lo conforman los lotes 1 y 2 y se valida uno a uno con los lotes restantes, tal como se explica en la **Figura 19.** En este caso la disposición de los datos respeta las agrupaciones originales de la base de datos, pero de igual forma se generan subgrupos de validación que en este caso serán los lotes de San Diego, por lo tanto, en este caso también se aplica una ventana deslizante, pero por en una disposición por lotes.

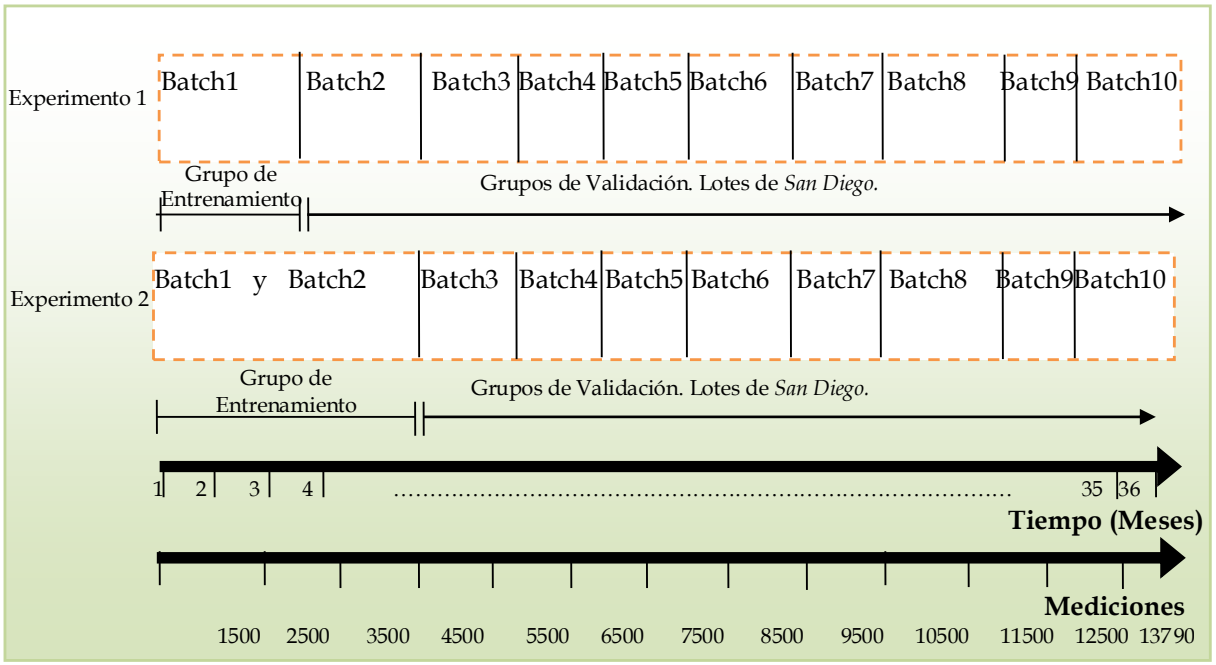


Figura 19. Partición de los datos de San Diego, aplicando la ventana deslizante en el tiempo para los grupos de validación.

2.3.3 Módulo para corrección de derivas (`moduleDrift.R`).

Este módulo contiene dos bloques independientes para el proceso de corrección de las derivas, el primero de ellos emplea el método de corrección de componentes por análisis de componentes principales (CC-PCA), que se basa en el método de corrección de las derivas propuesto por (Arthurson, y otros, 2000). En este bloque se toman las medidas de un gas de referencia al que se le aplica PCA y de allí se toma la primera componente para realizar la

corrección de componentes. En este caso, con el fin de determinar el mejor gas de referencia se prueba el método con las 6 clases de gases que contiene la base de datos San Diego.

Los datos de entrada a este bloque son los datos que provienen del módulo de particiones (`moduleSplit.R`), de allí, se toman los datos que conforman el conjunto de entrenamiento, que en el caso de San Diego corresponden en el primer experimento al lote 1 y en el segundo experimento involucra los lotes 1 y 2.

El segundo bloque realiza el tratamiento a la deriva a partir del método de corrección de componentes por análisis de componentes principales comunes CC-CPCA propuesto por (Ziyatdinov, y otros, 2009). En este caso, como entrada además de los datos de entrenamiento en cada experimento (lote 1 o lotes 1 y 2) se ingresará el número de componentes a remover en la corrección de la deriva.

Para encontrar las componentes principales comunes se realiza la diagonalización de la matriz **X** en los datos de entrenamiento, el resultado de la diagonalización es una matriz cuadrada de **128x128** dimensiones, que presenta las 128 componentes principales comunes resultantes, de la cual se extraen las componentes principales que se desean remover en los datos por el método de corrección de componentes y lograr la mitigación de las derivas. Las componentes que se desean remover para aplicar corrección de componentes de almacenan en la matriz **E1.**

La matriz **E1** generada en este módulo, que contiene las componentes principales comunes a remover es llevada al módulo de clasificación para aplicar de la misma manera con estas componentes extraídas la corrección de las derivas a los datos en los conjuntos de validación, que en este caso específico serán los lotes de San Diego que se han reservado para este propósito de acuerdo al experimento procesado, es decir si se entrena solo con el primer lote (experimento 1) o si se entrena con los dos primeros lotes (experimento 2).

Lo novedoso en este trabajo, consiste en remover no solo la primera componente principal común extraída de CCPCA, sino además, remover otras componentes principales comunes que acumulen varianza adicional y que contengan información relacionada con las componentes de deriva con el fin de mejorar la respuesta del sistema. Se utiliza para determinar el número de componentes el criterio de separabilidad que determina el número máximo de componentes a remover.

El método de preprocesado que se emplea en los dos casos anteriores es el de normalización de los datos (escalado y centrado), este proceso se realiza en los datos de entrenamiento antes de hacer la corrección de derivas mediante la resta de la media a los datos originales (**matriz** X), luego se divide este resultado por la desviación estándar de los datos contenidos en esta misma matriz **X** del entrenamiento; luego de corregir la deriva en los datos de entrenamiento, los datos fueron des-escalados y des-centrados nuevamente para ser llevados a su representación original. Los valores para escalar y centrar los datos de los conjuntos de validación se toman a partir de lo obtenido en el conjunto de entrenamiento, por lo tanto se almacenan en **Xscale** y **Xcenter** respectivamente, para ser direccionados al

módulo de clasificación y de esta manera poder aplicarlos a los **n** conjuntos de validación. De la misma forma, se resalta que la fase inicial en el preprocesado corresponde a realizar remoción de datos anómalos de la base de datos antes de realizar cualquier otro tratamiento a los datos.

Las salidas de este módulo son las componentes principales comunes extraídas del CPCA en el grupo de entrenamiento y asimismo, la información para el escalado y el centrado a usarse como modelo en los datos de validación, que corresponde a la desviación estándar y a la varianza de los datos de entrenamiento. De igual forma, este módulo entrega la componente principal o las componentes principales comunes dependiendo de si se ha usado el método CC-PCA o CC-CPCA respectivamente. En la **Figura 20** se presenta el diagrama esquemático de esta fase.

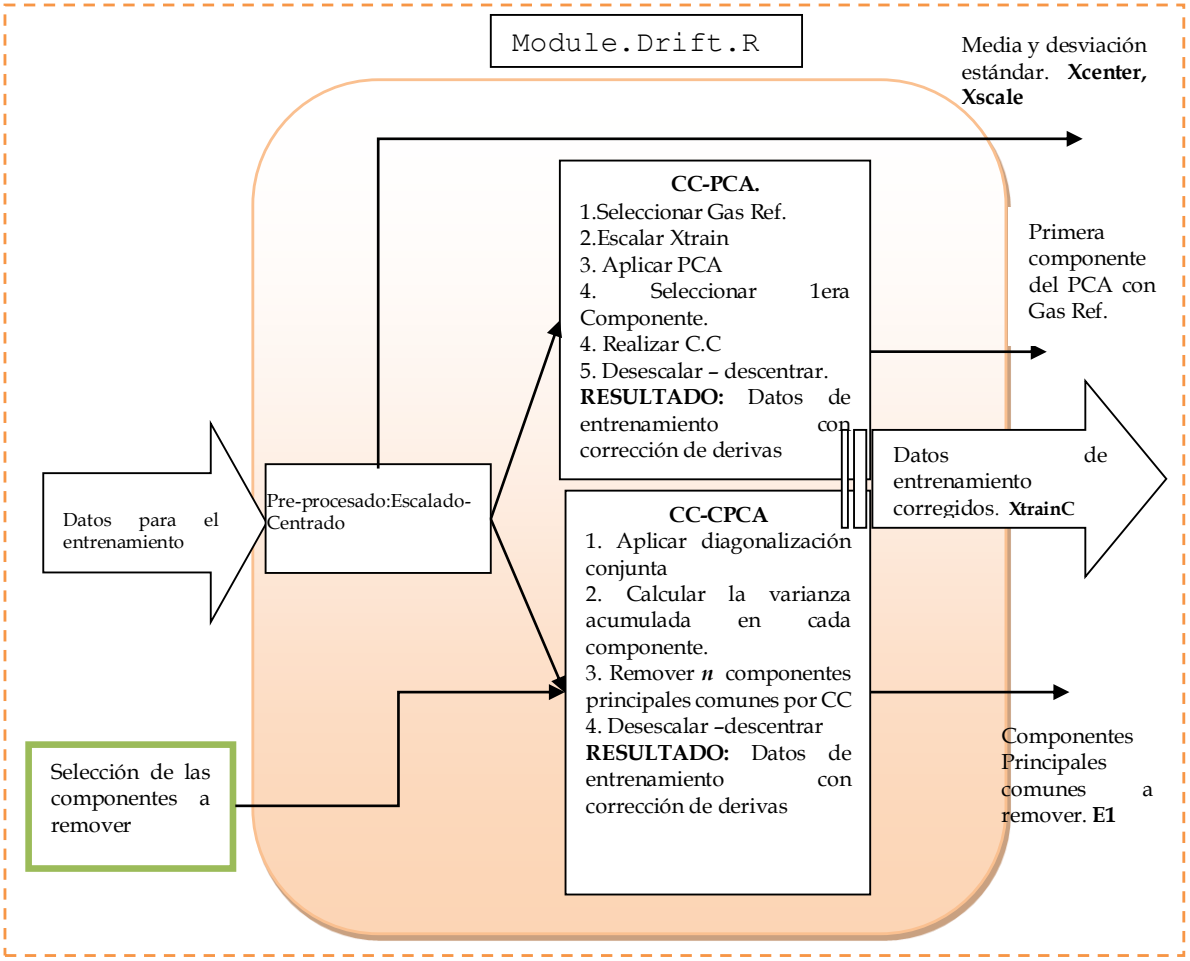


Figura 20. Diagrama esquemático del modulo de trabajo `moduleDrift.R`

2.3.4 Módulo para Validación (`moduleClass.R`)

Este módulo de trabajo en `R` realiza el proceso de encontrar el modelo para realizar la predicción de los datos a partir de los datos de entrenamiento, lo anterior teniendo en cuenta el modelo de extracción de componentes de deriva encontrado en el módulo de corrección de las derivas.

Las entradas al módulo son los datos de entrenamiento con corrección de derivas (bien sea por CC-CPCA o por CC-PCA), los lotes que se utilizan para la validación los cuales se toman del módulo de particiones (`moduleSplit.R`) y la matriz **E1** que contiene él o los componentes de deriva que se deben remover en cada uno de los lotes de validación cuando se aplique la corrección de componentes. De igual forma, para la normalización de los datos de validación, se utiliza la media y la desviación estándar obtenida a partir del espacio original de los datos de entrenamiento para garantizar que los datos de validación estén en una representación aproximada a los datos de entrenamiento antes de realizar la corrección de derivas.

Con el propósito de buscar una reducción de dimensionalidad en los datos, se aplica en la matriz de medidas para el entrenamiento un PCA, haciendo uso de la función `preProcess` del paquete `caret` en `R`, lo anterior con el propósito de obtener un nuevo espacio de representación en PCA, que acumulando el 99% de la varianza de información contenida en los datos originales logre reducir la dimensionalidad de la matriz **X**. Como resultado de lo anterior, se obtiene en este nuevo espacio de representación una matriz con el misma número de filas pero ahora con un número de columnas o componentes principales menor, las cuales contienen el 99% de la varianza de los datos, según se fijó en el parámetro `thresh` = 0.99. Basados en la transformación PCA de los datos de entrenamiento, se realiza a partir de este mismo espacio la transformación de los datos para la validación. Con la anterior, se busca mejorar el desempeño en las tareas de clasificación y garantizar que el costo computacional no sea elevado.

El modelo de clasificación se genera a partir del entrenamiento de los datos con corrección de derivas destinados para tal propósito, se emplea un clasificador k-NN, donde `k`, corresponde al número de vecinos a escoger. Para sintonizar este parámetro `k` se emplea una validación cruzada en el entrenamiento de 10 particiones y 10 repeticiones, haciendo uso de la función `traincontrol` del paquete `caret` y realizando el barrido entre valores de k = 2,3,4,5,6,7,8,9, de esta forma la función `train` es la encargada de generar el modelo para la clasificación haciendo el sintonizado del mejor *k*.

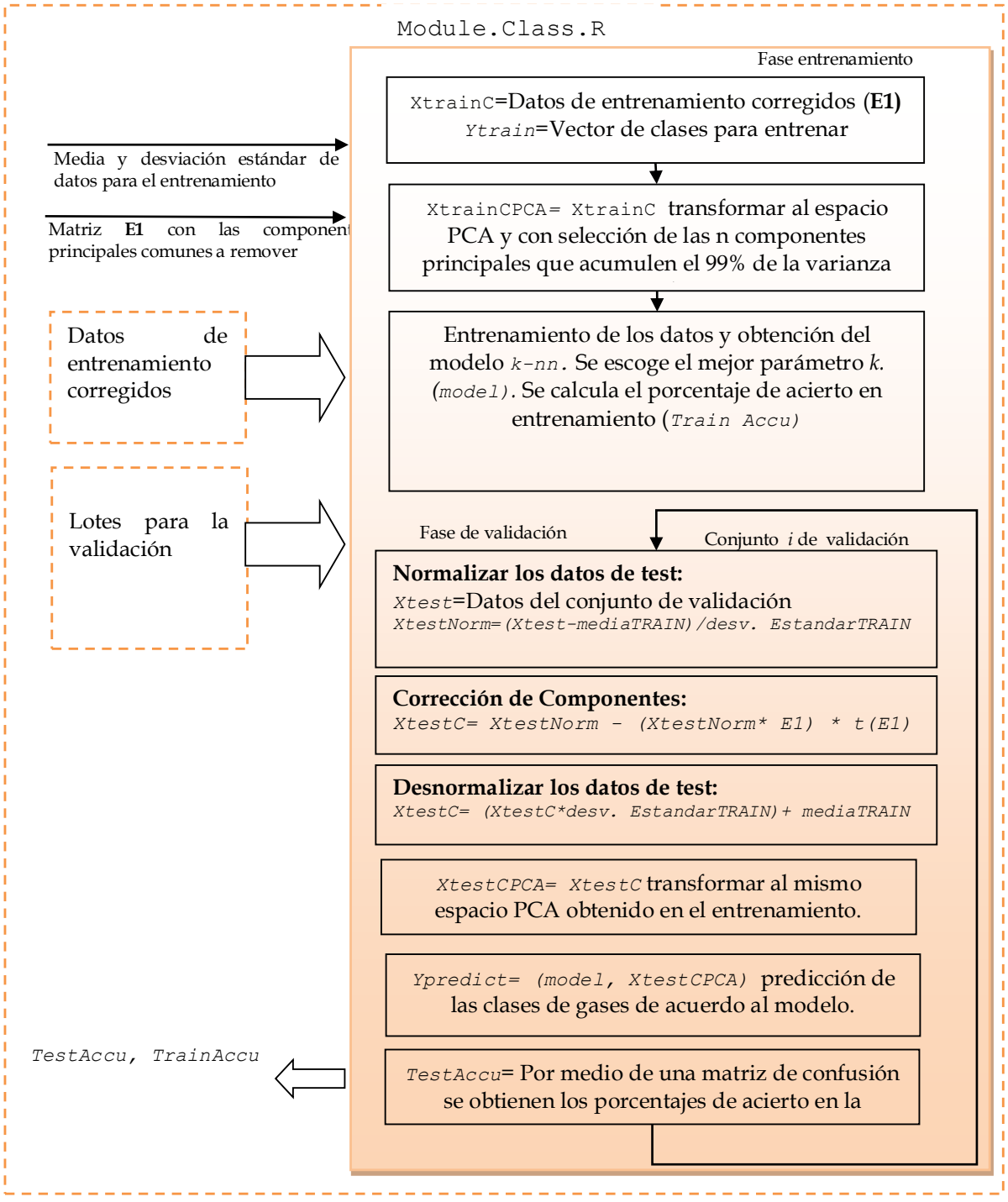


Figura 21. Diagrama de bloques del módulo para el entrenamiento y validación.

A continuación se explica el proceso realizado en los grupos de prueba a partir del modelo encontrado en la fase de entrenamiento, aplicado a los lotes restantes de la base de datos, destinados para la validación. Empleando un ciclo que toma uno a uno y en orden

ascendente los lotes destinados para tal fin, se realiza la validación de los datos. El primer paso a realizar luego de ser leído el lote *i*, es aplicarle normalización a partir de la media y la desviación estándar obtenida en el espacio original de los datos de entrenamiento, estos datos de validación normalizados son tratados por medio de la corrección de componentes, removiendo aquellos componentes de deriva que fueron pre-establecidos en el entrenamiento, es decir, la matriz **E1;** una vez corregidos los datos del conjunto *i* de validación, se des-escalan y se des-centran con el fin de tener nuevamente los datos en su representación inicial. A continuación, se transforman los datos al mismo modelo de espacio PCA hallado con los datos de entrenamiento y posteriormente se llevan al modelo del predictor k-NN obtenido igualmente en el entrenamiento.

Como parte final del módulo, se realiza una matriz de confusión para determinar los porcentajes de aciertos en los *i* conjuntos de validación y, asimismo obtener el acierto obtenido en el entrenamiento. Una diagrama de bloques que explica el módulo de observa en la **Figura 21.**

2.3.5 Selección de las componentes

El criterio para determinar el número máximo de componentes a remover en un conjunto de medidas de sensores químicos usados para la clasificación de compuestos volátiles, se determinó a partir del concepto de agrupamiento. La corrección de componentes por medio de CPCA, permite extraer un gran número de componentes y mediante este criterio se determina el número máximo de componentes a substraer con el propósito de mejorar la respuesta del sistema. El proceso llevado a cabo para determinar el criterio que permitió limitar el número de componentes a remover, basados en los conceptos de agrupamiento o clustering a partir del método de suma de cuadrados para hallar la matriz de dispersión dentro de las clases y la de dispersión entre clases, se presenta a continuación y se ha tomado de (Bishop, 2006).

Sea un conjunto de *n* datos: x_1, x_2, …, x_n. La matriz de covarianza de la muestra está dada por la ecuación 17

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)(X_i - m)^T \qquad (17)$$

Donde *m* es la media de la muestra y se expresa en la ecuación 18.

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (18)$$

Siendo *g* el número de grupos o clusters, la matriz de dispersión dentro de clases (S_w) se observa en la siguiente ecuación:

$$S_w = \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} Z_{ji}(X_i - m_j)\,(X_i - m_j)^T \qquad (19)$$

Donde, $Z_{ji} = 1$ si *X* pertenece al grupo *j,* 0 de lo contrario.

$m_j = \frac{1}{n_j}\sum_{i=1}^{n} Z_{ji}\, X_i$ es la media del cluster *j,* y $n_j = \sum_{i=1}^{n} Z_{ji}$ es el número de individuos en el cluster *j*.

La matriz de dispersión entre clases se presenta en la ecuación

$$S_B = \hat{\Sigma} - S_w = \sum_{j=1}^{g} \frac{n_j}{n}(m_j - m)(m_j - m)^T \qquad (20)$$

y describe la dispersión de la media de los cluster alrededor de la media total.

Luego de tener S_W y S_B se calcula la traza de la dispersión intra-clase $Tr(S_W)$ y la traza de la dispersión entre-clases $Tr(S_B)$ como la suma de los elementos de la diagonal, según se expresa en las ecuaciones 18 y 20, respectivamente.

$$Tr(S_w) = \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} Z_{ji}\left|X_i - m_j\right|^2 \qquad (21)$$

$$Tr(S_w) = \frac{1}{n}\sum_{j=1}^{g} S_j \qquad (22)$$

donde $S_j = \sum_{i=1}^{n} Z_{ji}\left|X_i - m_j\right|^2$, es la suma de cuadrados dentro del grupo, para el grupo *j*.

$$Tr(S_B) = \sum_{j=1}^{g} \frac{n_j}{n}\left|m_j - m\right|^2 \qquad (23)$$

Finalmente el criterio se obtiene de hallar la razón entre la traza entre-clases y la traza intra-clases, tal como se expresa en la ecuación 24.

$$C_s = \frac{Tr(S_B)}{Tr(S_W)} \qquad (24)$$

$\boldsymbol{C_s}$ se ha denominado en esta investigación como **criterio de separabilidad** y se utiliza para determinar el máximo número de componentes principales comunes a remover usando la técnica de CC-CPCA.

La función `daFischer` del paquete en `R` llamado `robCompositions` fue la empleada para calcular el criterio de separabilidad que determina el número máximo de componentes a remover.

Usando el conjunto de datos para el entrenamiento (con centrado y escalado), sin tratamiento de derivas y los mismos datos después de remover las componentes principales comunes escogidas para el tratamiento de derivas, se obtienen, para cada caso o experimento a partir de la función `daFischer`, las matrices de dispersión entre-clases e

intra-clases; posteriormente se halla la traza de cada una de estas matrices mediante la suma de los elementos de su diagonal y se calcula la razón entre la traza de la dispersión entre-clases y la traza de dispersión intra-clases; este último valor determina el **criterio de separabilidad.**

El criterio de separabilidad mejora cuanto más se aproxime a cero, dado que unos datos sin deriva son datos que se encuentran en el espacio de representación fuertemente cohesionados en su estructura interna de clases y a su vez las diferentes clases están claramente separadas o definidas. A partir de este concepto se determinó remover en cada experimento un número mayor de componentes hasta que el criterio de separabilidad en lugar de disminuir, empezase a incrementar. Se escoge el número de componentes en aquellas cuyo valor del criterio de separabilidad dejó de disminuir, de la misma forma se determina el promedio de los porcentajes de acierto para concretar que realmente se observan las mejoras en el desempeño del clasificador al tener unos datos con mínimo efecto de las derivas.

El criterio de separabilidad se prueba utilizando los datos de `chemosensors`, en los que se puede conocer de forma previa el número de compontes introducidos a los datos, la magnitud de los ruidos del sensor, la concentración del ruido de deriva y de esta forma tener la certeza de lo efectivo del método de remoción de componentes principales comunes y del criterio de selección de un número máximo de componentes al tener el conocimiento previo de la estructura misma de los datos usados. El método se verifica y se comprueba posteriormente en la base de datos experimental usada en este trabajo.

3. RESULTADOS

A continuación se presentan los resultados obtenidos en los experimentos realizados con las dos bases de datos que se emplearon, se inicia con las pruebas de corrección de derivas en los datos sintéticos de `chemosensors` y posteriormente se hace una comparación y validación de la metodología propuesta con las medidas experimentales de la Universidad de California.

3.1 PRUEBAS CON CHEMOSENSORS

Para iniciar el correspondiente análisis en el tema de las derivas se trabajó con los datos sintéticos de *chemosensors*, por lo tanto, se generó un escenario de trabajo con tres gases A, B y C a diferentes concentraciones. Se emplearon en orden cronológico 1100 muestras para el entrenamiento, estas se escogieron después de probar otras cantidades siendo este valor el mas adecuado y 1560 para la validación; en éstos últimos se realizaron 10 particiones, es decir, se generan 10 sub-conjuntos de 156 medidas cada uno, con el propósito de aplicar la ventana deslizante y analizar el efecto de las derivas a través del tiempo.

En la **Figura 22** se aprecia el espacio de trabajo generado para los datos sintéticos, el cual corresponde a tres analitos a diferentes concentraciones; en el lado izquierdo de la imagen se encuentra el espacio de los datos de entrenamiento y del lado derecho de la imagen se encuentra el de validación. El eje vertical representa la concentración (adimensional) y el eje horizontal corresponde al orden cronológico en el que se realizaron las medidas.

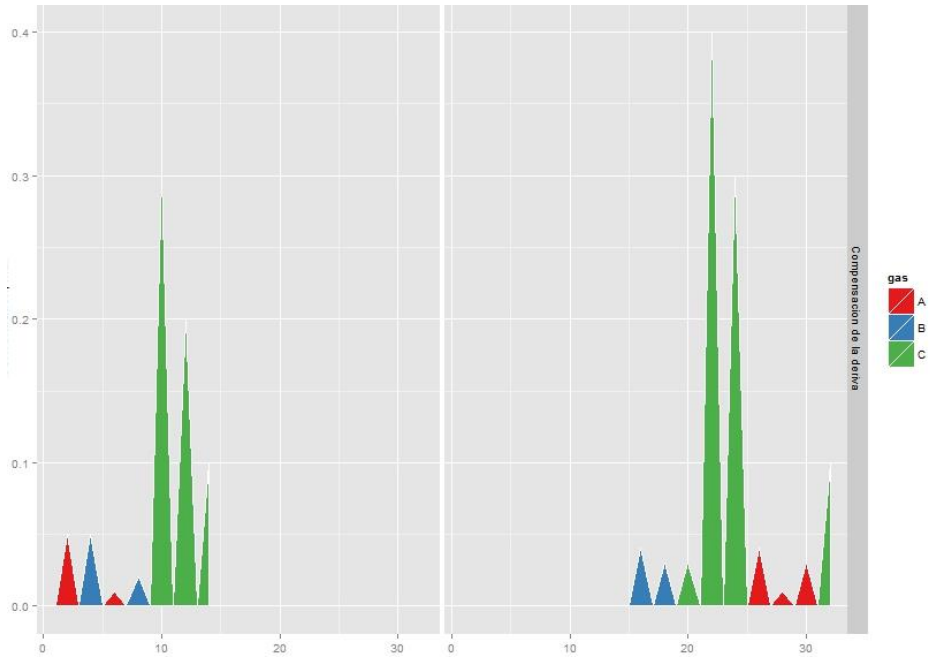


Figura 22. Escenario de trabajo generado con los datos sintéticos de `chemosensors`.

A partir de éste escenario, se realizaron una serie de experimentos los cuales consistieron en generar datos con mayor y menor ruido de deriva y que además incluyeran entre una y tres componentes añadidas de la misma; posteriormente se aplicó el método de corrección de componentes CC-CPCA para mitigar el efecto de las derivas en los datos. Adicionalmente, se empleó como método para determinar la mejora introducida al remover mayor cantidad de componentes, un criterio de agrupamiento basado en las matrices de dispersión entreclases e intraclases. Lo que se busca es, que al remover un mayor número de componentes, exista mayor cohesión interna en cada una de las clases o grupos y una separación claramente definida entre los mismos. A éste indicador basado en Fischer se le ha denominado en este trabajo **criterio de separabilidad**.

En la **Tabla 5** se presentan los parámetros utilizados en cada uno de los experimentos realizados con `chemosensors` con el propósito de validar la efectividad del criterio de separabilidad. Se generaron diferentes conjuntos de datos, variando los niveles de contaminación de derivas, donde, `ndcomp` representa el número de componentes de deriva, `dsd` corresponde al ruido de deriva, `ssd` hace referencia al ruido del sensor y `csd` corresponde a la concentración de ruido. Además, se usaron en todos los experimentos 16 sensores de las diferentes clases disponibles y la longitud de pulso `tunit=1`.

Parámetro / Experimento	ndcomp [1,2,3]	dsd [0,∞)	ssd [0,∞)	csd [0,∞)
1	3	2	0	0
2	2	2	0	0
3	1	2	0	0
4	3	0,1	0	0
5	2	0,1	0	0
6	1	0,1	0	0
7	1	0,1	0,1	0,1
8	3	2	2	2

Tabla 5. Detalle de los parámetros utilizados en los diferentes experimentos realizados con chemosensors.

3.1.1 Experimento 1. Datos de chemosensors con `dsd=1, ndcomp=2`

A continuación se expone el proceso realizado paso a paso en el experimento 1, los efectos de deriva incluidos en los datos corresponden a un ruido de deriva (`dsd`) con valor igual a 2 y 3 componentes de deriva (`ndcomp`).

La Figura 23, expone las dos primeras componentes del resultado del análisis PCA realizado al conjunto de entrenamiento de los datos generados para este primer experimento.

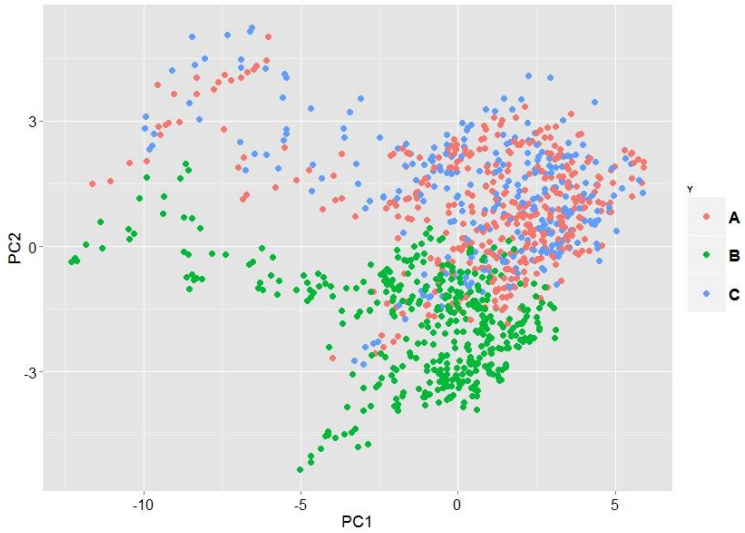


Figura 23. Resultados del PCA **(**componentes 1 y 2 **)** aplicado al conjunto de entrenamiento de chemosensors con `dsd=2` y 3 componentes de deriva.

Con los datos en este estado, es decir, sin corrección de las derivas, se realizó una prueba de clasificación en los 10 grupos de prueba y se obtuvieron los resultados expuestos en la **Tabla 6.** Allí se observan claramente los efectos de la deriva en la respuesta del clasificador.

Grupo	Porcentaje de acierto
Entrenamiento	87,273
Validación 1	42,949
Validación 2	54,487
Validación 3	39,744
Validación 4	41,667
Validación 5	38,462
Validación 6	26,282
Validación 7	35,897
Validación 8	32,051
Validación 9	23,077
Validación 10	28,846
Criterio de separabilidad	3,523

Tabla 6. Resultados obtenidos en experimento 1, sin tratar las derivas, dsd=2, ncomp=3.

Al observarse las dos primeras componentes del resultado del PCA del décimo grupo de validación presentado en la **Figura 24,** se detalla que las medidas están evidentemente solapadas, lo que hace más ardua la tarea de clasificación.

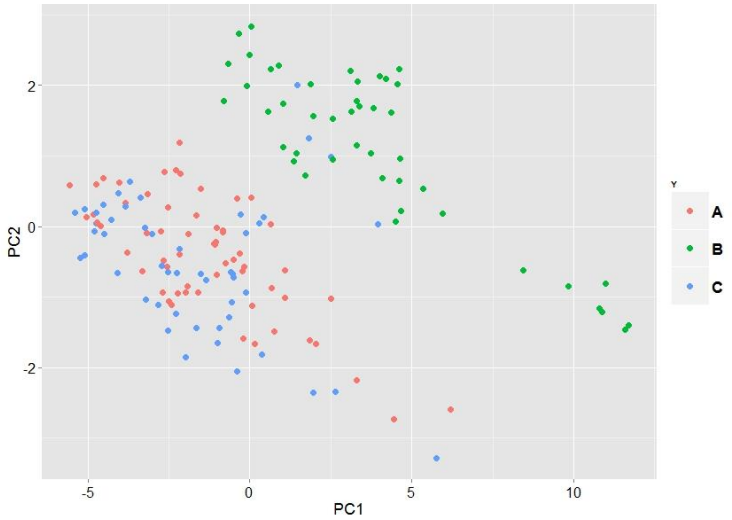


Figura 24. Resultado del PCA (componentes 1 y 2) del décimo grupo de validación sin corrección de la deriva.

Antes de aplicar CC-CCPA en la remoción de las derivas, se computan las varianzas acumuladas del conjunto de entrenamiento, estos resultados se pueden apreciar en la **Figura 25,** donde las 3 primeras componentes acumulan casi la totalidad de ella.

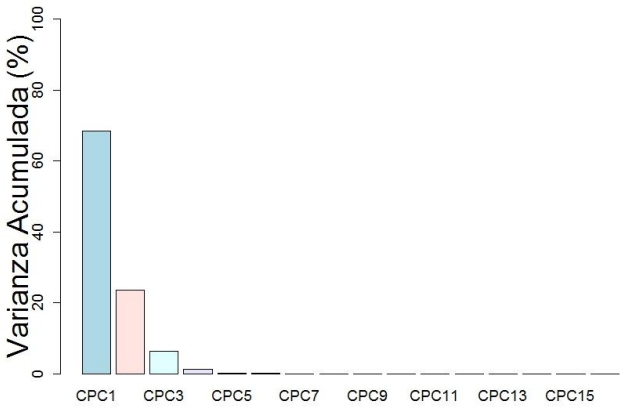


Figura 25. Porcentajes de varianza acumulada en las 16 componentes principales comunes de los datos de entrenamiento del experimento 1.

Con el propósito de mejorar la respuesta obtenida y de observar la influencia al remover diferentes componentes de deriva, se inició con la remoción de la primera componente, siendo esta la que mayor varianza acumula y luego se probó removiendo otras componentes adicionales para observar el efecto generado en los datos, los resultados se observan en la **Figura 26** y en la **Tabla 7.**

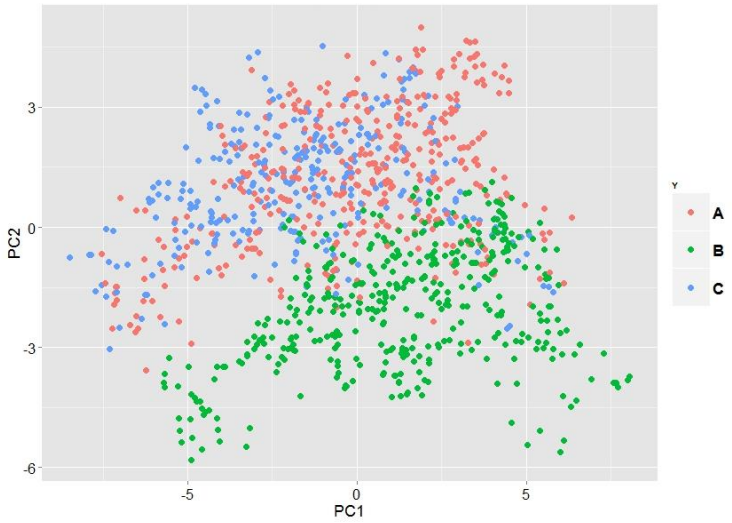


Figura 26. Resultado del PCA (componentes 1 y 2) aplicado al conjunto de datos de entrenamiento del experimento 1 con `dsd=2` y 3 componentes de deriva, después de remover la primera componente CPCA.

Grupo	Porcentaje de acierto
Entrenamiento	96,818
Validación 1	96,795
Validación 2	83,974
Validación 3	50,641
Validación 4	59,615
Validación 5	78,846
Validación 6	33,974
Validación 7	35,897
Validación 8	32,051
Validación 9	23,077
Validación 10	28,846
Criterio de Separabilidad	2,992

Tabla 7. Resultados obtenidos en experimento 1 con `dsd=2, ndcomp=3` al remover la primera componente del análisis CPCA.

Se observó también el cambio ocurrido en el grupo de validación 10 al remover esta primera componente, mediante el cómputo del PCA graficado en sus dos primeras componentes en la **Figura 27.** Como era de esperarse en este primer intento de mejorar la respuesta del clasificador al remover solo la primera componente, los resultados de la T**abla 6** no son tan notorios en los últimos subconjuntos de prueba como en los primeros lotes, donde sí aumentó el porcentaje de acierto con respecto a los datos sin tratamiento de las derivas; lo

anterior es debido a que los datos poseen 3 componentes de deriva y hasta el momento solo se ha removido la primera de ellas.

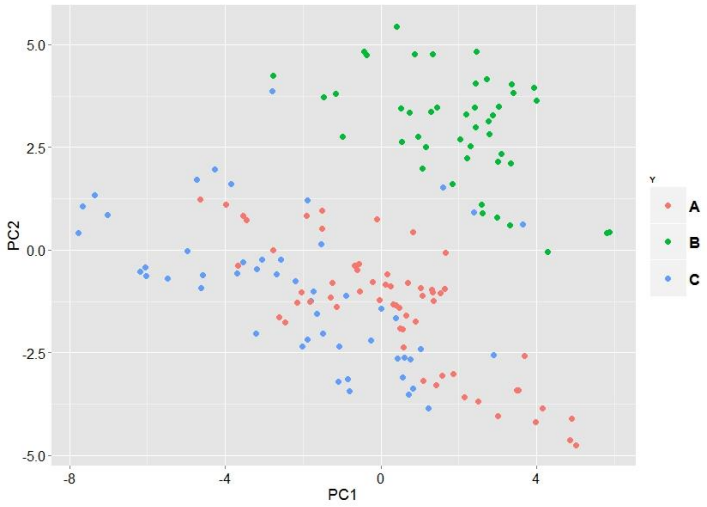


Figura 27. Resultados del PCA (componentes 1 y 2) del décimo grupo de validación al remover la primera componente por CC-CPCA.

Al remover la segunda componente, los resultados obtenidos en el conjunto de entrenamiento y al validar en los subconjuntos de prueba se observan en la **Figura 28** y en la **Tabla 8.**

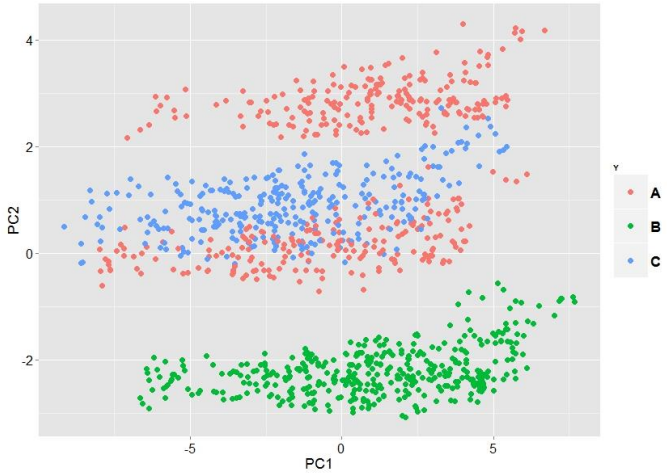


Figura 28. PCA aplicado al conjunto de entrenamiento del experimento 1 con `dsd=2` y 3 componentes de deriva, después de remover las dos primeras componentes principales comunes por CC-CPCA.

Grupo	Porcentaje de acierto
Entrenamiento	100,000
Validación 1	98,077
Validación 2	98,077
Validación 3	73,718
Validación 4	91,026
Validación 5	92,308
Validación 6	67,308
Validación 7	55,128
Validación 8	46,154
Validación 9	62,179
Validación 10	60,897
Criterio de separabilidad	2,632

Tabla 8. Resultados obtenidos en el experimento 1 con `dsd=2, ndcomp=3`, al remover las dos primeras componentes principales comunes.

Se observa también en la **Figura 29** el PCA del décimo grupo de validación, con el fin de apreciar la mejora introducida.

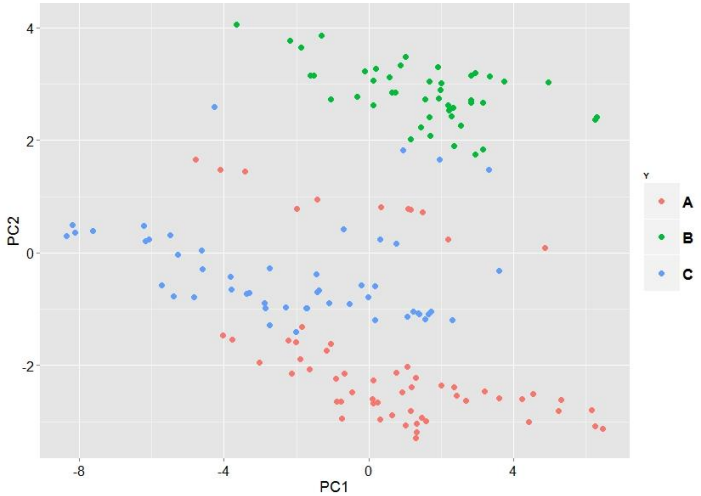


Figura 29. Resultados del PCA (componentes 1 y 2) del décimo grupo de validación al remover las dos primeras componentes principales comunes por CC-CPCA.

Continuando con el orden establecido, se removieron las tres primeras componentes principales comunes, como era de esperarse el clasificador mejoró su respuesta de forma significativa y el criterio de separabilidad tiende a cero al removerse el mismo número de componentes que le fueron introducidas a los datos cuando se generó este espacio de medidas sintéticas.

En la **Figura 30** y en la **Tabla 9** se observan los efectos que causa el remover del conjunto de datos las tres primeras componentes principales comunes mediante la técnica de CC-CPCA.

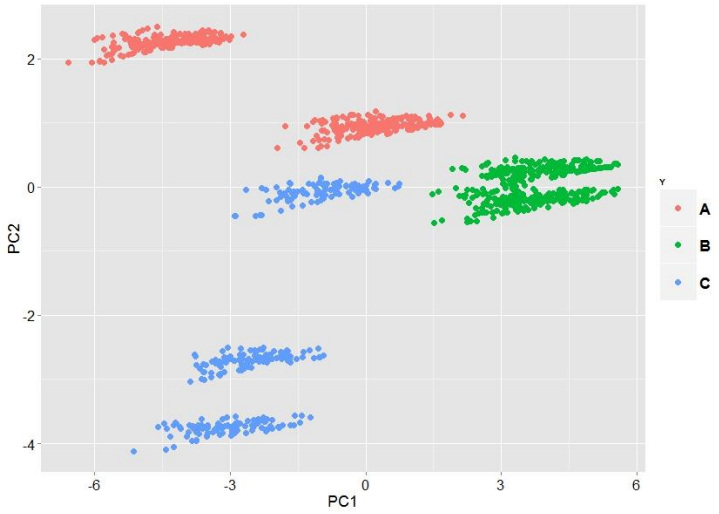


Figura 30. Resultados PCA (componentes 1 y 2) aplicado al conjunto de entrenamiento l, con `dsd=2` y 3 componentes de deriva, después de remover las tres primeras componentes principales comunes.

Grupo	Porcentaje de acierto
Entrenamiento	100,000
Validación 1	99,359
Validación 2	98,718
Validación 3	95,513
Validación 4	99,359
Validación 5	95,513
Validación 6	87,821
Validación 7	57,051
Validación 8	57,051
Validación 9	73,718
Validación 10	66,026
Criterio de separabilidad	1,492

Tabla 9. Resultados obtenidos en el experimento 1, con `dsd=2, ncomp=3` al remover las tres primeras componentes principales comunes.

Al igual que en las dos pruebas anteriores, se graficó el PCA del grupo número de 10 de los subconjuntos de validación, éste resultado se presenta en la **Figura 31,** donde se observa la separabilidad de las diferentes clases de gases presentes en este grupo**.**

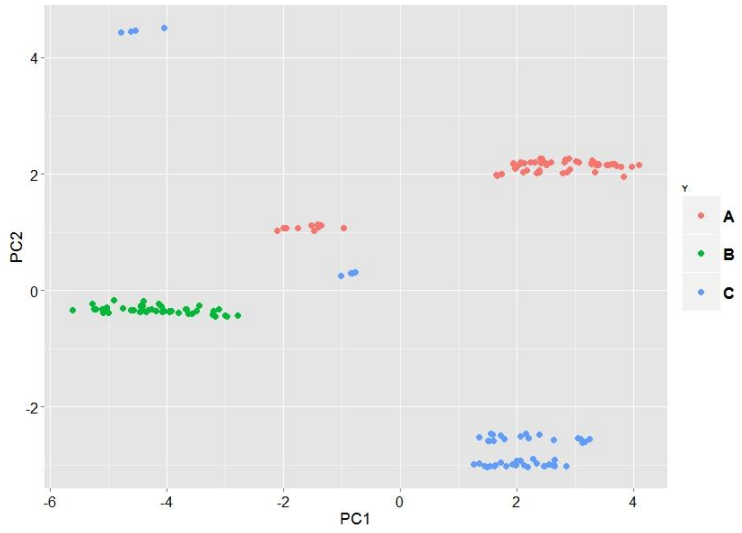


Figura 31. Resultado del PCA (componentes 1 y 2) del décimo grupo de validación al remover las tres primeras componentes principales comunes por CC-CPCA.

Se añade al conjunto de componentes por remover la componente número 4. Los resultados obtenidos en el conjunto de validación y en los subconjuntos de prueba se observan en la **Figura 32** y en la **Tabla 10.**

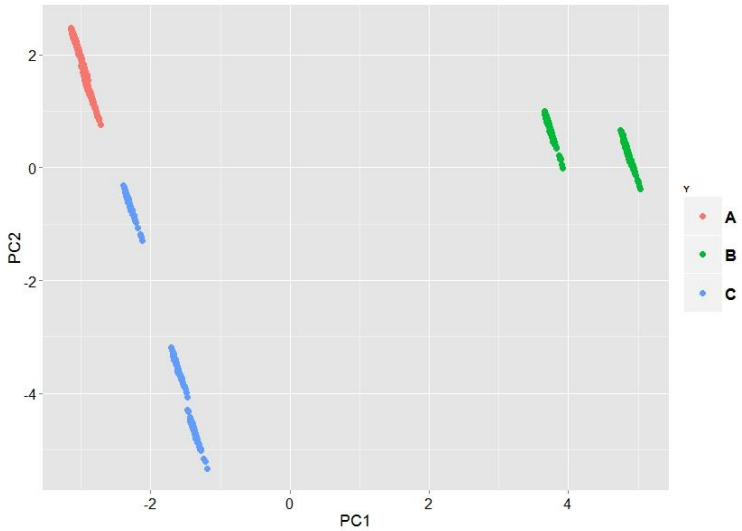


Figura 32. Resultado del PCA (componentes 1 y 2) aplicado al conjunto de entrenamiento en el experimento 1, con `dsd=2` y 3 componentes de deriva, después de remover las cuatro primeras componentes principales comunes.

Grupo	Porcentaje de acierto
Entrenamiento	100
Validación 1	98,077
Validación 2	98,718
Validación 3	95,513
Validación 4	100
Validación 5	96,795
Validación 6	98,077
Validación 7	97,436
Validación 8	96,795
Validación 9	98,077
Validación 10	97,436
Criterio de Separabilidad	0,131

Tabla 10. Resultados obtenidos en experimento 1, `dsd=2, ndcomp=3` al remover las cuatro primeras componentes principales comunes.

En este caso el criterio de optimización es muy cercano a cero, lo que indica que remover las 4 primeras componentes resultó ser representativo para mejorar el sistema y mitigar las derivas. El PCA del décimo conjunto de validación refleja este efecto, al concentrar los datos en cada grupo y separar cada uno de ellos.

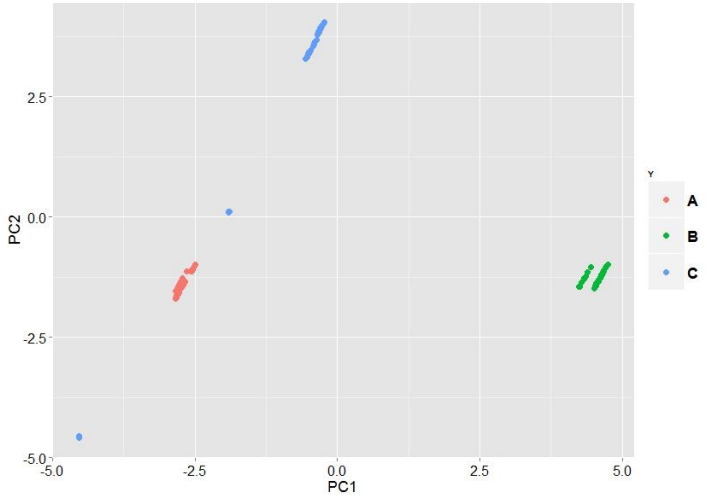


Figura 33. Resultado del PCA (componentes 1 y 2) del décimo grupo de validación al remover las cuatro primeras componentes principales comunes por CC-CPCA.

En este caso, en el PCA del subconjunto de validación 10, presentado en la **Figura 33,** presentó mejoras con respecto al PCA inicial (**Figura 24)**, en el que a los datos no se les había suprimido componente alguna.

Con el propósito de obtener una respuesta aún más cercana a lo ideal, es decir lograr unos datos sin deriva, se remueve otra componente adicional usando la técnica CC-CPCA. Los datos se presentan en la **Figura 34** y se detallan en la **Tabla 11**.

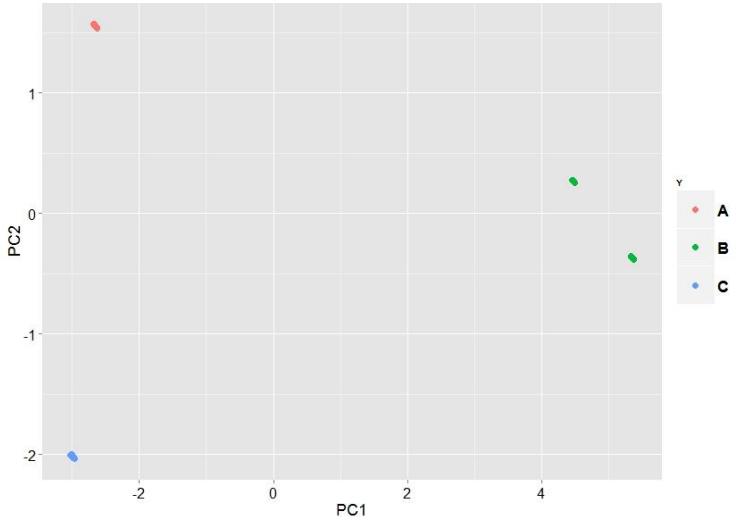


Figura 34. Resultado del PCA (componentes 1 y 2) aplicado al conjunto de entrenamiento del experimento 1 `con dsd=2` y 3 componentes de deriva, después de remover las cinco primeras componentes principales comunes.

Grupo	Porcentaje de acierto
Entrenamiento	100
Validación 1	100
Validación 2	100
Validación 3	100
Validación 4	100
Validación 5	100
Validación 6	100
Validación 7	100
Validación 8	100
Validación 9	100
Validación 10	100
Criterio de separabilidad	~ 0

Tabla 11. Resultados obtenidos el experimento 1 con `dsd=2, ncomp=3` al remover las cinco primeras componentes principales comunes.

En la **Figura 35,** se observa que los datos del subconjunto de validación número 10, estaban lo suficientemente cohesionados entre sí y a su vez separados por clases.

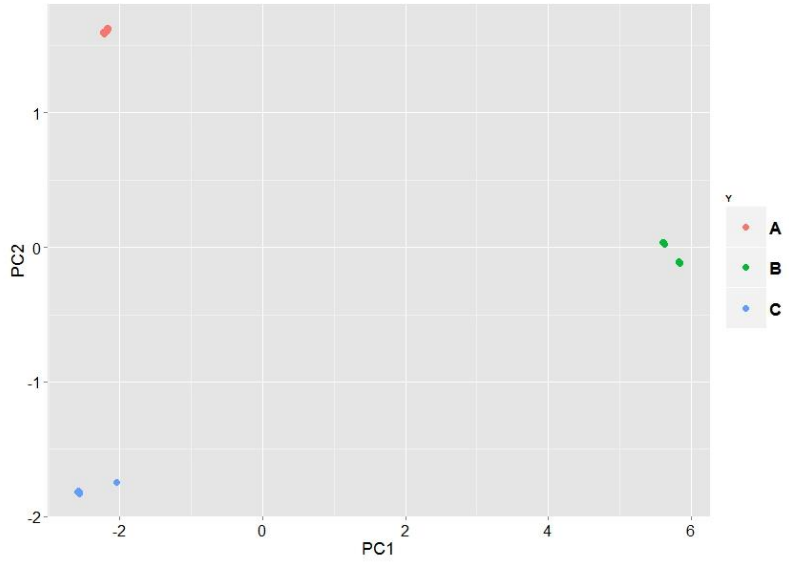


Figura 35. PCA del décimo grupo de validación al remover las cinco primeras componentes principales comunes por CC-CPCA.

La **Figura 36,** presenta un comparativo en el proceso de remover en cada paso del experimento una mayor cantidad de componentes de deriva y sus efectos en el sistema de clasificación; la mejora introducida en la remoción de componentes es medible bajo el **criterio de separabilidad** establecido**.** Esta gráfica y las demás de este estilo presentadas en este libro, poseen en el eje de vertical los porcentajes de acierto del clasificador k-NN en escala de 0 a 100 y en el eje horizontal se numeran los grupos de validación empleados. Este tipo de gráficas permiten realizar un análisis en el tiempo de las derivas, su influencia en los sensores y por ende analizar la repetitividad de la respuesta de los sistemas de reconocimiento de olores, especialmente en aquellos grupos que se alejan en el tiempo del conjunto de datos con los que se entrenó el clasificador.

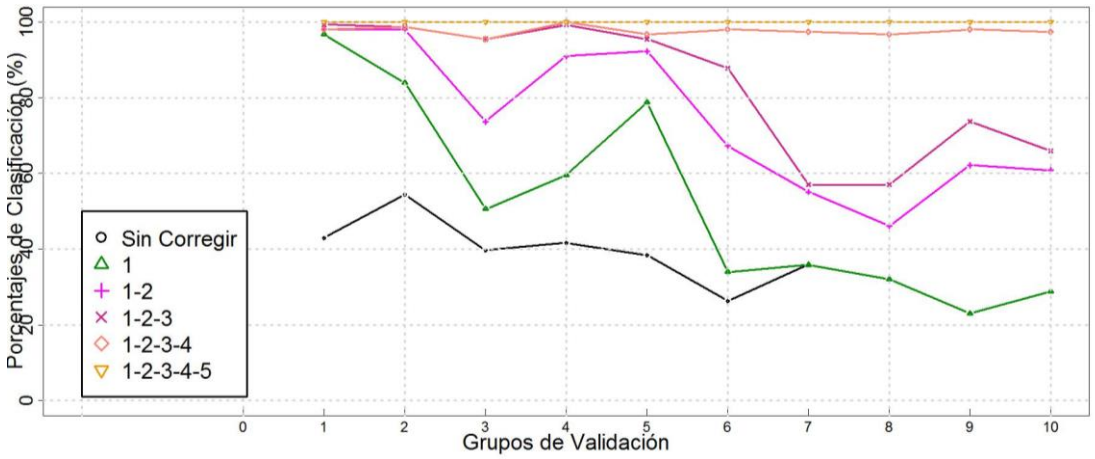


Figura 36. Gráfica comparativa de la remoción de componentes principales comunes en los datos del experimento 1 con `dsd=2` y `ndcomp=3`.

Se observó la mejora en la respuesta del sistema clasificador al aumentar el número de componentes removidas y de forma cuantitativa se muestran los mismos resultados en la **Tabla 12**. Aun cuando el número de componentes añadidas a los datos fueron en total 3, se removieron 5 de ellas para llegar al 100% de los aciertos en el clasificador, esto es causado por el factor ruido de deriva que se manejó con el parámetro `dsd`.

Grupo	Porcentaje de acierto					
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4	C 1 a 5
Entrenamiento	42,948	96,794	98,076	99,358	98,076	**100**
Validación 1	54,487	83,974	98,076	98,717	98,717	**100**
Validación 2	39,743	50,641	73,717	95,512	95,512	**100**
Validación 3	41,666	59,615	91,025	99,358	100	**100**
Validación 4	38,461	78,846	92,307	95,512	96,794	**100**
Validación 5	26,282	33,974	67,307	87,820	98,076	**100**
Validación 6	35,897	35,897	55,128	57,051	97,435	**100**
Validación 7	32,051	32,051	46,153	57,051	96,794	**100**
Validación 8	23,076	23,076	62,179	73,717	98,076	**100**
Validación 9	28,846	28,846	60,897	66,025	97,435	**100**
Validación 10	42,948	96,794	98,076	99,358	98,076	**100**
Promedio	36,346	52,371	74,487	83,012	97,692	**100**
Criterio de Separabilidad	3,523	2,992	2,632	1,492	0,131	**0**

Tabla 12. Resumen de los resultados obtenidos en el experimento 1, usando chemosensors con `dsd=2, ndcomp=3`, al remover desde la primera hasta las cinco primeras componentes principales comunes.

3.1.2 Experimento 2. Datos de chemosensors con `dsd=2, ndcomp=2`

Siguiendo la misma metodología y secuencia de pasos presentados en forma detallada en el **Experimento 1** (Sección 3.1.1)**,** se presentan en la **Figura 37** y en la **Tabla 13** el consolidado de lo obtenido usando datos con ruido de deriva igual a 2 y con 2 componentes de deriva añadidos a los datos sintéticos. En este caso se presenta el mismo efecto de mejoras en la respuesta del sistema clasificador cuando se remueve un mayor número de componentes.

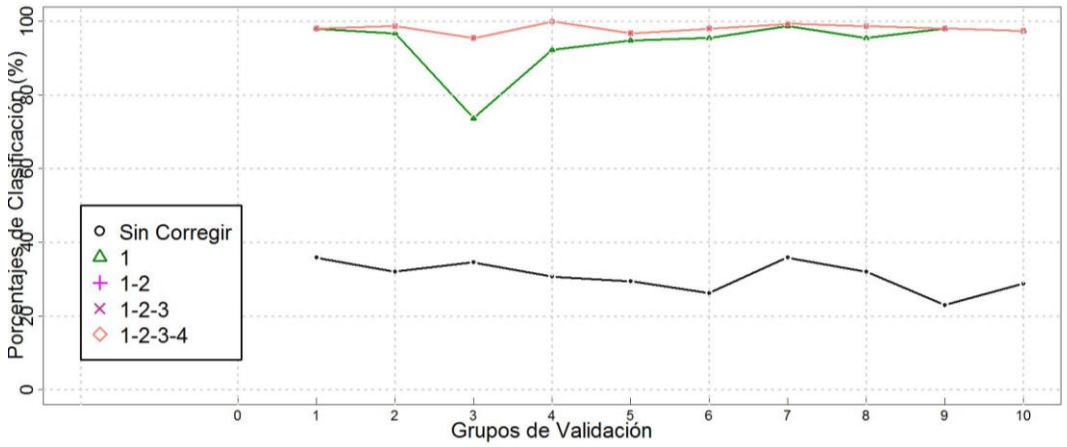


Figura 37. Gráfica comparativa de la remoción de componentes principales comunes por CC-CPCA en los datos de chemosensors con `dsd=2` y `ndcomp=2`, generados en el experimento 2.

Grupo	Porcentaje de acierto				
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4
Entrenamiento	70,818	100	100	100	**100**
Validación 1	35,897	98,076	98,077	98,077	**98,077**
Validación 2	32,051	96,794	98,718	98,718	**98,718**
Validación 3	34,615	73,717	95,513	95,513	**95,513**
Validación 4	30,769	92,307	100	100	**100**
Validación 5	29,487	94,872	96,794	96,795	**96,795**
Validación 6	26,282	95,513	98,077	98,077	**98,077**
Validación 7	35,897	98,718	99,356	99,359	**99,359**
Validación 8	32,051	95,513	98,718	98,718	**98,718**
Validación 9	23,076	98,077	98,077	98,077	**98,077**
Validación 10	28,846	97,436	97,436	97,436	**97,436**
Promedio	30,897	94,102	98,077	98,077	**98,077**
Criterio de separabilidad	4,064	1,801	0,427	0,142	**0,011**

Tabla 13. Resumen de los resultados obtenidos en el experimento 2, usando chemosensors con `dsd=2`, `ndcomp=2`, al remover desde la primera hasta las 4 primeras componentes principales comunes por CC-CPCA.

3.1.3 Experimento 3. Datos de chemosensors con `dsd=2, ndcomp=1`

En la **Figura 38** y en la **Tabla 14** se presentan los resultados obtenidos cuando el ruido de deriva es igual a 2 y los datos contienen solo una componente añadida de deriva. Se destaca en estos resultados que al remover una componente más por encima del mejor **criterio de separabilidad** obtenido, el promedio de porcentajes de acierto en el clasificador decrece, lo cual refleja que este criterio evidencia el máximo número de componentes de deriva que se

requieren suprimir en la corrección de componentes para obtener una mejor respuesta del sistema.

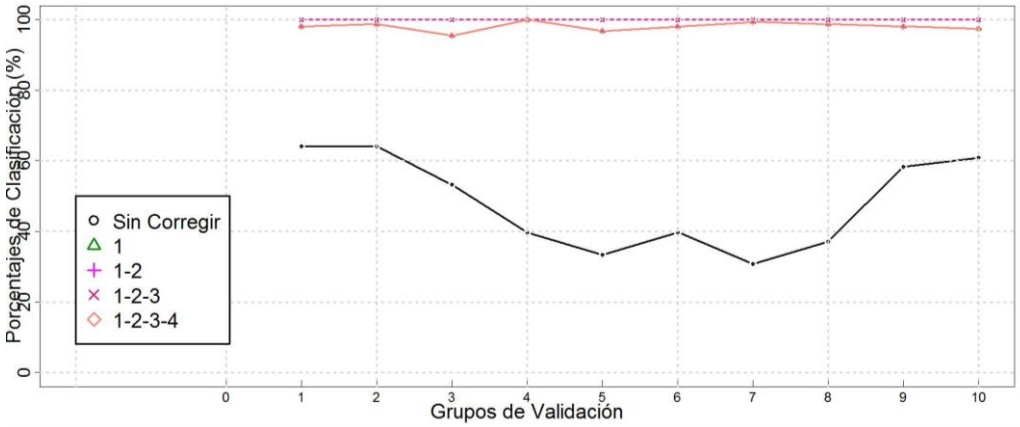


Figura 38. Gráfica comparativa de la remoción de componentes principales comunes por CC-CPCA en los datos de chemosensors con `dsd=2` y `ndcomp=1`, generados en el experimento 3.

Grupo	Porcentaje de acierto				
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4
Entrenamiento	94,091	100	100	**100**	100
Validación 1	64,102	98,076	98,076	**100**	98,076
Validación 2	64,102	98,717	98,717	**100**	98,717
Validación 3	53,205	95,512	95,513	**100**	95,512
Validación 4	39,743	100	100	**100**	100
Validación 5	33,333	96,794	96,795	**100**	96,794
Validación 6	39,743	98,076	98,077	**100**	98,076
Validación 7	30,769	99,358	99,356	**100**	99,358
Validación 8	37,179	98,717	98,718	**100**	98,717
Validación 9	58,333	98,076	98,077	**100**	98,076
Validación 10	60,897	97,435	97,436	**100**	97,435
Promedio	48,1406	98,0761	98,0765	**100**	98,0761
Criterio de Separabilidad	1,397	0,197	0,123	**0,001**	0,112

Tabla 14. Resumen de los resultados obtenidos en el experimento 3, usando chemosensors con `dsd=2`, `ndcomp=1`, al remover desde la primera hasta las 4 primeras componentes principales comunes por CC-CPCA.

3.1.4 Experimento 4. Datos de chemosensors con `dsd=0.1, ndcomp=3`

En la **Figura 39** y en la **Tabla 15** se muestran los resultados con ruido de deriva de 0.1 y 3 componentes de deriva.

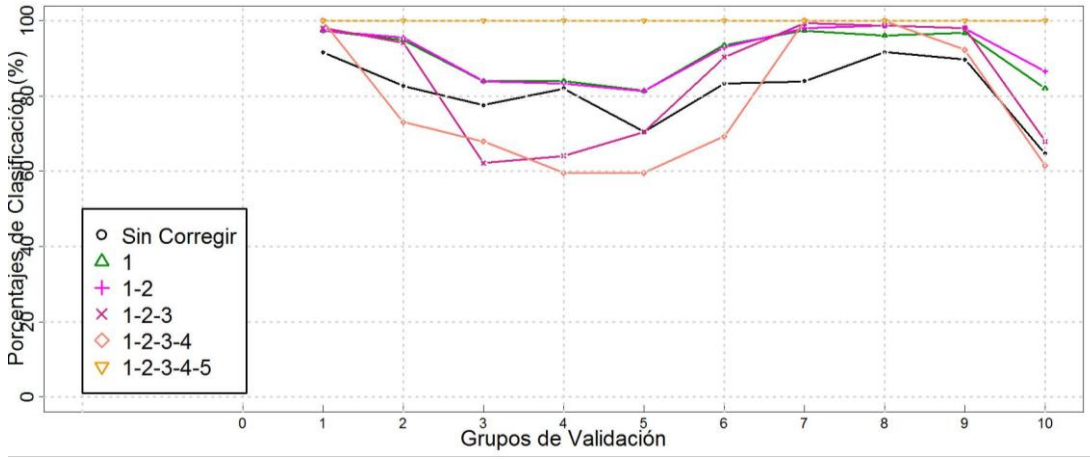


Figura 39. Gráfica comparativa de la remoción de componentes principales comunes en los datos de chemosensors con `dsd=0.1` y `ndcomp=3`.

Grupo	Porcentaje de acierto					
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4	C 1 a 5
Entrenamiento	98,181	99,727	100	100	100	**100**
Validación 1	91,667	97,435	97,435	98,076	100	**100**
Validación 2	82,692	94,871	95,512	94,230	73,076	**100**
Validación 3	77,564	83,974	83,974	62,179	67,948	**100**
Validación 4	82,051	83,974	83,334	64,102	59,615	**100**
Validación 5	70,512	81,411	81,411	70,512	59,615	**100**
Validación 6	83,334	93,589	92,949	90,384	69,231	**100**
Validación 7	83,974	97,435	98,077	99,358	100	**100**
Validación 8	91,667	96,153	98,718	98,717	100	**100**
Validación 9	89,744	96,794	98,077	98,076	92,307	**100**
Validación 10	64,744	82,051	86,539	67,948	61,538	**100**
Promedio	81,7949	90,7687	91,6026	84,3582	78,333	**100**
Criterio de separabilidad	1,506	0,889	0,282	0,114	0,110	**0,025**

Tabla 15. Resumen de los resultados obtenidos en el experimento 4, usando chemosensors con `dsd=0.1`, `ndcomp=3`, al remover desde la primera hasta las 5 primeras componentes principales comunes.

3.1.5 Experimento 5. Datos de chemosensors con `dsd=0.1, ndcomp=2`

En la **Figura 40** y en la **Tabla 16** se muestran los resultados con ruido de deriva de 0.1 y 2 componentes de deriva. En este caso nuevamente se observa que al aumentar el número de componentes removidas por CC-CPCA el criterio de separabilidad tiende a cero y el porcentaje de aciertos en el clasificador por ende se aproximan al 100 por ciento.

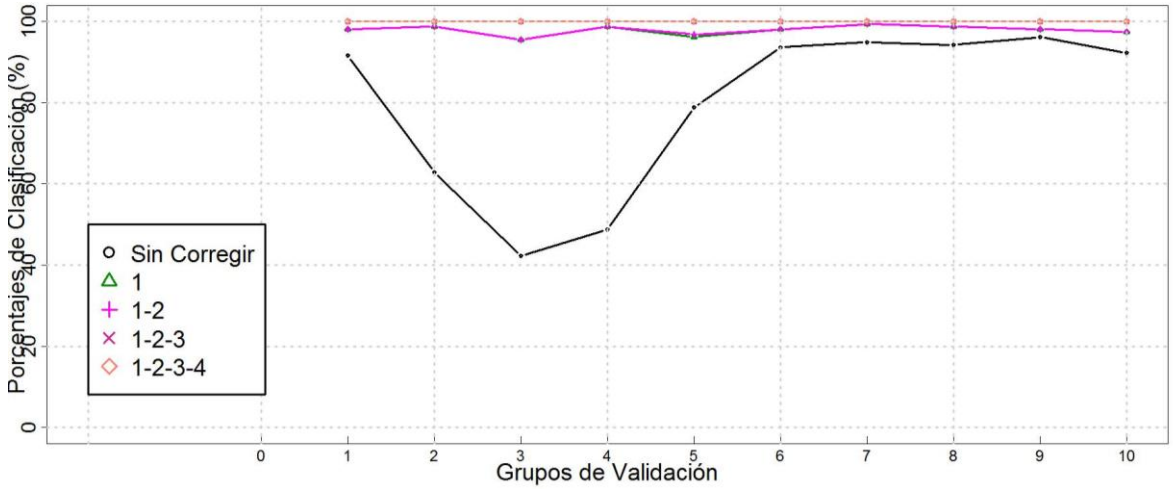


Figura 40. Gráfica comparativa de la remoción de componentes principales comunes en los datos de chemosensors con `dsd=0.1` y `ndcomp=2`.

Grupo	Porcentaje de acierto				
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4
Entrenamiento	96,545	100	100	100	**100**
Validación 1	91,667	98,077	98,076	100	**100**
Validación 2	62,821	98,718	98,717	100	**100**
Validación 3	42,307	95,512	95,512	100	**100**
Validación 4	48,717	98,718	98,717	100	**100**
Validación 5	78,846	96,154	96,794	100	**100**
Validación 6	93,589	98,077	98,076	100	**100**
Validación 7	94,871	99,359	99,358	100	**100**
Validación 8	94,231	98,718	98,717	100	**100**
Validación 9	96,153	98,077	98,076	100	**100**
Validación 10	92,307	97,436	97,435	100	**100**
Promedio	79,551	97,885	97,948	100	**100**
Criterio de separabilidad	1,203	0,442	0,251	0,237	**0,015**

Tabla 16. Resumen de los resultados obtenidos en el experimento 5, usando chemosensors con `dsd=0.1`, `ndcomp=2`, al remover desde la primera hasta las 4 primeras componentes principales comunes.

3.1.6 Experimento 6. Datos de chemosensors con `dsd=0.1, ndcomp=1`

Los resultados con 0.1 de ruido de deriva y una componente de deriva incluida en los datos, se presentan en la **Figura 41** y de forma cuantitativa se expresan en la **Tabla 17.** Nuevamente se logra remover el total de las componentes de deriva, para generar un desempeño del 100% en el clasificador.

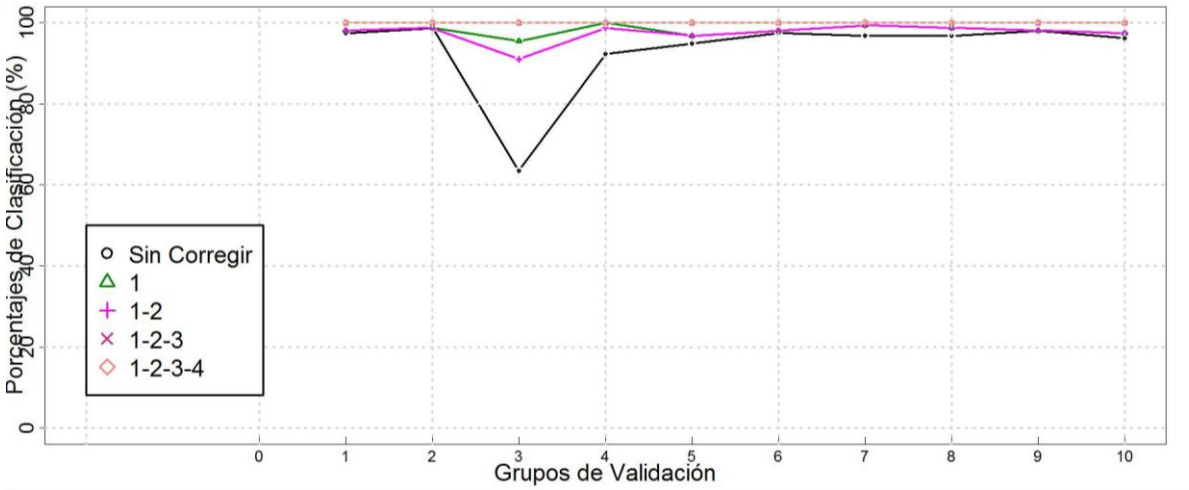


Figura 41. Gráfica comparativa de la remoción de componentes principales comunes en los datos de chemosensors con `dsd=0.1` y `ndcomp=1`.

Grupo	Porcentaje de acierto				
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4
Entrenamiento	99,909	100	100	100	**100**
Validación 1	97,436	98,077	98,077	100	**100**
Validación 2	98,718	98,718	98,718	100	**100**
Validación 3	63,462	95,513	91,026	100	**100**
Validación 4	92,308	100	98,718	100	**100**
Validación 5	94,872	96,795	96,795	100	**100**
Validación 6	97,436	98,077	98,077	100	**100**
Validación 7	96,795	99,359	99,359	100	**100**
Validación 8	96,795	98,718	98,718	100	**100**
Validación 9	98,077	98,077	98,077	100	**100**
Validación 10	96,154	97,436	97,436	100	**100**
Promedio	93,2053	98,077	97,501	100	**100**
Criterio de separabilidad	1,148	0,319	0,304	0,012	**0,000**

Tabla 17. Resumen de los resultados obtenidos en el experimento 6, usando chemosensors con `dsd=0.1`, `ndcomp=1`, al remover desde la primera hasta las 4 primeras componentes principales comunes.

3.1.7 Experimento 7. Datos de chemosensors con `dsd=0.1, csd=0.1, ssd= 0.1` y `ndcomp=1`

Con el propósito de analizar el comportamiento de los datos al poseer, además de ruido de deriva, ruido del sensor (`ssd`) y concentración de ruido (`csd`), se generaron los experimentos 7 y 8. La **Figura** 42, muestra el comparativo al remover componentes mediante CC-CPCA y en la T**abla 18** se exponen los resultados en forma cuantitativa para el experimento número 7.

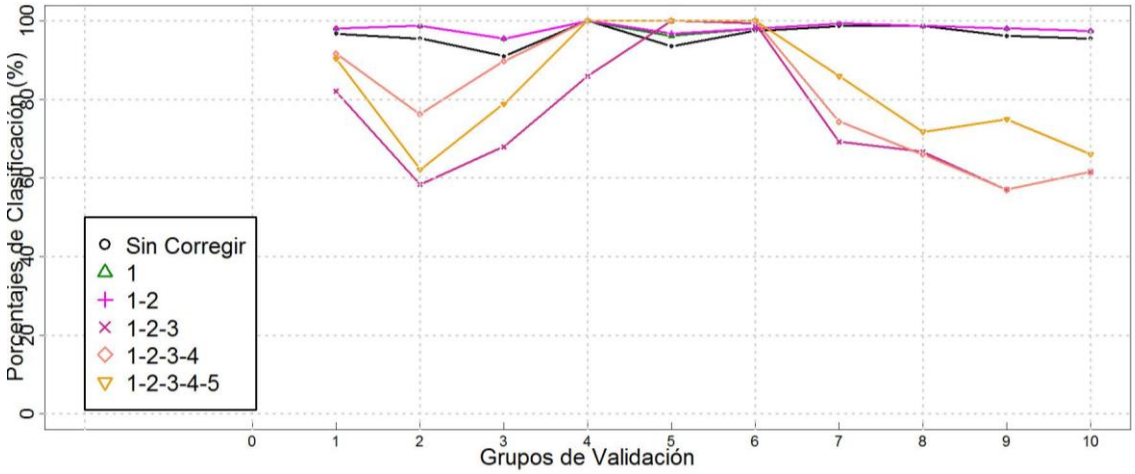


Figura 42. Gráfica comparativa de la remoción de componentes principales comunes en los datos de chemosensors con `dsd=0.1, csd=0.1, ssd=0.1` y `ndcomp=1`.

Grupo	Porcentaje de acierto					
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4	C 1 a 5
Entrenamiento	99,727	100	**100**	100	100	100
Validación 1	96,795	98,077	**98,077**	82,051	91,667	90,385
Validación 2	95,513	98,718	**98,718**	58,333	76,282	62,179
Validación 3	91,026	95,513	**95,513**	67,949	89,744	78,846
Validación 4	100	100	**100**	85,897	100	100
Validación 5	93,590	96,154	**96,795**	100	100	100
Validación 6	97,436	98,077	**98,077**	99,359	100	100
Validación 7	98,718	99,359	**99,359**	69,231	74,359	85,897
Validación 8	98,718	98,718	**98,718**	66,667	66,026	71,795
Validación 9	96,154	98,077	**98,077**	57,051	57,051	75
Validación 10	95,513	97,436	**97,436**	61,538	61,538	66,026
Promedio	96,346	98,0129	**98,077**	74,807	81,666	83,012
Criterio de separabilidad	1,743	0,251	**0,144**	0,193	0,103	0,027

Tabla 18. Resumen de los resultados obtenidos en el experimento 7, usando chemosensors con `dsd=0.1, csd=0.1, ssd=0.1` y `ndcomp=1`, al remover desde la primera hasta las 5 primeras componentes principales comunes.

3.1.8 Experimento 8. Datos de chemosensors con `dsd=4, csd=2, ssd= 2 y ndcomp=3`

En este experimento al igual que en el experimento 7, se buscó comprobar el efecto de tener mayor ruido del sensor y concentración de ruido, lo obtenido se muestra en la **Figura 43** y se expone en la **Tabla 19.** Se observa que el criterio de separabilidad define el número de componentes a remover y el añadir mas componentes desmejora la respuesta del clasificador, factor que se evidencia por el promedio obtenido de los diferentes grupos de validación.

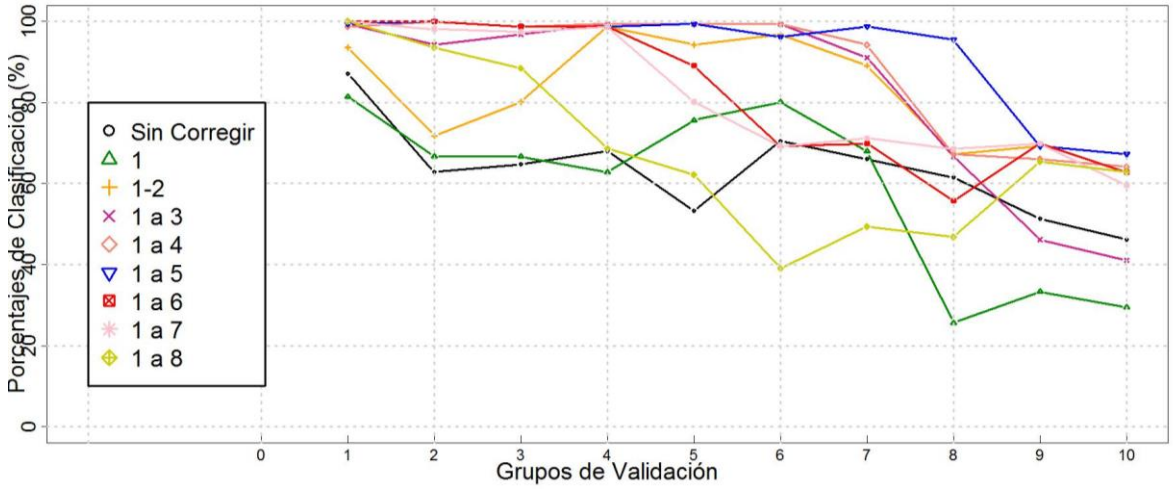


Figura 43. Gráfica comparativa de la remoción de componentes principales comunes en los datos de chemosensors con `dsd=4, csd=2, ssd=2` y `ndcomp=3`.

Grupo	Porcentaje de acierto								
	Sin remover componentes	C 1	C 1 y 2	C 1 a 3	C 1 a 4	C 1 a 5	C1 a 6	C1 a 7	C1 a 8
Entrenamiento	99,7	99,3	99,5	99,4	9,87	**99,2**	99,3	99,4	99,4
Validación 1	87,179	81,410	93,590	99,359	98,718	**99,359**	100,000	100,000	100,000
Validación 2	62,821	66,667	71,795	94,231	100,000	**100,000**	100,000	98,077	93,590
Validación 3	64,744	66,667	80,128	96,795	98,718	**98,718**	98,718	97,436	88,462
Validación 4	67,949	62,821	98,718	99,359	99,359	**98,718**	98,718	98,718	68,590
Validación 5	53,205	75,641	94,231	99,359	99,359	**99,359**	89,103	80,128	62,179
Validación 6	70,513	80,128	96,795	99,359	99,359	**96,154**	69,231	69,231	39,103
Validación 7	66,026	67,949	89,103	91,026	94,231	**98,718**	69,872	71,154	49,359
Validación 8	61,538	25,641	67,308	66,667	67,308	**95,513**	55,769	68,590	46,795
Validación 9	51,282	33,333	69,231	46,154	66,026	**69,231**	69,872	69,872	65,385
Validación 10	46,154	29,487	67,308	41,026	64,103	**67,308**	62,821	59,615	62,821
Promedio	63,1411	58,9744	82,8207	83,3335	88,7181	**92,307**	81,411	81,282	67,628
Criterio de separabilidad	20.997	5.410	1.981	1.965	1.390	**1.339**	1.578	1.425	1.410

Tabla 19. Resumen de los resultados obtenidos en el experimento 8, usando chemosensors con `dsd=4, csd=2, ssd=2` y `ndcomp=3`, al remover desde la primera hasta las 8 primeras componentes principales comunes.

En esta serie de experimentos con la base de datos sintéticos del paquete `chemosensors`, se logró observar que el efecto de remover las componentes principales comunes por el método de CC-CPCA, mejora claramente la respuesta del clasificador. El llamado **criterio de separabilidad** define el número de componentes a remover, tomando como estimador de parada en la remoción del número de componentes por CC-CPCA, la disminución o la aproximación de éste valor a cero. Cuando el **criterio de separabilidad** en lugar de continuar la disminución, se incrementa, indica que las componentes removidas en el punto donde se obtuvo el menor valor, son suficientes para mejorar la respuesta del sistema. En datos cuya contaminación por efecto de derivas y de ruido es alta, la respuesta del sistema al remover las componentes principales también es representativa, tal como se observó en los experimentos 7 y 8 (secciones 3.1.7 y 3.1.8 respectivamente), donde se pasa de un promedio de aciertos en los grupos de validación del 63% al 92% cuando se remueven las cinco primeras componentes principales comunes.

3.2 PRUEBAS CON LA BASE DE DATOS DE LA UNIVERSIDAD DE CALIFORNIA (SAN DIEGO)

Para esta base de datos se llevan a cabo pruebas o experimentos realizando corrección de las derivas por los métodos de CC-PCA y CC-CPCA. Se partió del análisis de resultados obtenidos sin corrección de derivas, con el propósito de escoger el mejor número o conjunto de datos a usarse en el entrenamiento, para ello se llevan los datos por lotes al módulo de clasificación sin realizar ninguna corrección previa a las derivas; lo anterior, con el propósito de determinar el número de lotes que contengan una componente adecuada de deriva, de modo que este conjunto de datos sea adecuadamente seleccionado.

En la **Figura 43**, se visualizan en color verde los resultados obtenidos tomando como conjunto de entrenamiento el lote 1 y los subconjuntos de validación son los 9 lotes subsiguientes; la línea de color negro presenta los porcentajes de acierto en la clasificación para los 8 grupos de validación cuando el entrenamiento es realizado con los dos primeros lotes de la base de datos. Cabe recordar que cada grupo de validación corresponde a uno de los **n** lotes restantes de San Diego, luego de seleccionados los lotes para el entrenamiento.

Se observa en la **Figura 43,** que la repuesta del sistema mejora al entrenar con los lotes 1 y 2, con respecto a la prueba de entrenamiento solo con el lote 1; además la marcada influencia de las derivas ocasiona que a partir de los lotes 6 y 7 la respuesta en porcentaje de clasificación del sistema se vea claramente afectada, dado que caen notoriamente los aciertos del predictor en estos últimos lotes hasta aproximadamente un 20% de aciertos.

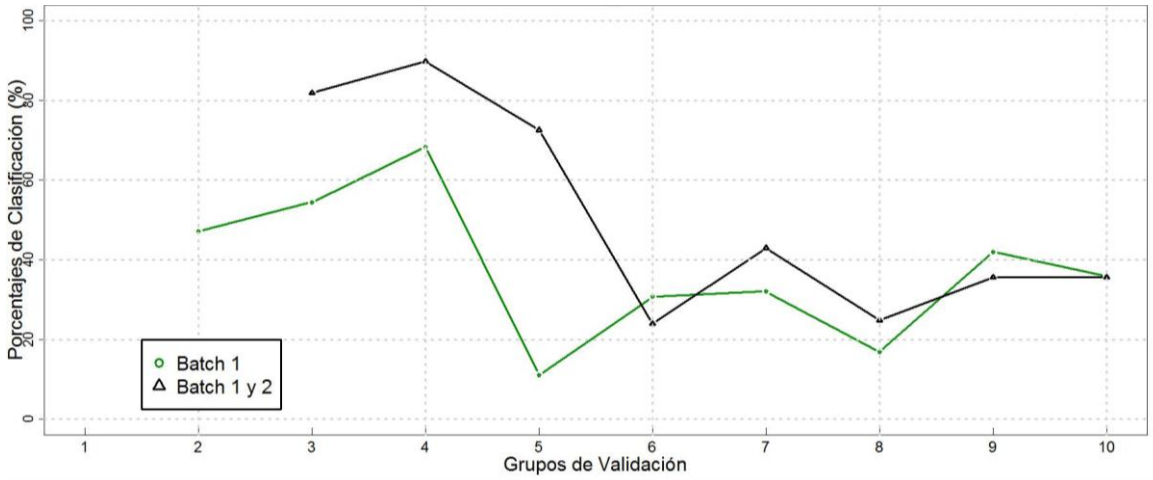


Figura 43. Análisis de la base de datos de San Diego por lotes, para escoger el conjunto de entrenamiento en los datos sin corrección de derivas.

En la **Tabla 20** se muestran de forma cuantitativa los resultados mostrados en la **Figura 43.** Las celdas que contienen la letra **T** indican que el lote fue usado como conjunto de entrenamiento.

NÚMERO DE LOTES USADOS PARA EL ENTRENAMIENTO	LOTES DE USADOS PARA LA VALIDACIÓN										PORCENTAJE PROMEDIO
	1	2	3	4	5	6	7	8	9	10	
1	**T**	48,2	55,4	68,7	13,1	32,2	33,1	18,3	43,4	37,8	38,99
2	**T**	**T**	81,9	91,9	73,3	24,9	43,9	26,9	38,9	37,0	52,27

Tabla 20. Porcentajes de acierto en los conjuntos de validación de San Diego, sin realizar ningún tratamiento de corrección de derivas a los datos.

Con el propósito de estudiar el efecto de la deriva en la separación de los grupos para la adecuada clasificación, se realiza PCA a los datos de entrenamiento y se observa la separabilidad de las 6 clases de gases. Asimismo, esta visualización de los datos en el espacio de representación PCA permite inferir acerca de que tan marcado se encuentra el efecto de la deriva en los datos analizados.

En la **Figura 44** se presentan las dos primeras componentes del PCA (PC1 Vs PC2), realizado a los datos del Lote 1. En esta representación se observa que la separabilidad de las clases aún es buena, con excepción de los grupos que corresponden a los gases F, D, E, que tienden a traslaparse en el espacio donde los gases poseen menor concentración. Se aclara que el comportamiento creciente que se observa en los otros gases corresponde a las diferentes concentraciones que se usaron para las mediciones de cada clase de gas, lo cual marca esa tendencia en cada una de ellas. En consecuencia, el sistema debe ser tan robusto como para discriminar entre las seis clases de gases independientes de la concentración en la que se ha medido y además no puede confundir el efecto de la concentración con el efecto de la deriva.

La información en San Diego no contiene los datos de concentración en cada gas, por lo que las clases de gases medidos, como se explicó anteriormente, corresponden a 6 diferentes gases independientes de la concentración en la que se encuentren. Esta importante característica ocasiona que el análisis de esta clase de datos sean tan importante para entrenar un sistema de reconocimiento de olores con gran capacidad de generalización.

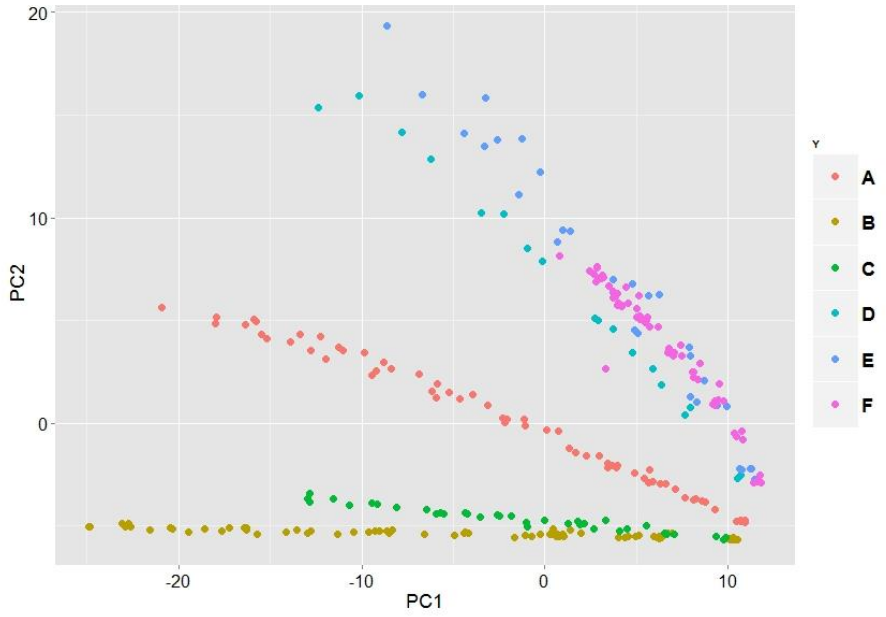


Figura 44. PCA realizado en el Lote 1, los datos de San Diego no poseen corrección de deriva. Se aplica la normalización como técnica de pre-procesado.

A continuación en la **Figura 45,** se plasma de la misma manera que en el caso anterior, el PCA visualizando las dos primeras componentes, pero en este caso se leen los datos de los lotes 1 y 2.

Se observa claramente en la **Figura 45,** como los datos de algunos de los gases empiezan a tener un efecto marcado por las derivas, tal es el caso de los gases D y E, cuyas representaciones en el espacio PCA evidencian un corrimiento hacia el eje inferior de su correspondiente representación. En el caso de los gases A y B también se observa mayor dispersión en las observaciones nuevas; estos efectos de corrimiento mencionados causan que la separabilidad de las clases se desmejore, a causa del traslape que comienza a ocurrir entre ellas por la intromisión de las derivas.

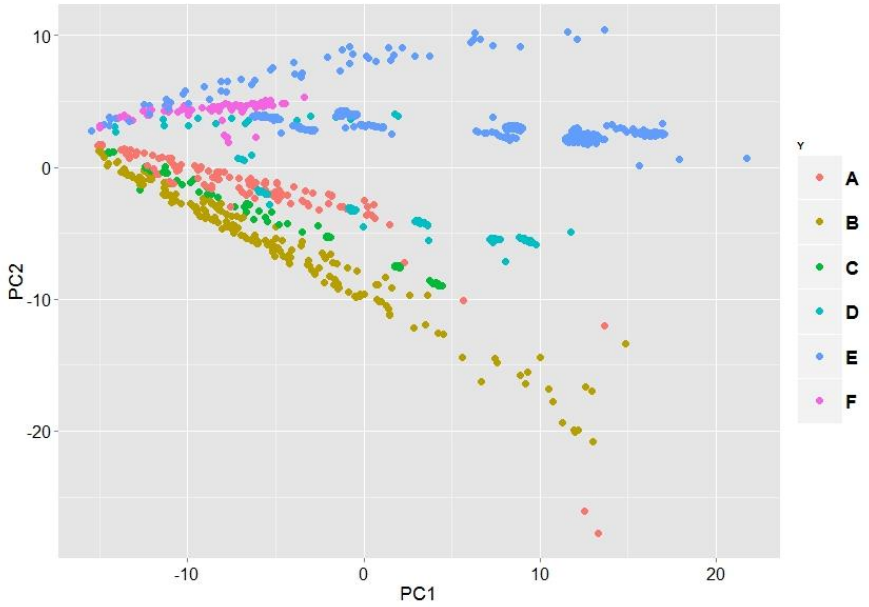


Figura 45. PCA realizado en los Lotes 1 y 2, los datos de San Diego no poseen corrección de deriva. Se aplica la normalización como técnica de pre-procesado antes de hacer el PCA.

En consecuencia, a partir de estas dos representaciones de PCA del lote 1 y de los lotes 1 y 2 visualizadas en las **Figuras 43, 44** y **45,** se resolvió tomar como grupo de entrenamiento al conjunto conformado por los lotes 1 y 2, en donde ya se evidencia la presencia significativa de derivas.

Se observa en la **Figura 45,** el efecto marcado de la deriva en casi la totalidad de los gases, degradando por completo el sistema. En cuanto al gas F, que corresponde a Tolueno al no poseer gran cantidad de medidas realizadas en este periodo de tiempo, se ve menos influenciado, esto es debido a que ninguna medición de este gas se realizó desde el mes 4 hasta el mes de 19, lo cual ocasiona que la corrección de derivas en este gas no tenga el impacto esperado por no existir una cantidad de medidas representativas de este gas en el conjunto usado para entrenamiento. El análisis a desarrollarse en la etapa de resultados emplea la estructura por lotes, dejando para entrenamiento los dos primeros lotes y la validación incluye los lotes del 3 al 10.

3.2.1 Corrección de derivas aplicando la técnica CC-PCA

Luego de escoger el número de lotes a usarse como conjunto de entrenamiento, se realiza el proceso de corrección de la deriva a partir del método de Corrección de Componentes por Análisis de Componentes Principales (CC-PCA), que consiste en seleccionar un gas de referencia de tal forma, que a partir de este gas se logre extraer la componente de deriva

asociada a las mediciones. Se realiza un experimento con cada uno de los 6 clases de gases A, B, C, D, E, F, dado que la metodología propuesta por (Arthurson, y otros, 2000) propone extraer la componente de corrección de la deriva a partir de un gas de referencia, en este caso el que mejor represente o tenga mayor información de ella.

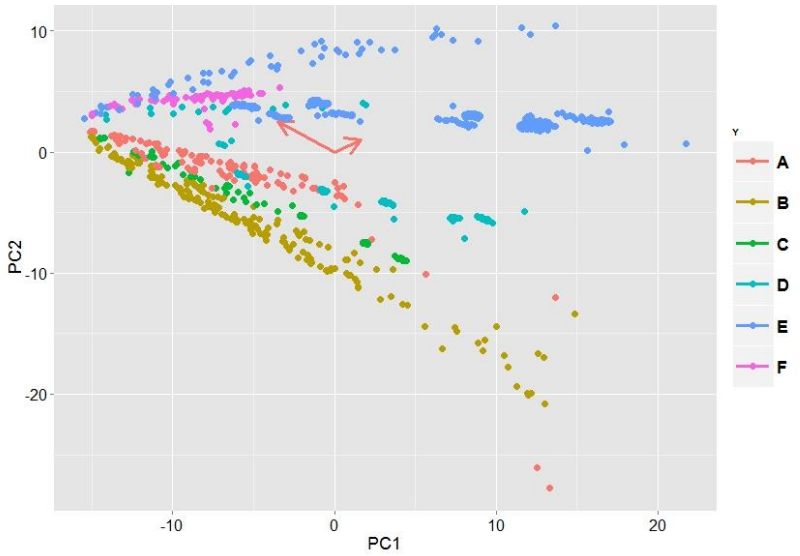


Figura 46. Espacio de representación PCA (CP1-CP2), de los datos de entrenamiento sin corrección de deriva, sobre el cual se proyectan los vectores de las dos primeras componentes comunes extraídas del gas de referencia A.

En la **Figura 46,** se presenta la dirección de las dos primeras componentes principales extraídas del espacio de representación PCA del gas de referencia A, y se han proyectado sobre el PCA del conjunto de datos de entrenamiento sin corrección de la deriva. Se identifican en la gráfica estas dos componentes como las flechas que tienen el mismo color de la observaciones del gas que se usó como referencia (en este caso Gas A). Obsérvese que la primera componente corresponde al vector de mayor magnitud e indica la dirección de deriva en el gas de referencia, ésta será restada según el método de Arthurson de todas las clases de gases de la base de datos.

En la **Figura 47,** se presenta el resultado de los datos de entrenamiento (lotes 1 y 2), llevados al espacio PCA (Componentes 1 y 2), luego de realizar la corrección de derivas con el gas de referencia A. En este caso, el gas A no es el que mejor representa la deriva en el conjunto de entrenamiento, situación que es claramente evidente al comparar el PCA mostrado en la **Figura 45** que representa el mismo conjunto pero sin corrección alguna de las derivas, razón por la que la corrección de componentes no logra efectos muy marcados en la separabilidad de las 6 clases de gases en el espacio PCA.

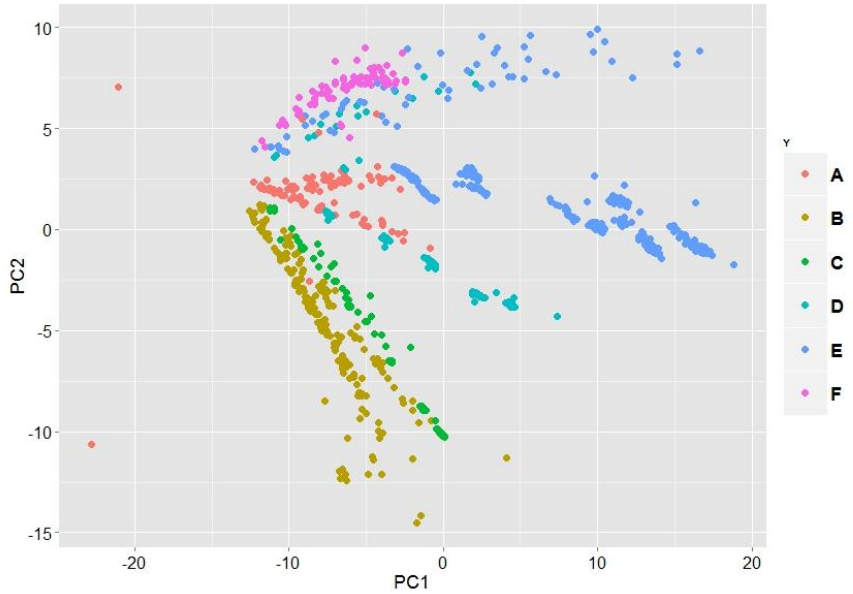


Figura 47. Espacio de representación PCA (CP1-CP2), de los datos de entrenamiento a los que se les aplicó corrección de componentes tomando como gas de referencia el gas A.

La **Figura 48**, muestra la varianza acumulada en cada una de las diez primeras componentes principales del PCA del conjunto de datos para el gas de referencia A, con respecto a la varianza en los datos originales normalizados. Se evidencia que el porcentaje de varianza acumulada se centra en la primera componente, sin embargo, éste porcentaje acumulado es de tan solo un 38% aproximado, lo que ratifica porque este gas no es representativo en la corrección de derivas.

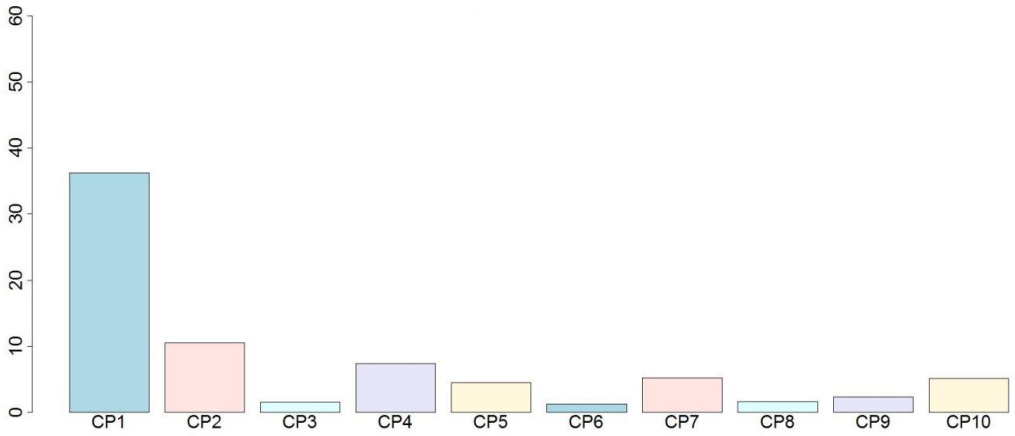


Figura 48. Diagrama de barras que representa la varianza acumulada en cada componente principal del gas de referencia A con respecto a los datos de entrenamiento originales normalizados. (Se grafican solo las 10 primeras componentes).

En la **Figura 49,** se presenta el espacio de representación PCA para los datos de entrenamiento; el gas tomado como referencia en este experimento es el gas B, se observa en este caso, que el gas de referencia en el espacio PCA tiende a tener un comportamiento más gaussiano con respecto a lo que se observa en la **Figura 45,** lo cual se infiere por la forma en que se agrupan las medidas de esa clase de gas con respecto a la representación de este mismo gas en el espacio PCA, este efecto causa que el gas C y gran parte de las muestras del gas A se encuentren adecuadamente separadas. En contraparte, aún se observa poca separación de clases entre los gases D, E, y F, lo cual no permite optimizar los resultados.

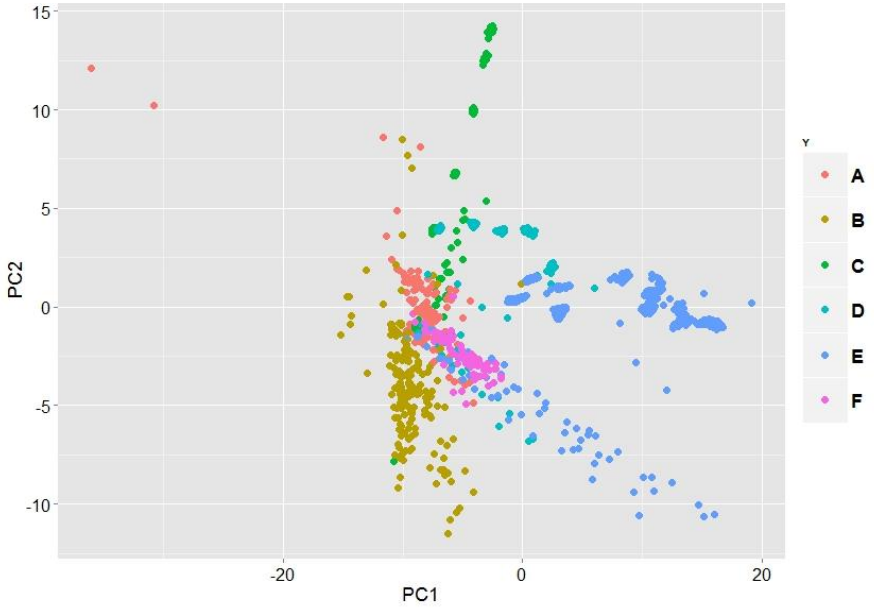


Figura 49. Espacio de representación PCA (CP1-CP2), de los datos de entrenamiento a los que se les aplicó corrección de componentes tomando como gas de referencia el gas B.

Al observar en la **Figura 50** la varianza acumulada en la primera componente, se concluye que ésta es adecuadamente significativa frente a las demás componentes que la conforman. Alcanza el 47.1%, mientras que en las otras componentes no logran exceder el 8%, lo anterior permite predecir el buen comportamiento de la corrección de componentes usando este gas como el de referencia.

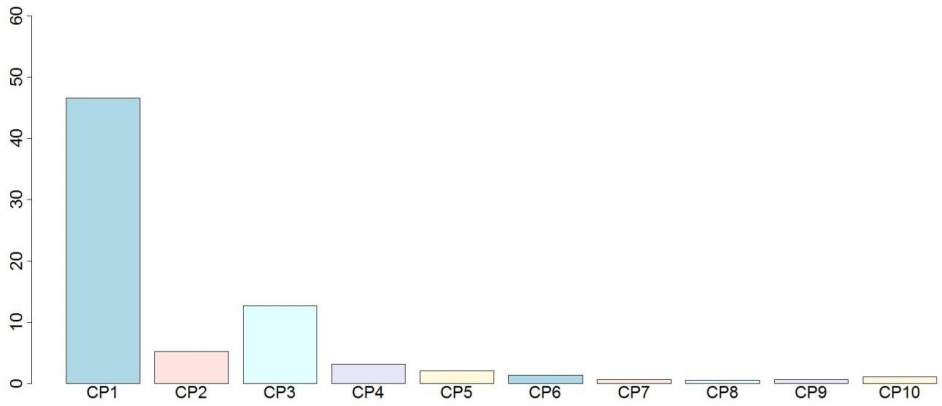


Figura 50. Diagrama de barras que representa la varianza acumulada en cada componente principal del gas de referencia B con respecto a los datos de entrenamiento originales normalizados. (Se grafican solo las 10 primeras componentes).

Al usar el gas C como gas de referencia, la representación PCA, arroja la evidente mejora en la gausianidad de los datos que corresponden a esta clase de gas, pero la separabilidad de las otras clases no mejora, además se siguen observando conjuntos de datos traslapados, por lo tanto con esta clase de gas no se predice un buen comportamiento en la etapa de clasificación.

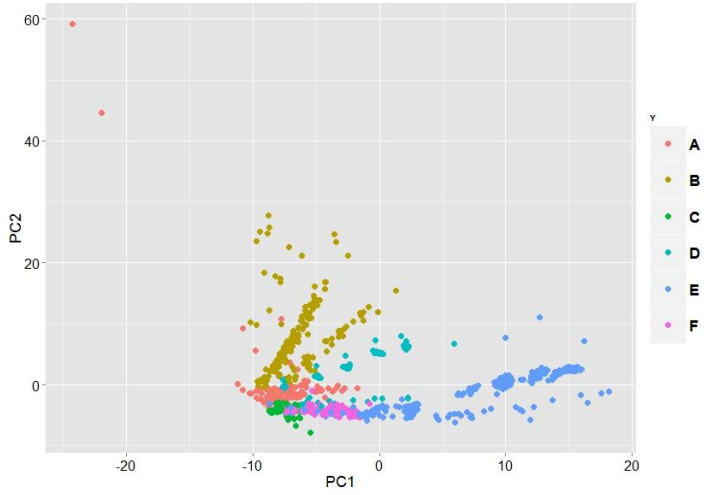


Figura 51. Espacio de representación PCA (CP1-CP2), de los datos de entrenamiento a los que se les aplicó corrección de componentes tomando como gas de referencia el gas C.

El diagrama de barras para el gas de referencia C mostrado en la **Figura 52,** refleja que la primera componente logra capturar una varianza en los datos, que en este caso es cercana al 45%. De acuerdo a lo anterior los resultados obtenidos en las dos primeras componentes del PCA son como se observan en la **Figura 51.**

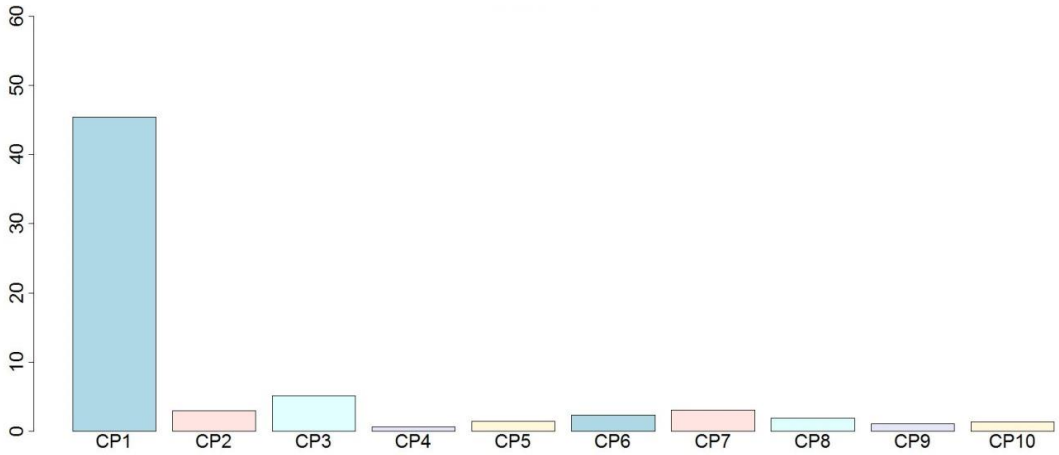


Figura 52. Diagrama de barras que representa la varianza acumulada en cada componente principal del gas de referencia C con respecto a los datos de entrenamiento originales normalizados. (Se grafican solo las 10 primeras componentes).

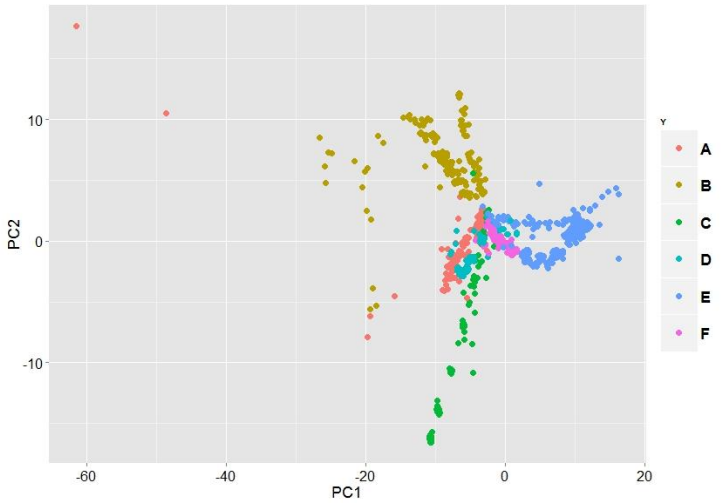


Figura 53. Espacio de representación PCA (CP1-CP2), de los datos de entrenamiento a los que se les aplicó corrección de componentes tomando como gas de referencia el gas D.

El espacio de representación obtenido para la corrección de los datos de entrenamiento tomando al gas D como el de referencia (**Figura 53**), evidencia que los gases de clases F y D son las dos clases de gases que se encuentran más traslapados y, en especial el gas D es el que más contamina a las otras clases de gases, aún cuando en este caso, éste gas es el gas de referencia, no logra mejorar su comportamiento al hacer la corrección.

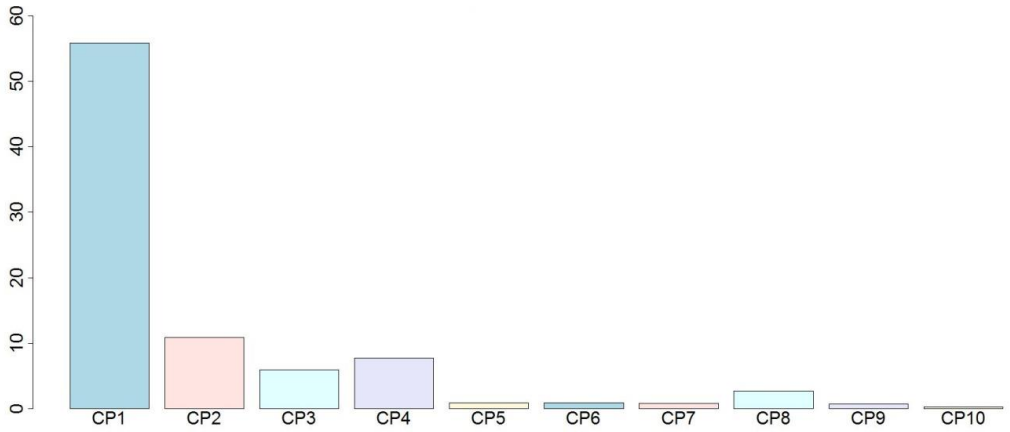


Figura 54. Diagrama de barras que representa la varianza acumulada en cada componente principal del gas de referencia D con respecto a los datos de entrenamiento originales normalizados. (Se grafican solo las 10 primeras componentes).

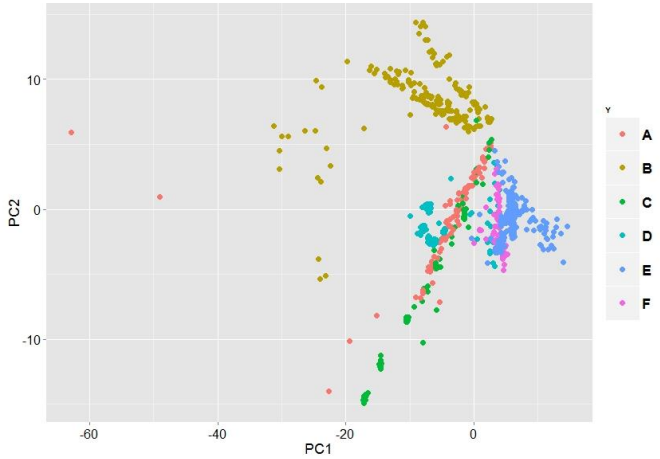


Figura 55. Espacio de representación PCA (CP1-CP2), de los datos de entrenamiento a los que se les aplicó corrección de componentes tomando como gas de referencia el gas E.

El espacio PCA de la **Figura 55** para el gas de referencia E, de la misma forma que en el caso anterior logra mejorar la separabilidad de los gases A y B pero aún las otras clases se traslapan y se repite nuevamente el caso de la superposición de los gases D y F. La varianza acumulada observada en la **Figura 56,** en este caso es cercana al 60%.

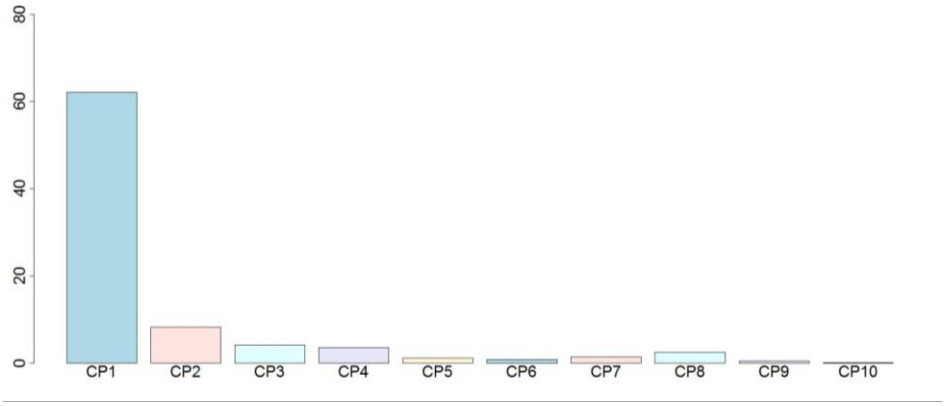


Figura 56. Diagrama de barras que representa la varianza acumulada en cada componente principal del gas de referencia E con respecto a los datos de entrenamiento originales normalizados. (Se grafican solo las 10 primeras componentes).

Finalmente, tomando como gas de referencia al gas F, como se esperaba, la mejora con respecto a los datos sin corrección de las derivas es pobre, esto es debido a que las medidas de este gas son las menos contaminadas con las derivas a causa de ser el gas que fue medido solo en el mes 2 del periodo de tiempo que involucra los datos del conjunto de entrenamiento que corresponde a 10 meses.

La baja influencia de las derivas en esta clase de gas se evidencia en la **Figura 57,** que comparada con la **Figura 45,** donde se presenta el PCA de los datos no corregidos, resultan ser ligeramente similares en el comportamiento de las agrupaciones de los gases. La **Figura 58,** muestra las varianzas acumuladas para cada componente tomando como referencia esta clase de gas.

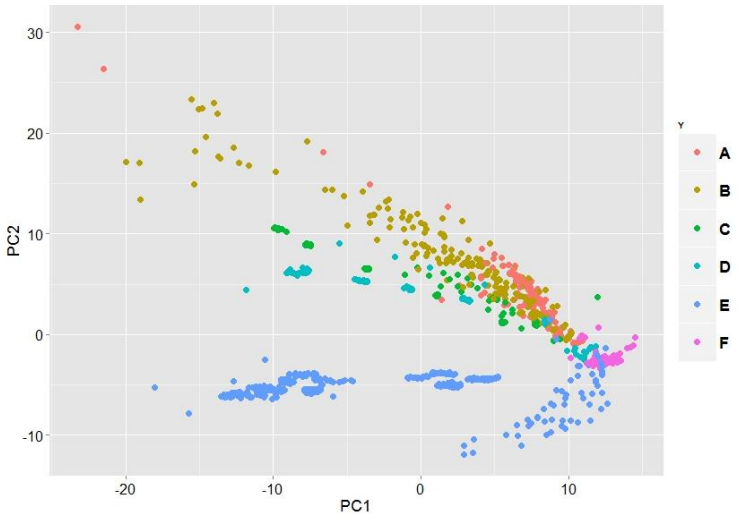


Figura 57. Espacio de representación PCA (CP1-CP2), de los datos de entrenamiento a los que se les aplicó corrección de componentes tomando como gas de referencia el gas F.

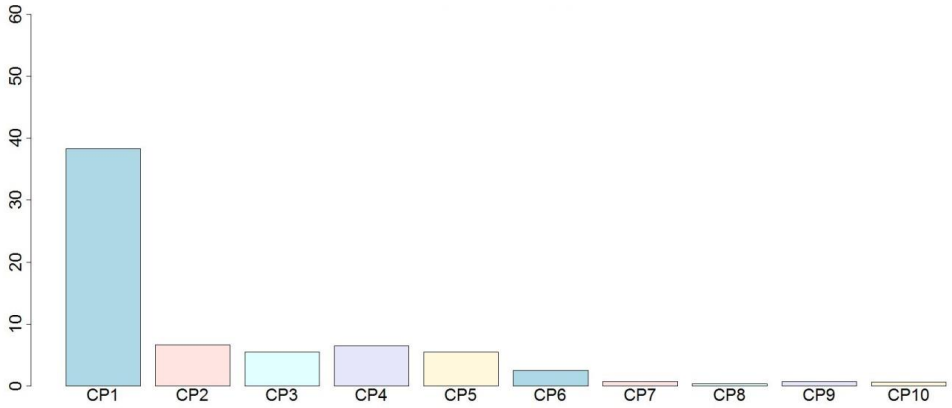


Figura 58. Diagrama de barras que representa la varianza acumulada en cada componente principal del gas de referencia F con respecto a los datos de entrenamiento originales normalizados. (Se grafican solo las 10 primeras componentes).

En la **Tabla 21** se incluyen los porcentajes de varianza acumulada en las 10 primeras componentes principales con respecto a cada gas de referencia, observándose que el gas que posee mayor varianza en la primera componente corresponde al gas E, por cuanto la corrección de las derivas tomando este gas como el de referencia hace que esta clase de gas genere una mejor respuesta en la agrupación de los datos.

COMPONENTES	GAS DE REFERENCIA- VARIANZA ACUMULADA %					
	gas A	gas B	gas C	gas D	gas E	gas F
CP1	36,2	46,6	45,4	55,8	**62,1**	38,3
CP2	10,5	5,2	2,9	10,9	8,2	6,6
CP3	1,6	12,7	5,2	5,9	4,2	5,5
CP4	7,4	3,2	0,6	7,7	3,6	6,5
CP5	4,5	2,1	1,4	0,9	1,2	5,5
CP6	1,2	1,3	2,4	0,9	0,9	2,5
CP7	5,2	0,6	3,0	0,8	1,4	0,8
CP8	1,6	0,5	1,9	2,7	2,5	0,3
CP9	2,3	0,6	1,1	0,7	0,5	0,8
CP10	5,2	1,1	1,3	0,3	0,1	0,7

Tabla 21. Porcentajes de varianza acumulada en las diez primeras componentes principales del gas de referencia con respecto a los datos originales de entrenamiento.

3.2.2 Validación del método de CC-PCA usando clasificador k-NN.

Con el objeto de verificar la mejora introducida en la corrección de la deriva usando el método del gas de referencia, a continuación se valida en los 8 lotes de la base de datos San Diego disponibles para tal fin. Se hace uso de un clasificador k-NN con sintonización del mejor `k`, además tal como se describió en el capítulo del diseño experimental (**sección 2.3.3),** se hace uso de la primera componente principal del gas de referencia encontrada en los datos de entrenamiento para, con ésta misma, corregir los datos de los subconjuntos de validación, igualmente la normalización se realiza a partir de la media y la desviación estándar de los datos del entrenamiento.

Al realizar este proceso de validación en los 8 lotes, se obtienen los resultados mostrados en la **Figura 59,** cuyos valores se detallan de forma cuantitativa en la T**abla 22**. En ésta se comparan los resultados obtenidos con el método CC-PCA en cada gas de referencia contra los resultados obtenidos en los mismos grupos de validación cuando no se aplicó ningún método de mitigación de las derivas.

En la F**igura 59** la línea de color negro que corresponde a los datos no corregidos, no logra ser superada en su totalidad por ninguno de las correcciones efectuadas con cada uno de los gases tomados como gas de referencia, tan solo en los lotes 6 y 8 los resultados con cada uno de los gases de referencia logran estar por encima de los resultados no corregidos. Lo anterior concuerda con lo expresado en (Ziyatdinov, y otros, 2009), donde se indica que componentes de ruido en el sistema hacen que los datos corregidos sean en algunos casos inferiores a los datos no corregidos.

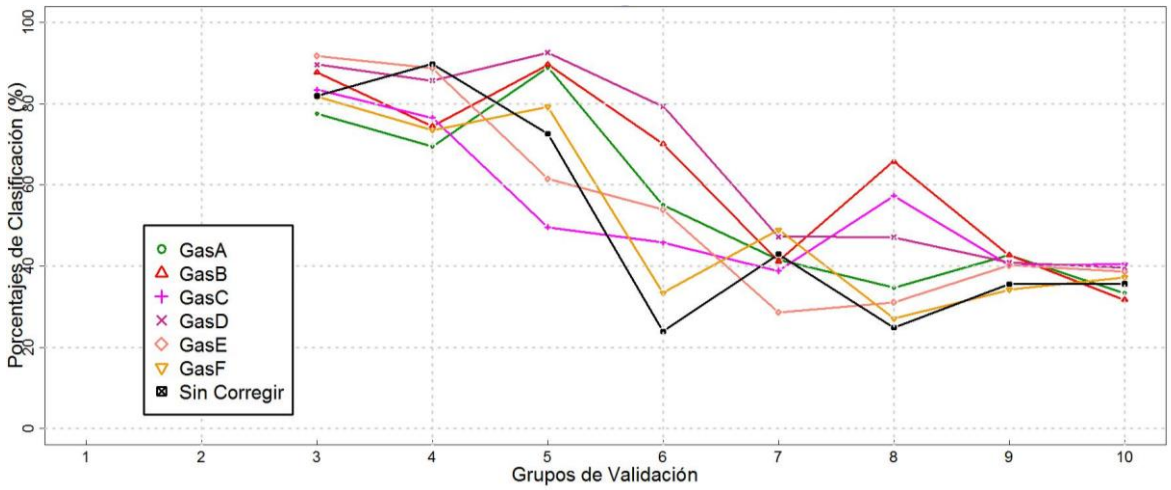


Figura 59. Resultados de validación aplicando el método de CC-PCA con los 6 gases de referencia, comparado con los resultados obtenidos sin corrección de derivas (línea de color negro).

LOTE	GAS DE REFERENCIA						SIN CORREGIR
	Gas A	Gas B	Gas C	Gas D	Gas E	Gas F	
1	ENTRENAMIENTO						
2	ENTRENAMIENTO						
3	77,6	87,7	83,4	85,6	91,8	81,8	81,9
4	69,4	74,5	76,5	84,7	88,8	73,5	89,8
5	88,9	89,6	49,6	90,6	61,5	79,3	72,6
6	55,0	70,1	45,8	78,4	53,9	33,4	24,0
7	41,6	41,2	38,8	47,2	28,6	48,9	43,0
8	34,7	65,8	57,3	45,1	31,1	27,1	24,9
9	42,7	42,7	40,2	40,9	40,2	34,2	35,6
10	33,4	31,7	40,6	39,7	38,6	37,2	35,6
PROMEDIO	55,4	62,9	54,0	**64,2**	54,3	51,9	50,9

Tabla 22. Resultados obtenidos en la validación de los lotes de *San Diego* con el método de CC-PCA, usando cada uno de los gases como el gas de referencia y comparando con los datos no corregidos.

Como era de esperarse, el gas que presenta el más bajo desempeño es el gas F ya que como se explicó con anterioridad es el que menor deriva tenía incorporada en el periodo de tiempo usado para entrenamiento; igualmente éste fue el gas que más efecto de solapamiento tuvo con las otras clases de gases en la representación del espacio PCA.

Se concluye, según los promedios obtenidos en los 8 grupos de validación, que el gas que acumula mayor varianza en su primera componente, es decir el gas **E,** no determina para el sistema la mejor representación al momento de extraer las derivas; por el contrario, el mejor promedio de porcentajes de clasificación lo obtuvo el gas de referencia **D**, por lo tanto, este método de selección del mejor gas de referencia, aunque mejora los resultados de clasificación en un 14% aproximado con respecto a los datos no corregidos, se torna complejo en la escogencia del mejor de ellos.

3.2.3 Corrección de derivas aplicando la técnica CC-CPCA

Esta sección presenta un análisis de la corrección a las derivas usando el método CC-CPCA. En este caso nuevamente se entrena con los lotes 1 y 2 y se valida con los 8 lotes subsiguientes en concordancia con la metodología planteada. Esta metodología, toma la matriz **X** de características, que en este caso serán los datos del entrenamiento y a ella se le aplica una diagonalización conjunta, encontrándose una matriz de 128x128 dimensiones que contiene los eigenvectores que representan las componentes principales comunes de la matriz original. El propósito es encontrar una o varias componentes comunes a todas las clases que representen adecuadamente la deriva en las 6 clases de gases.

Por medio de este trabajo se realizó una búsqueda estadística de las **n** componentes a ser removidas, dando un tratamiento no lineal a la corrección de derivas al estimar la substracción de más de una componente. De la misma forma que en `chemosensors`, se utilizó el **criterio de separabilidad** para determinar el mejor comportamiento de los datos al remover un mayor número de componentes. El número de componentes a elegir correspondió a aquellas que generan el valor del **criterio de separabilidad** más cercano a cero antes de empezar a incrementarse nuevamente.

Luego de computar la matriz de diagonalización **E**, se obtiene la varianza capturada en cada una de las componentes con el fin de realizar el análisis inicial de cuáles componentes substraer para la corrección de deriva. En la ecuación 21, se muestra la expresión que indica que la varianza se calcula como la varianza proyectada por la matriz **E** de diagonalización, con respecto a la varianza total proyectada por el conjunto de datos del entrenamiento.

$$var = \frac{Varianza\ proyectada_{(E)}}{Varianza\ total_{Xtrain}} \qquad (21)$$

En la **Figura 60**, se muestra el diagrama de barras con los porcentajes de varianza acumulada en los 16 primeros componentes y de la misma manera en la **Tabla 23,** se muestran de forma cuantitativa estos porcentajes. En la inspección de la gráfica es notable que las componentes que mayor varianza acumulan, en orden descendente, son la componente 1, la componente 5, la componente 2, 3 y la 4.

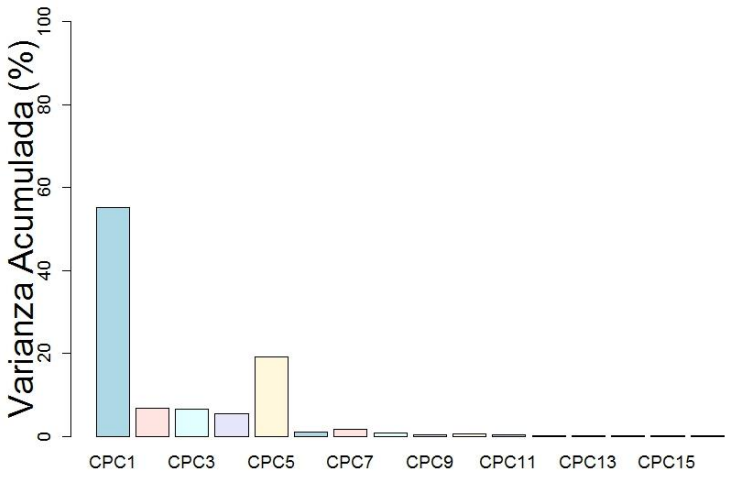


Figura 60. Porcentajes de varianza acumulados en las componentes 1 a la 16 resultantes de la diagonalización, con respecto a la que se proyecta por los datos de entrenamiento antes de diagonalizar.

Este análisis es importante porque conlleva a realizar una búsqueda sobre cuáles de esas componentes están contribuyendo realmente a la deriva y cuáles en realidad guardan información relevante. Se resalta de nuevo, que esta base de datos posee las medidas de los 6 gases en diferentes concentraciones, las cuales se desconocen, esto puede de cierta manera hacer que las derivas estén subyacentes en la información útil para el sistema de clasificación.

Las diez primeras componentes principales comunes obtenidas se organizan en la **Tabla 23** con el fin de poder seleccionar solo aquellas que acumulan la mayor cantidad de varianza. Éstas 10 primeras componentes acumulan un porcentaje de varianza equivalente al 97,3% de la varianza total, por lo tanto se consideran suficientes para realizar las pruebas con la técnica de corrección de componentes.

Las primeras componentes obtenidas son las que representan la deriva en el sistema, por lo tanto, se inició la búsqueda seleccionando solo la componente CPC1 que contiene una varianza del 55,13%; esta selección de la primera componente fue desarrollada en (Ziyatdinov, y otros, 2009), de tal forma que en este trabajo se seleccionó un mayor número de componentes para demostrar que esto ocasiona mejoras en el sistema de clasificación al tratar la deriva en su naturaleza no lineal.

VARIANZA EN LAS 10 PRIMERAS COMPONENTES PRINCIPALES COMUNES	
CPC1	55,13
CPC2	6,72
CPC3	6,51
CPC4	5,42
CPC5	19,19
CPC6	0,98
CPC7	1,67
CPC8	0,84
CPC9	0,24
CPC10	0,59
Total Varianza Acumulada=	**97,30**

Tabla 23. Varianza acumulada en las 10 primeras componentes principales comunes resultantes de aplicar la diagonalización conjunta a los datos de entrenamiento (Lotes 1 y 2).

De acuerdo a la metodología propuesta por (Ziyatdinov, y otros, 2009), la remoción de la primera componente principal común, es decir, la que acumula la mayor cantidad de varianza contribuye a la disminución de la deriva.

En las **Figuras 61** a la **65**, se muestran los resultados obtenidos al graficar la primera y segunda componente principal en el espacio de representación PCA de los datos corregidos a partir de las pruebas realizadas sobre el conjunto de entrenamiento, al extraer 1, 2 o hasta 6 componentes principales comunes en este grupo de datos. Estas graficas PCA (componente 1 Vs. componente 2) comparadas con la F**igura 45,** evidencian una mayor separabilidad de las clases.

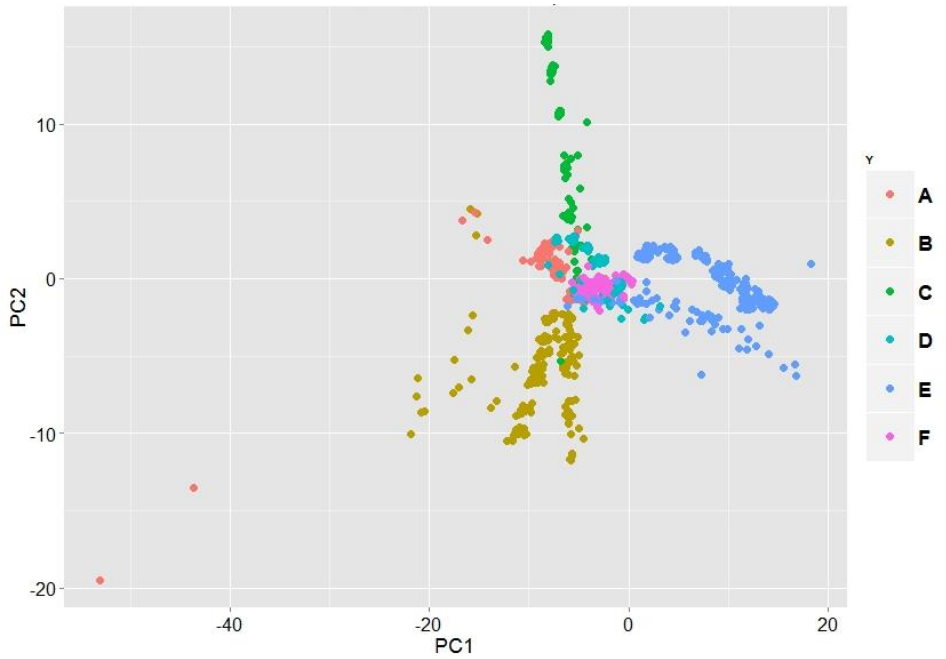


Figura 61. PCA del conjunto de entrenamiento (Lotes 1 y 2) removiendo la componente 1 por el método de CC-CPCA.

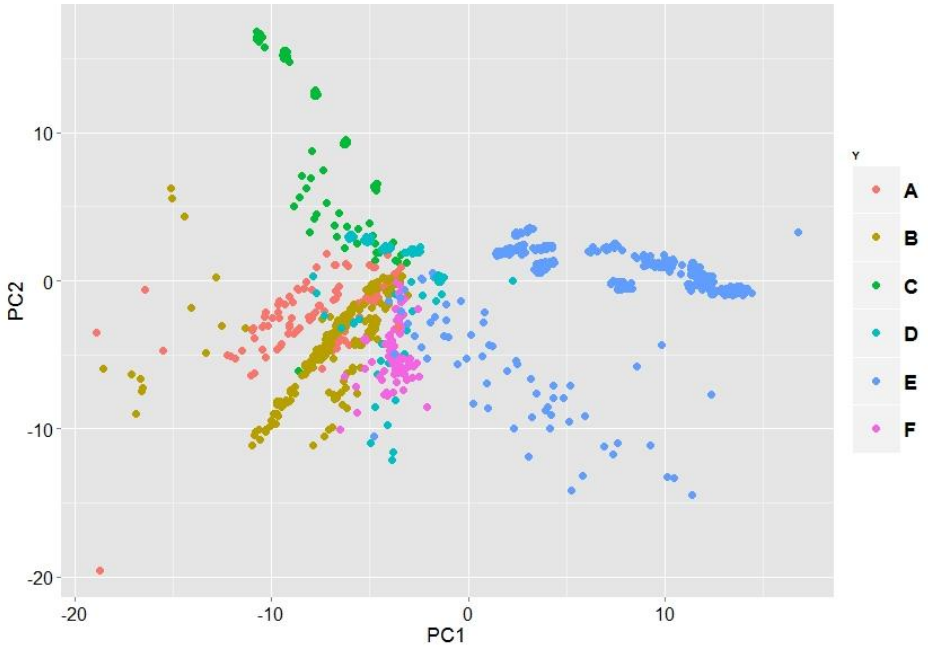


Figura 62. PCA del conjunto de entrenamiento (Lotes 1 y 2) removiendo las componentes 1 y 2 por el método de CC-CPCA.

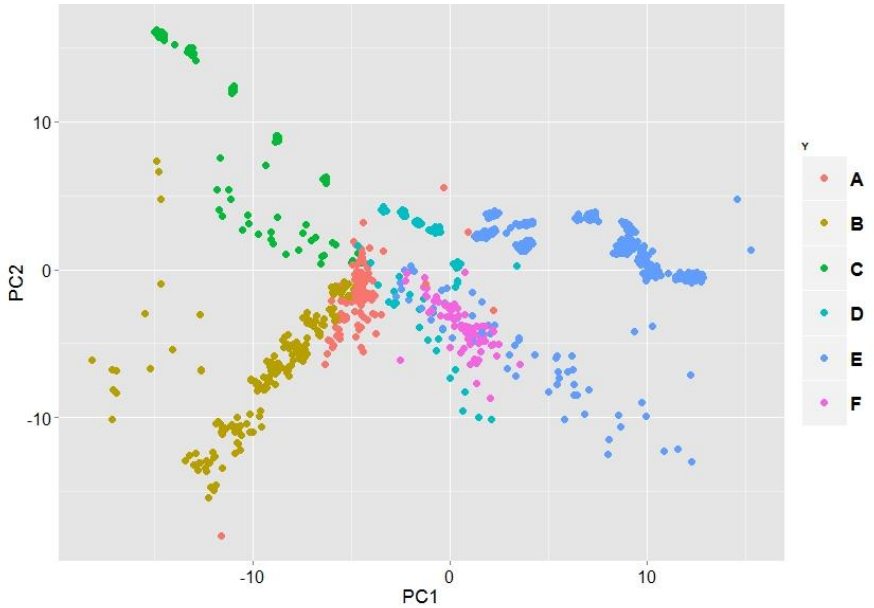


Figura 63. PCA del conjunto de entrenamiento (Lotes 1 y 2) removiendo las componentes 1,2,3 por el método de CC-CPCA.

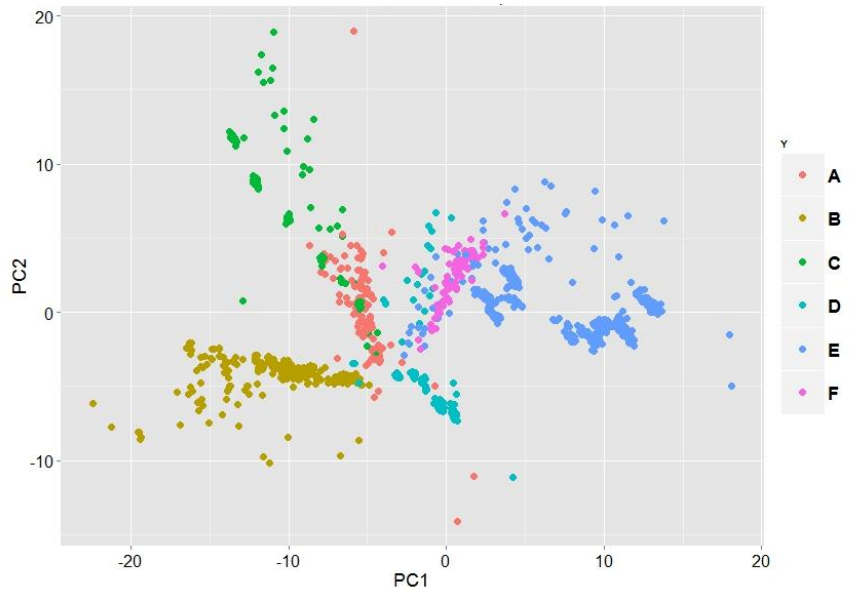


Figura 64. PCA del conjunto de entrenamiento (Lotes 1 y 2) removiendo las componentes 1,2,3,4 por el método de CC-CPCA.

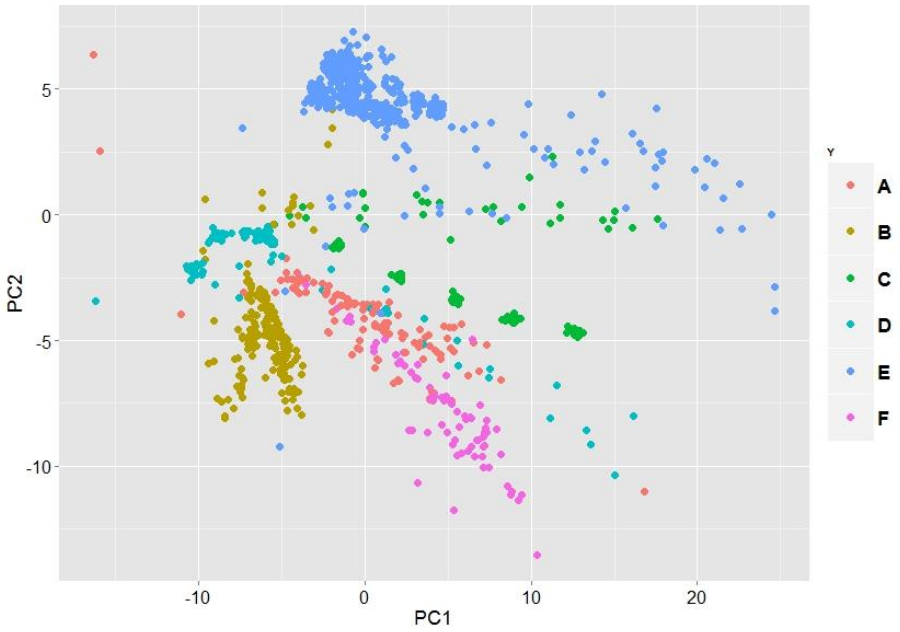


Figura 65. PCA del conjunto de entrenamiento (Lotes 1 y 2) removiendo las componentes 1,2,3,4,5 por el método de CC-CPCA.

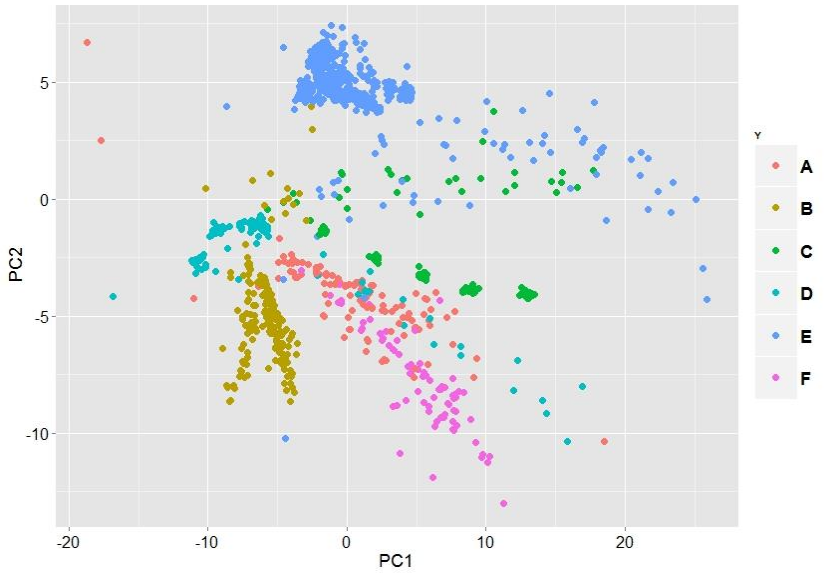


Figura 66. PCA del conjunto de entrenamiento (Lotes 1 y 2) removiendo las componentes 1,2,3,4,5,6 por el método de CC-CPCA.

A continuación se muestran los resultados obtenidos al comparar el análisis PCA sobre uno de los lotes de validación antes y después de la remoción de componentes por el método de CC-CPCA. Se presentan en las **Figuras 66** a la **70** el PCA (componente 1 Vs. componente 2) del lote de prueba 5 sin remoción de derivas y con corrección de derivas por el método CC-CPCA.

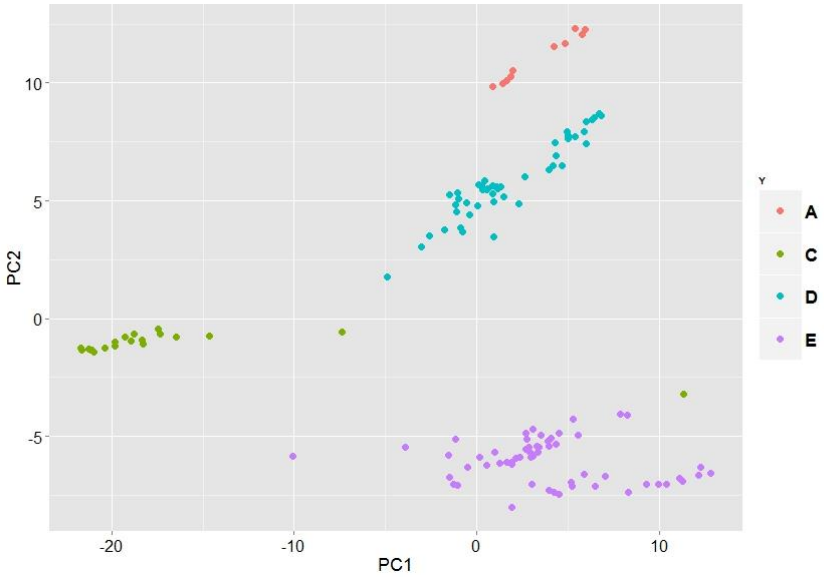


Figura 67. PCA del conjunto de prueba (Lote 5) sin tratamiento de derivas.

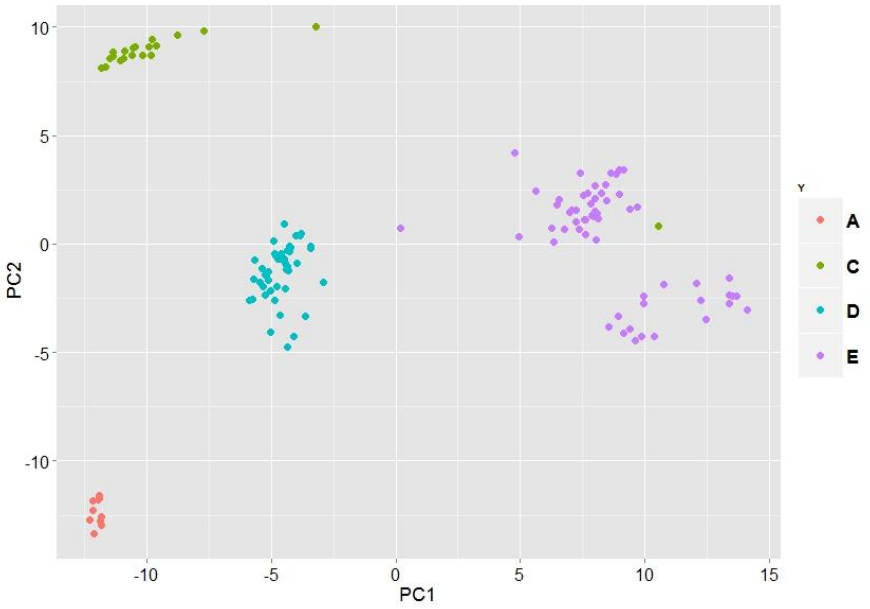


Figura 68. PCA del conjunto de prueba (Lote 5) con la componente 1 removida por el método CC-CPCA.

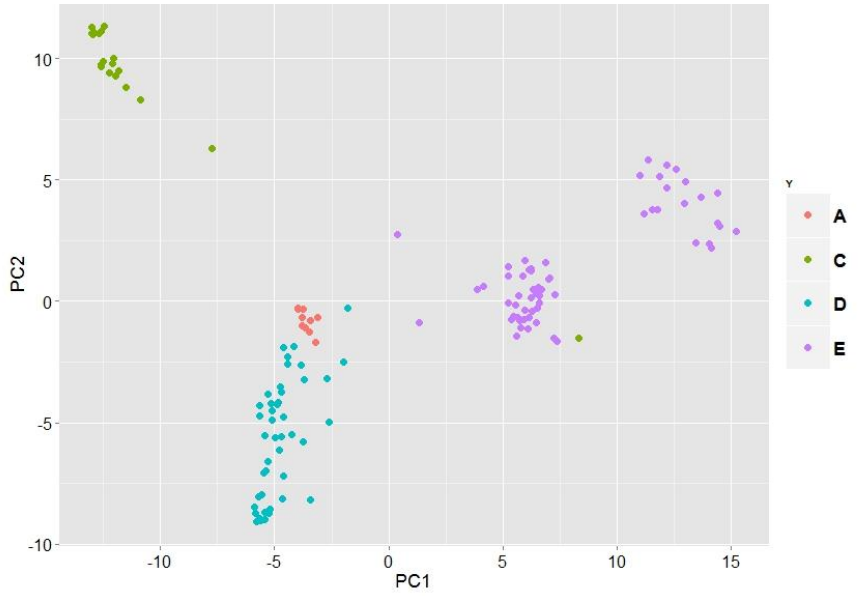


Figura 69. PCA del conjunto de prueba (Lote 5) con las componentes 1,2,3 removidas por el método CC-CPCA.

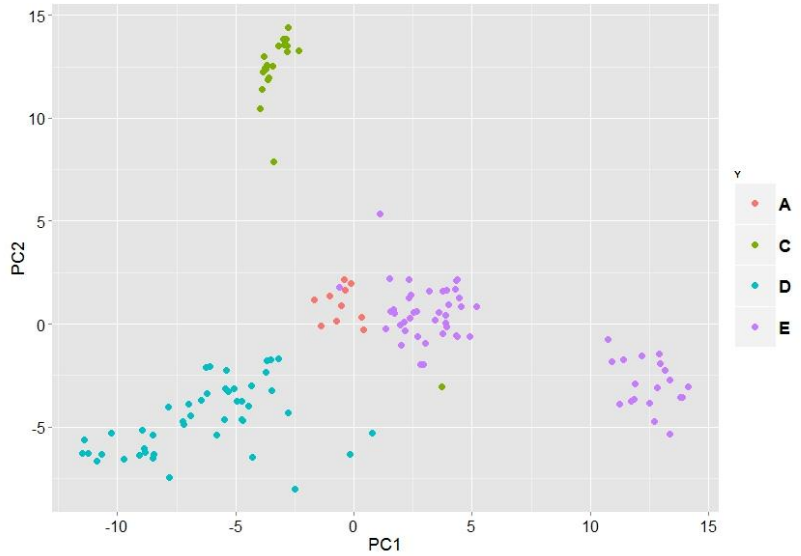


Figura 70. PCA del conjunto de prueba (Lote 5) con las componentes 1,2,3,4,5 removidas por el método CC-CPCA.

Se observa en las **Figuras 66** a **70** como la remoción de la primera componente principal es la que evidencia mejoras en la separabilidad de las clases en todos los grupos de gases; en las figuras presentadas se toman como ejemplo, las gráficas que corresponden a la remoción de la componente 1, las componentes 1, 2 y 3 y la remoción en conjunto de las componentes 1, 2, 3, 4 y 5, sin embargo, el análisis realizado involucra la remoción de hasta las seis primeras componentes principales comunes.

Aplicando el **criterio de separabilidad** planteado en esta investigación, se buscó explorar la mejor y mayor cantidad de componentes a remover en este conjunto de datos. Los resultados de la clasificación en los lotes de validación de la base de datos de San Diego, cuando se han removido incrementalmente desde la componente 1 hasta la 6, se visualizan en la **Figura 71** y de forma cuantitativa se exponen en la **Tabla 24.** Se determinó que en este caso específico, las componentes que presentan mayor relevancia en la corrección de derivas son las componentes 1 y 2, de acuerdo a lo obtenido mediante el **criterio de separabilidad** observado en la **Tabla 24.**

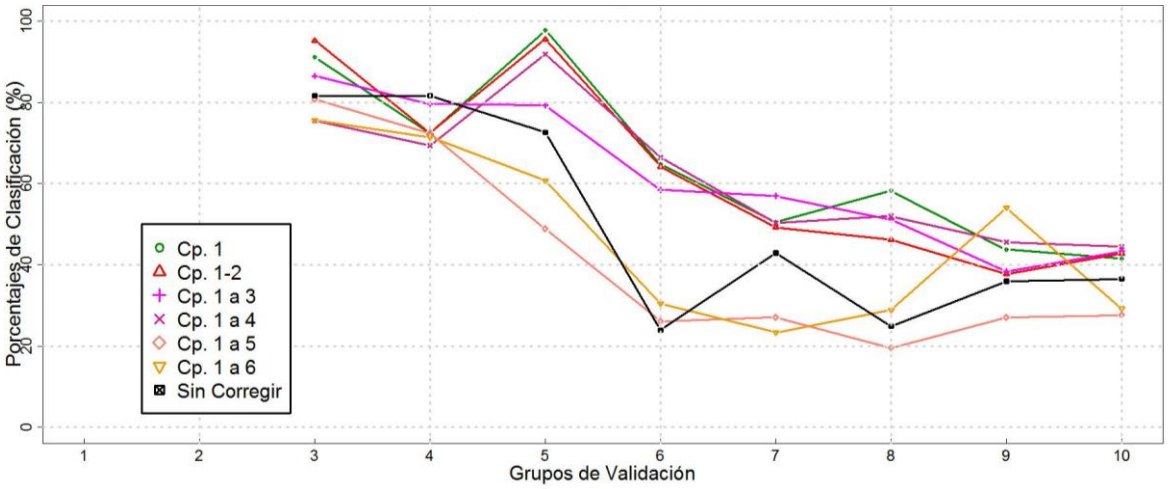


Figura 71. Porcentajes de clasificación usando corrección de componentes por CC-CPCA, removiendo 1 componente y hasta 6 componentes comparando con los datos no corregidos.

Aunque la componente 5 sea la segunda que mas varianza acumula según lo establecido en la **Tabla 23**, se visualiza al removerla la degradación del clasificador, por cuanto se estima que esta componente alberga información que puede corresponder a la concentración de los gases, lo que ocasiona el decaimiento en los porcentajes obtenidos en cada uno de los lotes.

En la **Tabla 24** se muestran en forma cuantitativa los resultados expuestos en la F**igura 71,** en donde la remoción de componentes incluye hasta 6 de ellas. Visiblemente los resultados no logran mejorar al introducir más componentes, por el contrario, tienden a la baja con respecto a los datos sin corrección.

El criterio de separabilidad en este caso estima que es suficiente al remover un máximo de dos componentes, sin embargo, al calcular los promedios de los ocho grupos de validación se obtuvo que el mejor promedio en los porcentajes de acierto en estos grupos se logra al remover sólo la primera componente, esto se debe a que en algunos lotes de validación se lograron obtener mejores porcentajes de acierto, pero también existen otros lotes en donde el porcentaje de acierto en lugar de aumentar disminuye.

Cabe anotar que las mediciones realizadas en San Diego poseen un alto grado de contaminación y saturación en algunos de los lotes, tal como lo refieren los autores (**Tabla No. 3**). En el lote 7 por ejemplo, se encuentran 3613 medidas realizadas durante el mes 21, en el mes 36 aparecen registradas 3600 mediciones 8 (lote 10) y de la misma forma el mes 20 (lote 6) posee 1625 medidas; de tal forma que las medidas durante estos 3 meses aportan mas de la mitad del conjunto total de mediciones. Al compararse en la **Tabla 23** las columnas dos y tres, los resultados a partir del lote 6, presentan una tendencia a la baja en la respuesta del clasificador, excepto en el lote diez, debido a que en los cinco meses anteriores no se

habían realizado medidas, lo que ayudó a la recuperación en la descontaminación de la capa de medida en los sensores. Por lo tanto, la saturación de la matriz de sensores en este periodo de tiempo ocasiona que sea más difícil la labor de mejorar la respuesta del clasificador en los últimos lotes de validación, por esta marcada influencia de los datos.

Lote No.	No corregidos	C1	C 1 ,2	C 1 a 3	C 1 a 4	C 1 a 5	C 1 a 6
1	ENTRENAMIENTO						
2							
3	81,9	**92,07**	**97,21**	86,53	75,57	80,76	75,65
4	89,8	**72,45**	**72,45**	79,59	69,39	72,45	71,43
5	72,6	**98,78**	**97,56**	79,26	91,85	48,89	60,74
6	24,0	**65,26**	**63,57**	58,46	66,42	26,14	30,46
7	43,0	**53,54**	**50,21**	56,96	50,27	27,08	23,33
8	24,9	**59,22**	**47,22**	51,11	52,00	19,56	28,89
9	35,6	**44,77**	**38,72**	38,43	45,55	27,05	54,09
10	35,6	**41,47**	**43,73**	43,35	44,48	27,65	29,25
Promedio	50,9	**66,01**	**63,91**	58,165	59,994	35,545	42,598
Criterio de separabilidad	1,001	**0,504**	**0,497**	0,565	0,574	1,476	1,367

Tabla 24. Resultados de las validaciones hechas con la remoción de hasta seis componentes principales comunes por CC-CPCA, en la base de datos de la Universidad de California.

En este punto, es importante recordar que el criterio de separabilidad se calculó a partir de los datos de entrenamiento, por lo tanto es un indicador para tomar decisiones en el número de componentes a remover, sin embargo, pueden existir variaciones en las validaciones debido a las características de las mediciones, del ruido presente y de la forma en que se recolectaron los datos; no obstante, es un parámetro importante a tener en cuenta en la selección del número de componentes a remover y que otorga con una baja taza de error el mayor número de componentes a ser removidas por CC-CPCA. En este caso específico con un 3% de error con respecto al mejor promedio de validación, indica que el número de componentes a remover en estos datos equivalen a las dos primeras de ellas, otorgando una mejora del 13,91 % en los promedios de los porcentajes obtenidos en los grupos de validación con respecto a los datos sin tratamiento de las derivas.

3.3 APORTE DE LA METODOLOGÍA PROPUESTA

Por medio de los resultados obtenidos se comprueba que al remover más de una componente principal común por el método de CC-CPCA, se puede mejorar la respuesta del sistema de forma representativa. Para ilustrar este efecto se comparan los resultados al remover una componente (Ziyatdinov, y otros, 2009), y al remover más de una componente principal común; en este último caso, la metodología propuesta para la selección del número de componentes a remover se basa en el cálculo de un indicador denominado **criterio de separabilidad,** explicado en la sección 2.3.5 de este trabajo. En la **Tabla 24** se aprecian los resultados de la aplicación de estas dos metodologías sobre los datos `chemosensors` en los ocho experimentos realizados, los cuáles se encuentran detallados en las secciones 3.1.1 a la 3.1.8 respectivamente.

Experimento	Sin corrección de derivas	CC-CPCA		CC-CPCA aplicando la metodología propuesta	
		Componentes removidas	**Promedio de porcentajes de acierto en validación**	**Componentes removidas**	**Promedio de porcentajes de acierto en validación**
1	36,34 %	Primera componente	52,37 %	Las cinco primeras	100,00 %
2	30,89 %		94,10 %	Las cuatro primeras	98,07 %
3	48,14 %		98,07 %	Las tres primeras	100,00 %
4	81,79 %		90,77 %	Las cinco primeras	100,00 %
5	79,55 %		97,88%	Las cuatro primeras	100,00 %
6	93,21 %		98,08%	Las cuatro primeras	10,00 %
7	96,35%		98,01 %	Las dos primeras	98,08 %
8	63,14 %		58,97%	Las cinco primeras	92,31%

Tabla 25. Comparación de los resultados obtenidos sobre la base de datos `chemosensors`.

En la **Tabla 25** se aprecia que con la metodología propuesta se obtienen mejores resultados en los porcentajes de acierto del clasificador para los diferentes experimentos realizados en la base de datos `chemosensors`. De lo anterior se concluye que la substracción de otras componentes adicionales a la primera, provoca la reducción de los efectos causados por las derivas en los sistemas de reconocimiento de olores.

CONCLUSIONES

En esta tesis de maestría se demostró que substraer más de una componente principal común utilizando la corrección de componentes por CPCA, resulta ser efectivo en el proceso de clasificación, tanto en los datos seleccionados para el entrenamiento como en los datos destinados para la validación. Lo anterior se refleja en el incremento de los porcentajes de acierto de cada uno de los lotes de las bases de datos en los que se aplicó el método propuesto. En consecuencia, se concluye que la presencia de derivas involucra además de la primera componente otras adicionales que deben ser removidas, con el propósito de obtener mejoras en la respuesta del clasificador y asimismo se establece que la remoción de más de una componente principal común es útil para reducir los efectos causados por las derivas en sistemas de olfato artificial

Por medio del análisis de componentes principales comunes (CPCA), comparado con la corrección de componentes por el método de gas de referencia (PCA), se demostró que el mejor desempeño se obtiene del primer método, en donde mediante la diagonalización conjunta se extrae la matriz de varianza acumulada en todas las componentes principales comunes, siendo las primeras de ellas las que marcan la concentración de la deriva en los conjuntos de las diferentes clases de gases medidos, mientras que en la búsqueda del mejor gas de referencia se requiere un análisis independiente con cada gas para encontrar el que mejor representa a las derivas en los datos de estudio. Por lo tanto, este estudio se basó en la estrategia basada en la corrección de componentes por CPCA para obtener el espacio de representación de los datos que tienen incorporadas las derivas y a partir de allí se determinó cuáles de ellas representaban las derivas en el sistema.

Utilizando el método de selección de características multivariantes CPCA con corrección de componentes, se diseñó el criterio de selección de las componentes principales comunes a remover para lograr mitigar el efecto de las derivas, a través de un indicador llamado en este trabajo **criterio de separabilidad.** El cálculo de este criterio se basó en conceptos de clustering, tal como se especificó en la **sección 2.3.5** y es hallado a partir de los datos de entrenamiento, por lo tanto es un indicador para tomar decisiones en el número de componentes a remover; sin embargo, pueden existir variaciones en las validaciones debido a las características de los datos, del ruido presente y de la forma en que éstos fueron recolectados; no obstante, es un parámetro importante a tener en cuenta en la selección del número de componentes a remover y a partir del cual se reduce la tasa de error en las tareas de clasificación.

La validación de la metodología de extracción de componentes principales comunes para corregir las derivas se realizó usando un clasificador k-NN en el que se sintonizó el parámetro de búsqueda de vecinos *k*, con el fin de garantizar un mejor desempeño. Con este clasificador se logró validar la influencia de la remoción de las componentes principales en

el procesado de los datos. Se destaca el hecho de observarse que aun usando un clasificador de bajo costo computacional se logran mejorar los porcentajes de acierto en los resultados de los grupos de validación, lo cual demuestra que la técnica de remoción de derivas, al usarse como etapa previa a la clasificación es un método potente que determina mejoras en la respuesta del sistema al ayudar a mejorar la estructura misma de los datos.

Al aplicar la metodología de selección de componentes principales comunes para la corrección de derivas, es necesario emplear un adecuado pre-procesamiento de los datos. En este trabajo de grado el pre-procesamiento abarcó tres aspectos que influyen positivamente en los resultados finales en el método de corrección de componentes; estos aspectos son el escalado, la normalización y la remoción de datos anómalos. Dado que los parámetros de los modelos de componentes principales dependen del escalado y normalización, estos se deben aplicar tanto a las muestras del conjunto de entrenamiento como en los subconjuntos de validación. Por otra parte los datos anómalos fueron previamente removidos para prevenir que la dirección de las componentes de deriva hubiesen sido influenciadas por estos datos. Como parte del pre-procesamiento, también se empleó la técnica de reducción de dimensionalidad denominada PCA, con el propósito de obtener un espacio de representación que suprima las características redundantes, lo cual se traduce en menor cantidad de datos para procesar, sin que esto afecte la calidad de la información.

Cabe anotar que cuando las mediciones poseen un alto grado de contaminación y saturación de los sensores, por un inadecuado manejo en las técnicas de recolección de datos los resultados de clasificación presentan una tendencia a la baja, debido a la contaminación de la matriz de sensores. Lo anterior, ocasiona que sea más difícil la labor de mejorar la respuesta del clasificador, especialmente cuando cronológicamente las medidas de prueba estén más alejadas del conjunto de entrenamiento. Por lo tanto, en la recolección de los datos, no se recomienda realizar un alto número de mediciones en un corto periodo de tiempo, debido a que los sensores requieren un tiempo de descontaminación y una fase de limpieza para que los resultados sean los esperados.

Un factor preponderante en las respuestas de los sistemas de detección de aromas, es el hecho de tener en los casos experimentales volátiles medidos en concentraciones de diferente valor, tal como se utilizó en la base de datos de la Universidad de California. La metodología propuesta permite que mediante la utilización de un clasificador convencional como lo es el k-nn se mejoren los porcentajes de acierto en la clasificación, aun cuando se desconocen las concentraciones de los gases.

Como trabajos futuros de esta tesis, se plantean la aplicación de esta misma metodología de corrección de componentes por CC-CPCA, realizando la validación con un clasificador más potente. Asimismo, se establece que éste mismo estudio puede realizarse sobre esta base de datos, una vez los autores publiquen la concentración en la que fue medida cada gas y de

esta manera poder realizar el análisis de las derivas en los datos conociendo previamente los subgrupos generados en cada clase de acuerdo a la concentración de los mismos. Lo anterior, permite discriminar las componentes correspondientes a la concentración de las componentes de deriva.

REFERENCIAS

Arthurson, T., Eklöv, T., Lundström, I., Marterson, P., Sjöstrom, M., & Holmberg, M. (2000). Drift correction for gas sensors using multivariate methods. *Journal of Chemometrics*, 711-723.

Bahraminejad, B., Basri, S., Isa, M., & Hambali, Z. (2011). Hydrogen detection in organic gas mixtures based on analyzing the transient response. *Sensors review*, 26-31.

Berna, A. (2010). Metal Oxide Sensors for Electronic Noses and Their Application to Food Analysis. *Sensors, Vol.10*, 3882-3910.

Berrueta, L. A., Alonso-Salces, R. M., & Héberger , K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A*, 196–214.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bógalo, J., & Quilis, E. (2003). *Estimación del ciclo económico mediante filtros Butterworth.* España: Instituto Nacional de Estadística.

Brezmes, J., López, M., Llobet, E., Vilabona, X., & et. al. (2005). Evaluation of an Electronic Nose to Assess Fruit Ripeness. *IEEE Sensors Journal*, 97-108.

Cho, J., Kim, Y., Jin Na, K., & Jeon, G. (2008). Wireless electronic nose system for real-time quantitative analysis of gas mixtures using micro-gas sensor array and neuro fuzzy network. *Sensors and actuators B: Chemical*, 104-111.

Di Carlo, S., & Falasconi, M. (2012). Drift correction methods for gas chemical sensors in artificial olfaction systems: techniques and challenges. En W. W. InTech, *Advances in chemical sensors* (págs. 305-326). Rijeka: Intech-Open-Acces.

Durán Acevedo, C. M. (2005). En C. M. Durán Acevedo, *Diseño y optimización de los subsistemas de un sistema de olfato electrónico para aplicaciones agroalimentarias e industriales.* Tarragona, España: Universitat Rovira I Virgili.

Durán, C., & Baldovino, D. (2009). Monitoring System to Detect the Maturity of Agro-industrial Products Through of an Electronic Nose. *Revista Colombiana de Tecnologías de Avanzada. Vol.1, No.13.*, 1-8.

Figaro Company. (s.f.). Obtenido de http://www.figarosensor.com/

Fu, J., Li, G., Qin, Y., & Freeman, W. (2007). A pattern recognition method for electronic noses based on an olfactory neural network. *Sensors and actuators*, 489-497.

Gonzalez-Jimenez, J., Monroy, J. G., & Blanco, J. L. (2011). The Multi-Chamber Electronic Nose-An Improved Olfaction Sensor for Mobile Robotics. *Sensors, 11*, 6145-6164.

Guadarrama, A., Fernández, J., Ïñiguez, M., Souto, J., & de Saja, J. (2001). Discrimination of wine aroma using an array of conducting polymer sensors in conjunction with solid-phase micro-extraction (SMPE) technique. *Sensors and Actuators*, 401-408.

Gualdrón G., O. E., Durán, C. M., Isaza, C. V., Carvajal F., A., & Uribe, C. (2011). Sistema de olfato electrónico de bajo costo para la detección de diferentes compuestos químicos contaminantes. *Revista Colombiana de tecnologías de avanzada*, 121-126.

Gualdron Guerrero, O. E. (2006). *Desarrollo de diferentes métodos de selección de variables para sistemas multisensoriales.* Tarragona, España: Universitat Rovira I Virgili.

Guiñón, J. L., Ortega, E., García-Antón, J., & Pérez-Herranz, V. (2007). Implementación y análisis del filtro de media móvil. *Filtrado de señales (I)*, 220-227.

Gutierrez-Osuna, R. (2000). Drift reduction for metal-oxide sensor arrays using canonical. *Proceedings of the 7th International Symp. On Olfaction and Electronic Nose,* (págs. 147-155). Brighton, Reino Unido: Institute of Physics Publishing.

Gutierrez-Osuna, R. (2002). Pattern analysis for machine olfaction: a review. *Sensors Journal, IEEE*, 189-202.

Gutierrez-Osuna, R. (18-21 de Mayo de 2003). Signal processing methods for drift compensation. 2nd NOSE II Workshop. Linköping, Suecia.

Huang, D., & Leung, H. (2009). Reconstruction of Drifting Sensor Responses Based on Papoulis-Gerchberg Method. *Sensros Journal, IEEE*, 595-604.

Kashwan, K., & Bhuyan, M. (2005). Robust electronic-nose system with temperature and humidity drift compensation for tea and spice flavour discrimination. *Sensors and the International Conference on new Techniques in Pharmaceutical and Biomedical Research*, 154-158.

Kermit, M., & Tomic, O. (2003). independent component analysis applied on gas sensor array measurement data. *Sensors Journal, IEEE*, 218-228.

Kwan, T., Boussaid, F., & Bermak, A. (2011). A CMOS Single-Chip Gas Recognition Circuit for Metal Oxide Gas Sensor Arrays. *Circuits and Systems IEEE*, 1569-1580.

Lonwongtragool, P., Wongchoosuk, C., & Kerdcharoen, T. (2011). Portable electronic nose for beverage quality assessment. *Electrical Engineering/Electronics, Computer, Telecomunications and Information Technology (ETCI-CON)*, 163-166.

Marco, S., & Gutierrez-Galvez, A. (2012). Signal and data processing for machine olfaction and chemical sensing: a review. *IEEE Sensors Journal*, 3189-3214.

Marín Diazaraque, J. M. (2013). *Universidad Carlos III de Madrid*. Recuperado el 18 de Octubre de 2013, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema3am.pdf

Moreno, I., Caballero, R., Galán, R., Matía, F., & Jiménez, A. (2009). La Nariz Electrónica: Estado del Arte. *Revista Iberoamericana de automática e Informática Inustrial*, 76-91.

Padilla, A., Perera, A., Montoliu, I., Chaudry, K., Persaud, K., & Marco, S. (2010). Drift compensation of gas sensor array data by Orthogonal Signal Correction. *Chemometrics and Intelligent Laboratory Systems*, 28-35.

Paniagua, M., Llobet, E., Brezmes, J., Vilanova, X., & et. al. (2003). On-line drift counteraction for metal oxide gas sensor arrays. *Electronics Letters IEEE*, 40-42.

Pearce, T., Schiffman, S., Nagle, H., & Gardner, J. (2003). *Handbook of Machine Olfaction: Electronic Nose Technology.* Michigan (EE.UU): Wiley B.C.H.

Perera, A., Papamichail, N., Barsan, N., Weimar, U., & et.al. (2006). On-line novelty detection by recursive dynamic principal component analysis and gas sensor arrays under drift conditions. *Sensors Journal, IEEE*, 770-783.

Quicazán, M., Díaz M., A., & Zuluaga D., C. (2010). La nariz electrónica, una novedosa herramienta para el control de procesos y calidad en la industria agroalimentaria. *Vitae, Revista de la Facultad de Química Farmaceútica de la Universidad de Antioquia*, 209-217.

Rodriguez Gamboa, J. C. (2013). Modelo de Inferencia de la Respuesta de un Sensor de gas de Estado Sólido para Sistemas de Olfato Electrónico. *Medellín: ITM*.

Rodriguez Gamboa, J. C., & Durán Acevedo, C. M. (2008). Sistema de olfato electrónico para la detección de compuestos volátiles. *Revista Colombiana de tecnologías de avanzada*, 20-26.

Rodríguez, J., Durán, C., & Reyes, A. (2010). Electronic Nose for Quality Control of Colombian Coffee through the Detection of Defects in "Cup Tests". *Sensors*, 36-46.

Rodríguez-Gamboa, J. C., Albarracín-Estrada, E. S., & Delgado-Trejos, E. (2011). Quality Control Through Electronic Nose System. En E. b. Eldin, *Modern Approaches To Quality Control* (págs. 505-522). Rijeka, Croatia: Intech Europe.

Song, K., Wang, Q., Zhang, H., & Cheng, Y. (2011). A Wireless Electronic Nose System Using a Fe2O3 Gas Sensing Array and Least Squares Support Vector Regression. *Sensors*, 485-505.

Tian, F., Yang, S., & Dong, K. (2005). Circuit and Noise Analysis of Odorant Gas Sensors in an E-Nose. *Sensors, Vol.5*, 85-96.

University of California. (25 de Abril de 2012). *Machine Learning Repository*. Obtenido de Gas Sensor Array Drift Dataset Data Set: http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset

Vergara, A., Vembu, S., Ayhan, T., Ryan, M., Homer, M., & Huerta, R. (2012). Chemical gas sensor drift compensation using classifier ensembles. *Sensor and Actuators*.

Vergara, A., Vembua, S., Ayhan, T., Ryan, M. A., Homer, M. L., & Huerta, R. (2012). Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 320–329.

Wilson, A., & Baietto, M. (2009). Applications and Advances in Electronic-Nose Technologies. *Sensors, Vol. 9*, 5099-5148.

Zhang, M., Wang, X., Liu, Y., Xu, X., & Zhou, G. (2012). Species discrimination among three kinds of puffer fish using an electronic nose combined with olfactory sensory evaluation. *PubMed*.

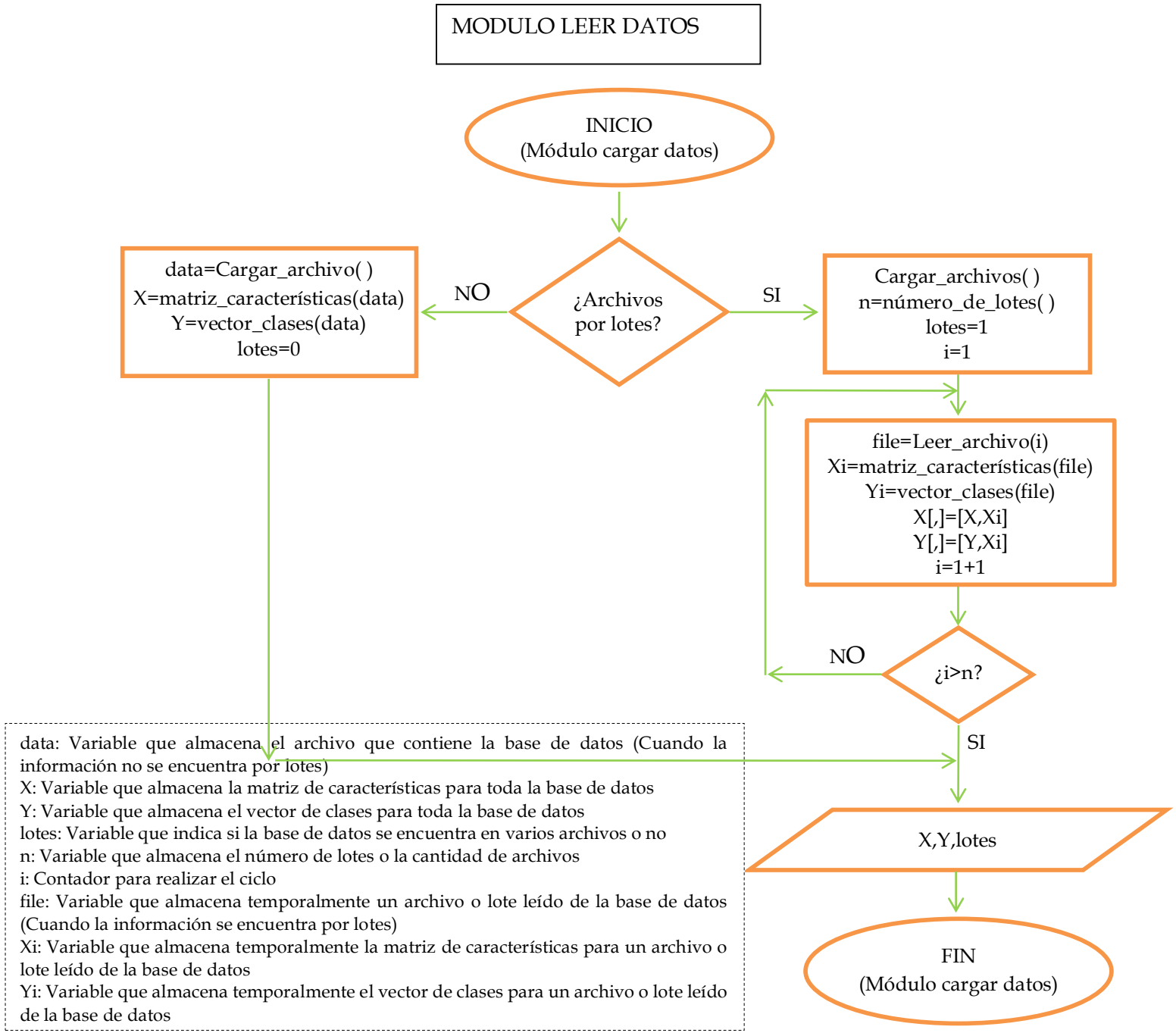
Zhong, M., & Girolami, M. (2012). A Bayesian Approach to Approximate Joint Diagonalization of Square Matrices. *Proceedings of the 29th International Conference of Machine Learning*. Edinburgh: Scotland, UK.

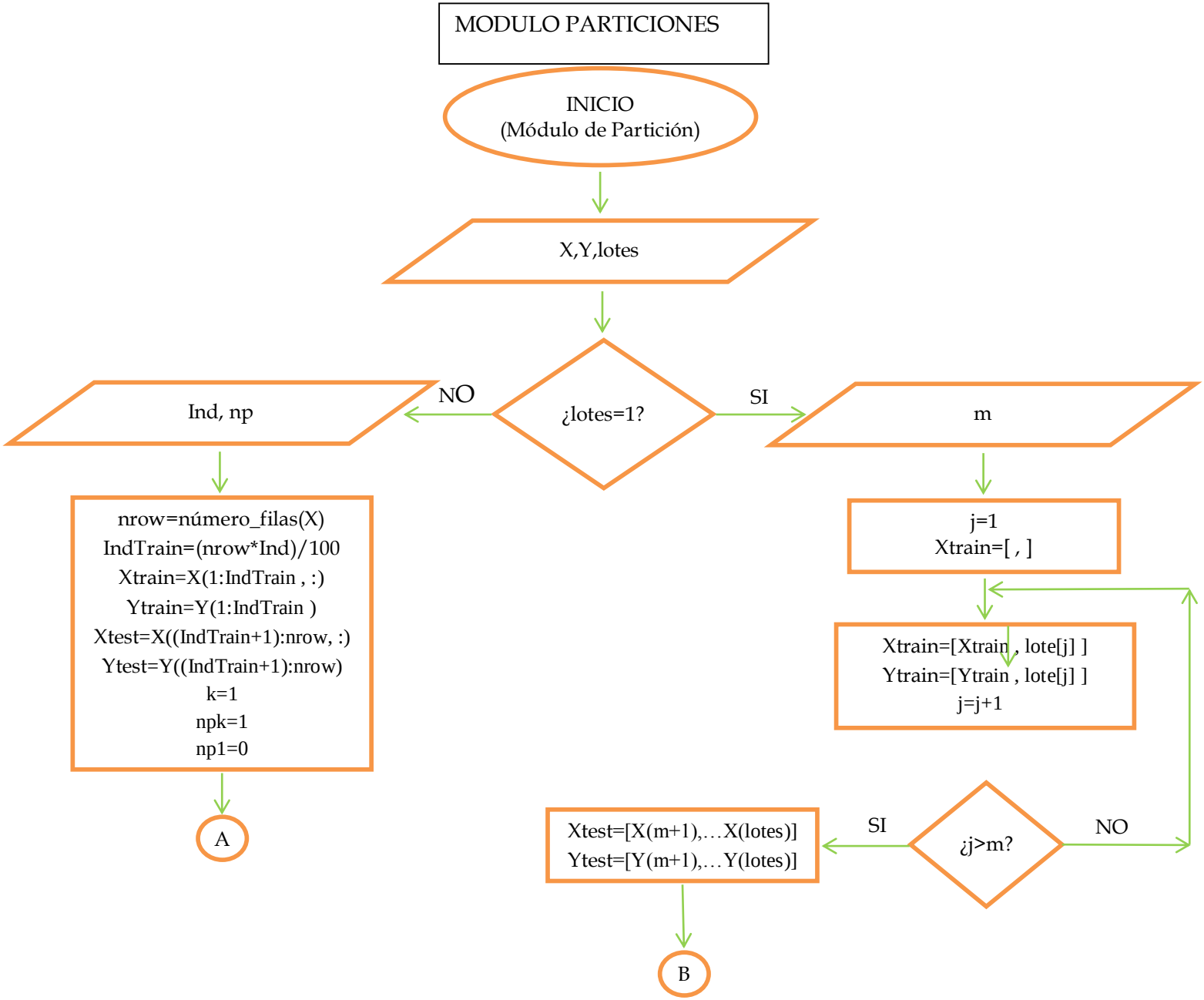
Zhou, H., Homer, M., Shevade, A., & Ryan, M. (2005). Nonlinear Least-Squares Based Method for Identifying and Quantifying Single and Mixed Contaminants in Air with an Electronic Nose. *Sensors, Vol. 6*, 1-18.

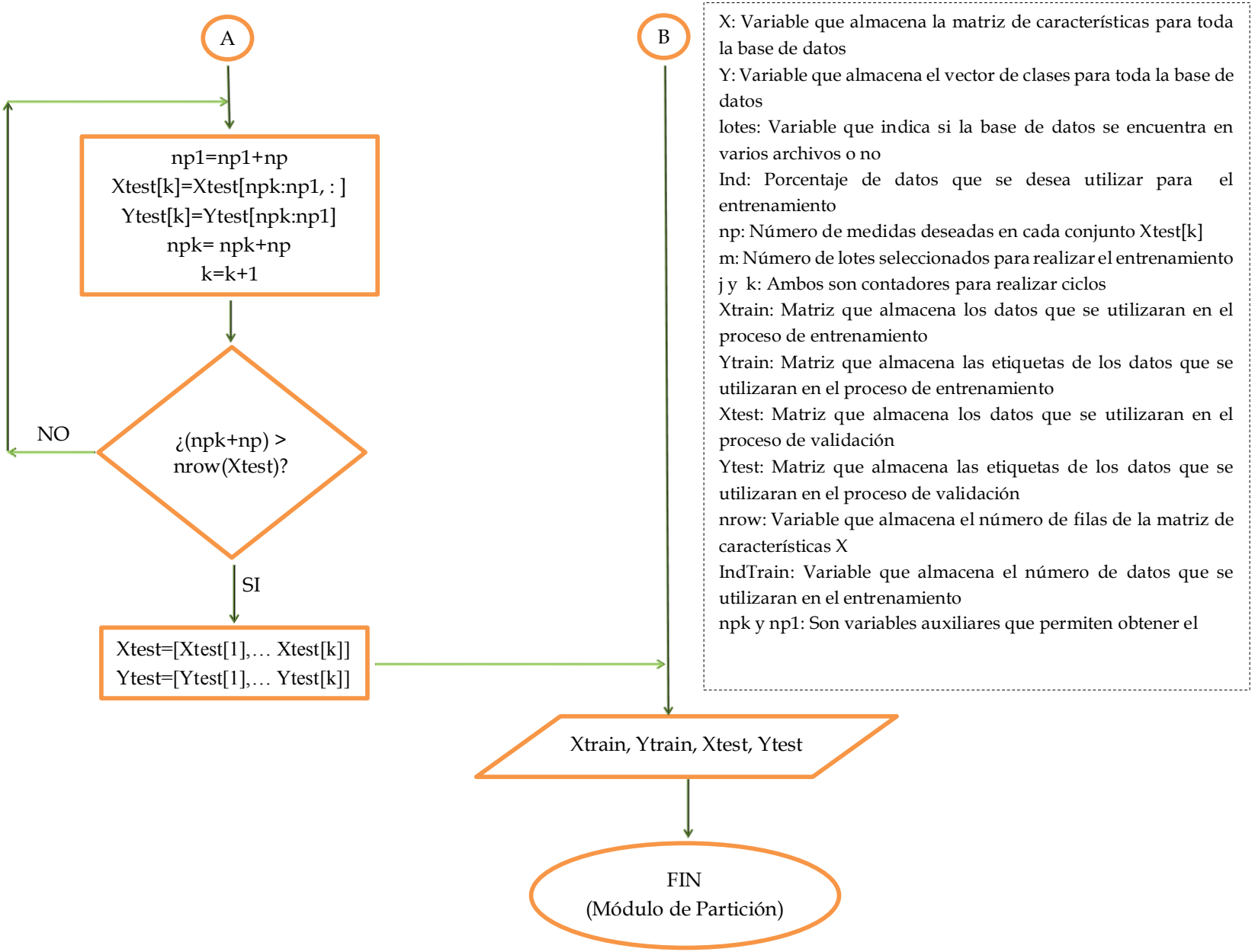
Ziyatdinov, A., & Perera-Lluna, A. (2014). Data simulation in machine olfaction with the R package Chemosensors. *PLOS-one*.

Ziyatdinov, A., Marco, S., Chaudry, A., Persaud, K., Caminal, P., & Perera, A. (2009). Drift compensation of gas sensor array data by common principal component analysis. *Sensors and Actuators*.

APÉNDICE A


MODULO LEER DATOS
INICIO
(Módulo cargar datos)
¿Archivos por lotes?
NO
SI
data=Cargar_archivo( )
X=matriz_características(data)
Y=vector_clases(data)
lotes=0
Cargar_archivos( )
n=número_de_lotes( )
lotes=1
i=1
file=Leer_archivo(i)
Xi=matriz_características(file)
Yi=vector_clases(file)
X[,]=[X,Xi]
Y[,]=[Y,Xi]
i=1+1
¿i>n?
NO
SI
X,Y,lotes
FIN
(Módulo cargar datos)
data: Variable que almacena el archivo que contiene la base de datos (Cuando la información no se encuentra por lotes)
X: Variable que almacena la matriz de características para toda la base de datos
Y: Variable que almacena el vector de clases para toda la base de datos
lotes: Variable que indica si la base de datos se encuentra en varios archivos o no
n: Variable que almacena el número de lotes o la cantidad de archivos
i: Contador para realizar el ciclo
file: Variable que almacena temporalmente un archivo o lote leído de la base de datos (Cuando la información se encuentra por lotes)
Xi: Variable que almacena temporalmente la matriz de características para un archivo o lote leído de la base de datos
Yi: Variable que almacena temporalmente el vector de clases para un archivo o lote leído de la base de datos

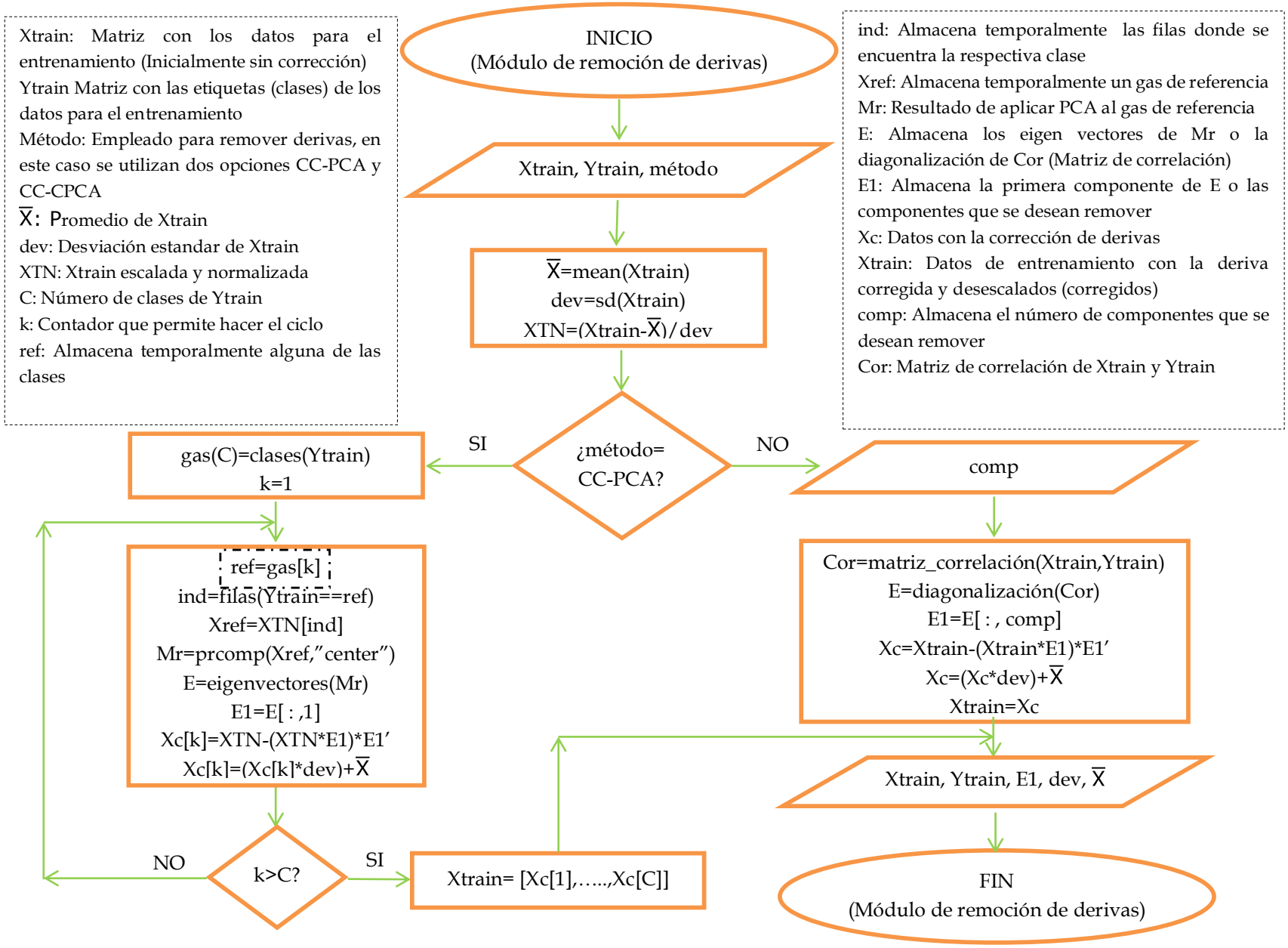

MODULO PARTICIONES
INICIO
(Módulo de Partición)
X,Y,lotes
¿lotes=1?
NO
SI
Ind, np
m
nrow=número_filas(X)
IndTrain=(nrow*Ind)/100
Xtrain=X(1:IndTrain , :)
Ytrain=Y(1:IndTrain)
Xtest=X((IndTrain+1):nrow, :)
Ytest=Y((IndTrain+1):nrow)
k=1
npk=1
np1=0
j=1
Xtrain=[,]
Xtrain=[Xtrain , lote[j]]
Ytrain=[Ytrain , lote[j]]
j=j+1
¿j>m?
SI
NO
Xtest=[X(m+1),…X(lotes)]
Ytest=[Y(m+1),…Y(lotes)]
A
B



X: Variable que almacena la matriz de características para toda la base de datos
Y: Variable que almacena el vector de clases para toda la base de datos
lotes: Variable que indica si la base de datos se encuentra en varios archivos o no
Ind: Porcentaje de datos que se desea utilizar para el entrenamiento
np: Número de medidas deseadas en cada conjunto Xtest[k]
m: Número de lotes seleccionados para realizar el entrenamiento
j y k: Ambos son contadores para realizar ciclos
Xtrain: Matriz que almacena los datos que se utilizaran en el proceso de entrenamiento
Ytrain: Matriz que almacena las etiquetas de los datos que se utilizaran en el proceso de entrenamiento
Xtest: Matriz que almacena los datos que se utilizaran en el proceso de validación
Ytest: Matriz que almacena las etiquetas de los datos que se utilizaran en el proceso de validación
nrow: Variable que almacena el número de filas de la matriz de características X
IndTrain: Variable que almacena el número de datos que se utilizaran en el entrenamiento
npk y np1: Son variables auxiliares que permiten obtener el

MODULO REMOCIÓN DE DERIVAS

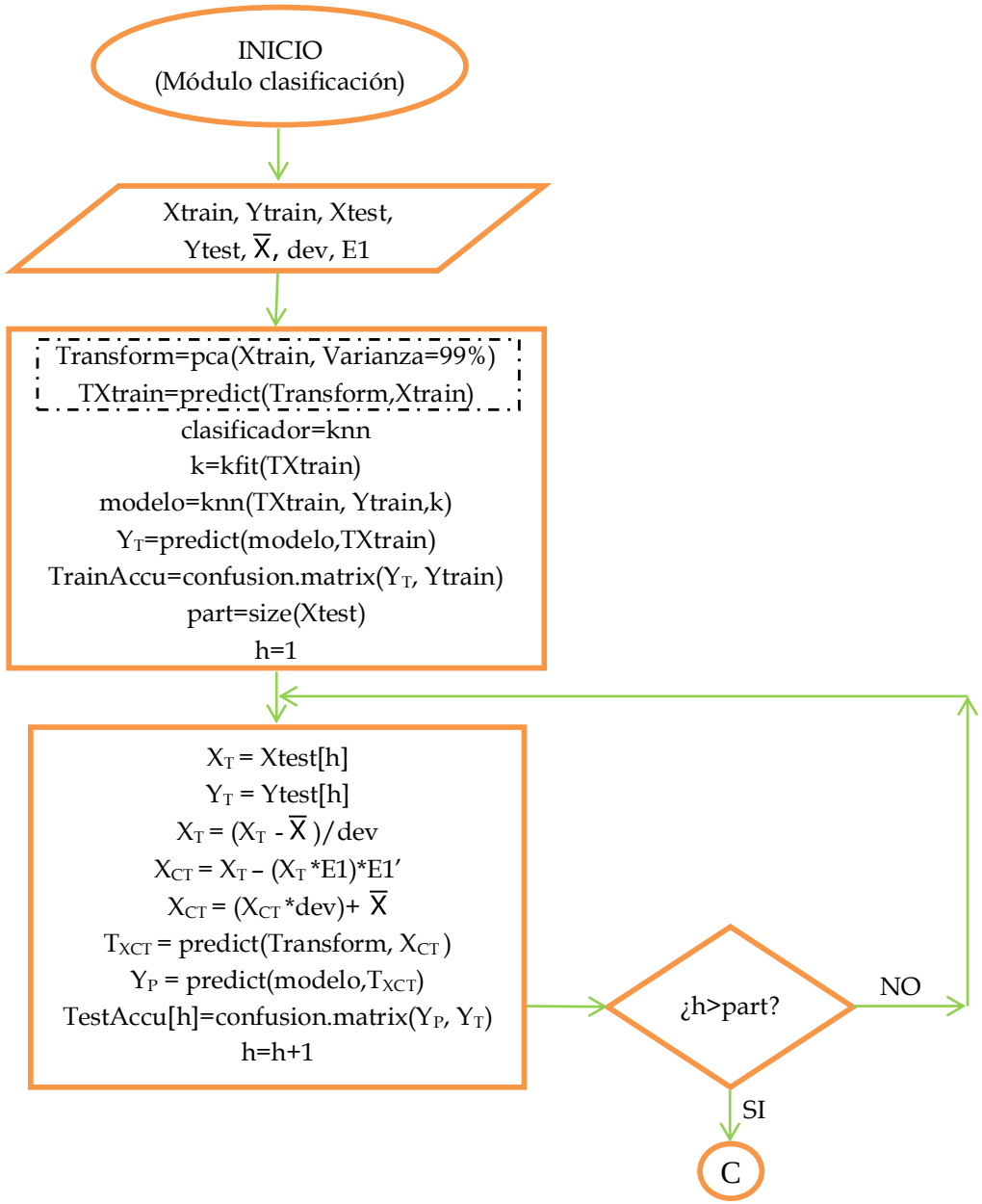
Xtrain: Matriz con los datos para el entrenamiento (Inicialmente sin corrección)
Ytrain Matriz con las etiquetas (clases) de los datos para el entrenamiento
Método: Empleado para remover derivas, en este caso se utilizan dos opciones CC-PCA y CC-CPCA
$\overline{X}$: Promedio de Xtrain
dev: Desviación estandar de Xtrain
XTN: Xtrain escalada y normalizada
C: Número de clases de Ytrain
k: Contador que permite hacer el ciclo
ref: Almacena temporalmente alguna de las clases

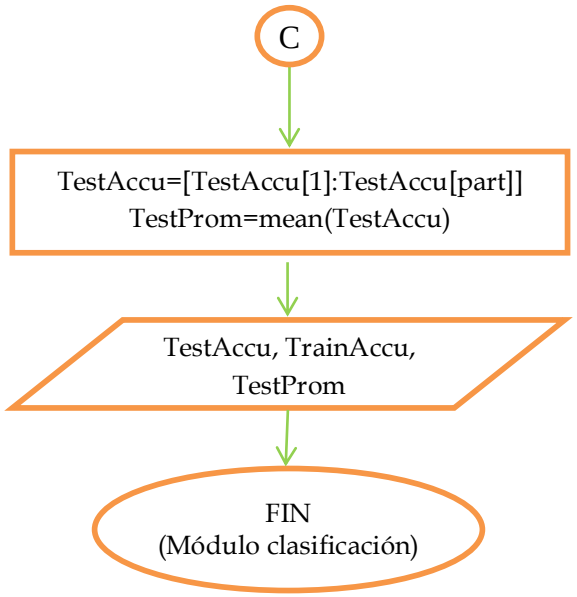


ind: Almacena temporalmente las filas donde se encuentra la respectiva clase
Xref: Almacena temporalmente un gas de referencia
Mr: Resultado de aplicar PCA al gas de referencia
E: Almacena los eigen vectores de Mr o la diagonalización de Cor (Matriz de correlación)
E1: Almacena la primera componente de E o las componentes que se desean remover
Xc: Datos con la corrección de derivas
Xtrain: Datos de entrenamiento con la deriva corregida y desescalados (corregidos)
comp: Almacena el número de componentes que se desean remover
Cor: Matriz de correlación de Xtrain y Ytrain

MODULO DE CLASIFICACIÓN

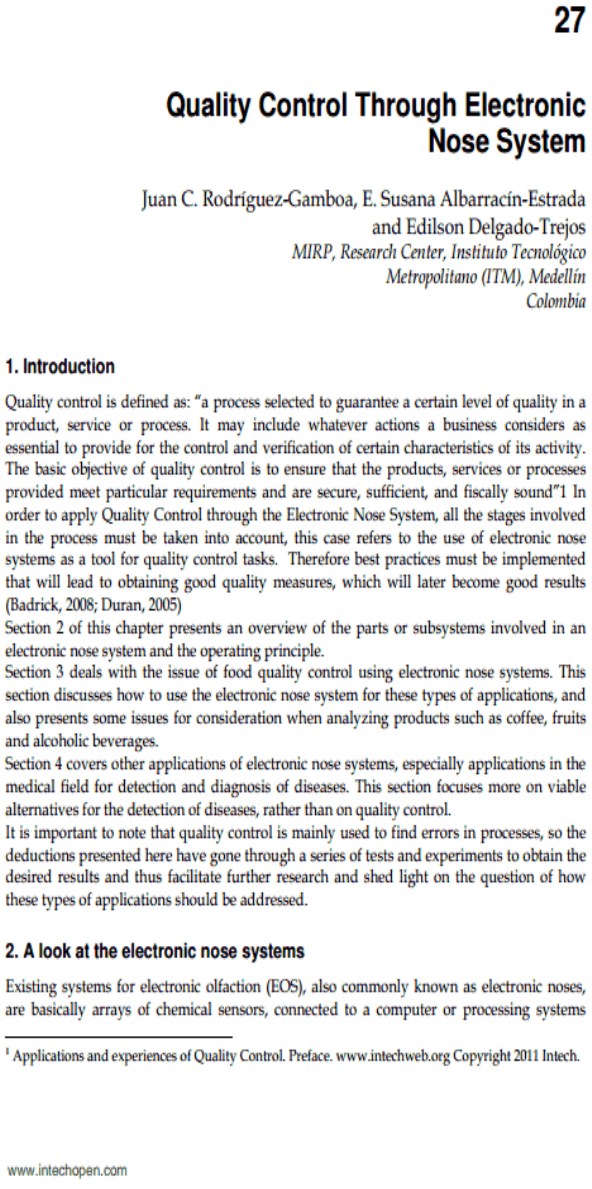
Xtrain: Datos de entrenamiento con la deriva corregida y desescalados.
Ytrain Matriz con las etiquetas (clases) de los datos de entrenamiento
Xtest: Matriz con los datos que se utilizaran en el proceso de validación
Ytest: Matriz con las etiquetas de los datos que se utilizaran en el proceso de validación
$\overline{X}$: Promedio de Xtrain
dev: Desviación estandar de Xtrain
E1: Almacena la primera componente de E o las componentes que se desean remover según el método escogido en el módulo de remoción de derivas
Transform: PCA con varianza capturada 99%
TXtrain: Datos de entrenamiento pasados al espacio PCA con un 99% de la varianza acumulada (Extrae las components)
k: El mejor k para el método knn
modelo: Se guarda el modelo del clasificador utilizando el mejor k
Y_T: Almacena el resultado del clasificador
TrainAccu: Porcentaje de acierto en el entrenamiento
part: Tamaño de Xtest
h: Contador para realizar ciclo




C
TestAccu=[TestAccu[1]:TestAccu[part]]
TestProm=mean(TestAccu)
TestAccu, TrainAccu,
TestProm
FIN
(Módulo clasificación)

APENDICE C
PRODUCTOS DERIVADOS DE ESTE TRABAJO DE GRADO

- ✓ Durante la ejecución de este proyecto de grado se elaboró y publico un capítulo de libro titulado Quality Control Through Electronic Nose System en el libro Modern Approaches To Quality Control con ISBN 978-953- 307-971-4 y DOI: 10.5772/22217.



27

Quality Control Through Electronic Nose System

Juan C. Rodríguez-Gamboa, E. Susana Albarracín-Estrada and Edilson Delgado-Trejos
MIRP, Research Center, Instituto Tecnológico Metropolitano (ITM), Medellín Colombia

1. Introduction

Quality control is defined as: "a process selected to guarantee a certain level of quality in a product, service or process. It may include whatever actions a business considers as essential to provide for the control and verification of certain characteristics of its activity. The basic objective of quality control is to ensure that the products, services or processes provided meet particular requirements and are secure, sufficient, and fiscally sound"1 In order to apply Quality Control through the Electronic Nose System, all the stages involved in the process must be taken into account, this case refers to the use of electronic nose systems as a tool for quality control tasks. Therefore best practices must be implemented that will lead to obtaining good quality measures, which will later become good results (Badrick, 2008; Duran, 2005)

Section 2 of this chapter presents an overview of the parts or subsystems involved in an electronic nose system and the operating principle.

Section 3 deals with the issue of food quality control using electronic nose systems. This section discusses how to use the electronic nose system for these types of applications, and also presents some issues for consideration when analyzing products such as coffee, fruits and alcoholic beverages.

Section 4 covers other applications of electronic nose systems, especially applications in the medical field for detection and diagnosis of diseases. This section focuses more on viable alternatives for the detection of diseases, rather than on quality control.

It is important to note that quality control is mainly used to find errors in processes, so the deductions presented here have gone through a series of tests and experiments to obtain the desired results and thus facilitate further research and shed light on the question of how these types of applications should be addressed.

2. A look at the electronic nose systems

Existing systems for electronic olfaction (EOS), also commonly known as electronic noses, are basically arrays of chemical sensors, connected to a computer or processing systems

[1] Applications and experiences of Quality Control. Preface. www.intechweb.org Copyright 2011 Intech.

www.intechopen.com

Figura C1. Captura de pantalla de la primera página del capítulo de libro.

- ✓ A la fecha, también está en proceso de producción un artículo titulado Techniques for drift correction gas sensor used in electronic nose system: A review.

Techniques for Drift Correction on Gas Sensor Used in Electronic Nose System: A Review

Status: Planned Paper
Section: *Sensors*
Type of Paper: Review Paper
Title: "Techniques for Drift Correction on Gas Sensor Used in Electronic Nose System: A Review."
Authors: Susana Albarracin-Estrada and Juan Rodríguez-Gamboa.
Affiliation: Instituto Tecnológico Metropolitano, Medellín - Colombia.

Abstract: The purpose of this paper is to present a review of techniques have been used to counteract the drift in chemical sensors commonly used in electronic nose systems. The drifts are changes in sensor response due to many factors and it is necessary to minimize their effects, one choice is controlling all parameters of the measure or applying processes and calibration algorithms to counteract the excesses that can not be controlled. This problem has been attacked from different approaches, the first of which corresponds to the correction of drift from the design and construction of new sensors, the second corresponds to the correction of the drifts in the classification stage, designing and implementing powerful classifiers improve the accuracy in the response and the discriminating power of electronic nose systems, the third approach addresses the problem from the space of representation of the feature set of the volatiles sensed, that is enhancing the response of the sensors in the characterization stage to perform the tasks of classifying and pattern recognition. This article presents a review of the most relevant works developed to mitigate this problem frequently presented in electronic nose systems with arrays of chemical sensors. Each approach has advantages and disadvantages for removing the drift, it will focus on techniques that address the problem from the preprocessed signal response of the sensor array, such as independent component analysis (ICA), principal component analysis (PCA), common principal component analysis (CPCA), between other techniques. This will allow us to conclude which is the best way to tackle the problem of drift and contribute to future work in this area.

Keywords: drift counteraction, Electronic Nose, gas-sensor

Figura C2. Captura de pantalla tomada del sitio web de Science Journal bajo la url: http://journal.insciences.org/techniques-for-drift-correction-on-gas-sensor-used-in-electronic-nose-system-a-review/

- ✓ Adicionalmente, se espera elaborar dos artículos con los resultados y aportes de este trabajo de grado y publicarlos en revistas indexadas.