



**UNIVERSIDAD
DEL AZUAY**

**UNIVERSIDAD DEL AZUAY
FACULTAD DE CIENCIA Y TECNOLOGÍA
ESCUELA DE INGENIERÍA EN ALIMENTOS**

**Modelado quimiinformático de índices de retención de
compuestos orgánicos volátiles del café.**

**Trabajo de graduación previo a la obtención del título de:
INGENIERO EN ALIMENTOS**

Autor:

CRISTHIAN JOEL CABRERA ERAS

Director:

Dr. CRISTIAN ROJAS VILLA

Co-Director:

Dr. PIERCOSIMO TRIPALDI

CUENCA, ECUADOR

2019

DEDICATORIA

El presente trabajo dedico a toda mi familia, en especial a Rosana Eras Agila y Carlos Joel Cabrera Peñaloza, mis padres, quienes han sido el máximo apoyo en todas las etapas de formación como persona y estudiante, además de mi esposa Jenniffer Medina quien con su amor incondicional siempre me motiva a superarme día a día.

AGRADECIMIENTO

En primer lugar quiero agradecer a Dios por darme la sabiduría y la perseverancia para culminar mi carrera universitaria como Ingeniero en Alimentos ya que sin él nada de esto sería posible.

A mis padres, Rosana Eras y Joel Cabrera por el apoyo brindado desde que inicie la universidad, a mi esposa Jenniffer Medina por estar en cada momento difícil de mi vida, a mis hermanas Jessica y Mónica quienes siempre estuvieron apoyándome comprendiendo mi ausencia como hermano en casa. A la Universidad del Azuay por los conocimientos impartidos hacia mí, al PhD Cristian Rojas Villa y todo el laboratorio de Análisis de Química Instrumental, que colaboraron con sus conocimientos y valioso tiempo para guiarme en el transcurso de este proyecto.

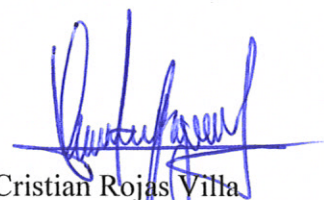
RESUMEN

En este trabajo se desarrolló una relación cuantitativa estructura-propiedad (QSPR) para modelar los índices de retención de 114 compuestos orgánicos volátiles del café, medidos en la columna polar HP-INNOWAX en cromatografía de gases. Se utilizaron los programas Dragon y PaDEL-Descriptor para calcular 18283 descriptores moleculares, los cuales fueron reducidos mediante el algoritmo V-WSP para posteriormente desarrollar un modelo con tres descriptores mediante el método de reemplazo. La estabilidad del modelo se verificó mediante validación interna y validación externa. También se definió el dominio de aplicabilidad del modelo y se brindó una explicación del mecanismo de acción de los descriptores involucrados.

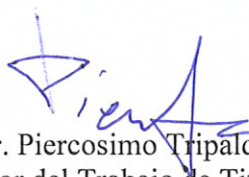
PALABRAS CLAVE: café, compuestos orgánicos volátiles, índice de retención, QSPR



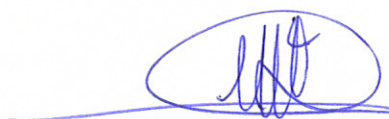
Ing. Maria Fernanda Rosales, MSc.
Directora de Escuela



Dr. Cristian Rojas Villa
Director del Trabajo de Titulación



Dr. Piercosimo Tripaldi
Codirector del Trabajo de Titulación

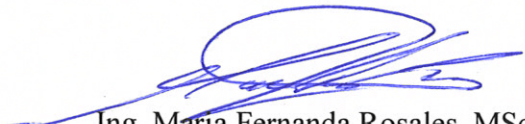


Cristhian Joel Cabrera Eras
Autor

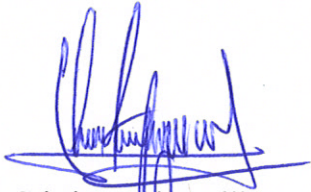
ABSTRACT

In this work, a quantitative structure-property relationship (QSPR) was developed to model the retention indices of 114 volatile organic compounds in coffee, measured in the HP-INNOWAX polar column in gas chromatography. The Dragon and PaDEL-Descriptor programs were used to calculate 18283 molecular descriptors, which were reduced through the V-WSP algorithm to later develop a model with three descriptors using the replacement method. The stability of the model was verified through both internal and external validation. The applicability domain of the model was also defined and an explanation of the action mechanism of the involved descriptors was provided.

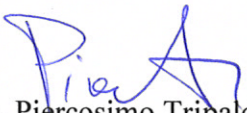
KEYWORDS: coffee, volatile organic compounds, retention index, QSPR




Ing. Maria Fernanda Rosales, MSc.
Faculty Director



Dr. Cristian Rojas Villa
Thesis Director



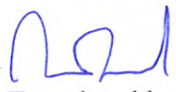
Dr. Piercosimo Tripaldi
Thesis Co-director



Cristhian Joel Cabrera Eras
Author



UNIVERSIDAD
Mayor de San Marcos
Dpto. Idiomas



Translated by
Ing. Paúl Arpi

TABLA DE CONTENIDO

DEDICATORIA.....	ii
AGRADECIMIENTO.....	iii
RESUMEN	iv
ABSTRACT	v
TABLA DE CONTENIDO	vi
INTRODUCCIÓN	1
MARCO TEORICO.....	3
1.1. EL CAFÉ.....	3
1.2. COMPUESTOS ORGÁNICOS VOLÁTILES DEL CAFÉ	4
1.3. CROMATOGRAFÍA DE GASES E ÍNDICES DE RETENCIÓN.....	4
1.4. RELACIONES CUANTITATIVAS ESTRUCTURA-PROPIEDAD	6
METODOLOGIA.....	8
2.1. GENERACIÓN DE LA BASE DE DATOS.....	8
2.2. CURADO DE LA INFORMACIÓN.....	8
2.3. DISEÑO MOLECULAR Y CÁLCULO DE DESCRIPTORES MOLECULARES.....	8
2.4. REDUCCIÓN NO SUPERVISADA DE DESCRIPTORES MOLECULARES	9
2.5. SELECCIÓN SUPERVISADA DE DESCRIPTORES MOLECULARES	10
2.6. VALIDACIÓN DEL MODELO	10
2.7. ANÁLISIS DEL DOMINIO DE APLICABILIDAD	10
2.8. MECANISMO DE ACCIÓN (INTERPRETACIÓN DE LOS DESCRIPTORES).....	11
RESULTADOS Y DISCUSIÓN	13
3.1. DESCRIPCIÓN DE LA BASE DE DATOS.....	13
3.2. CURADO DE LA BASE DE DATOS	13
3.3. DISEÑO MOLECULAR Y CÁLCULO DE DESCRIPTORES MOLECULARES.....	14
3.4. REDUCCIÓN NO SUPERVISADA DE DESCRIPTORES MOLECULARES	14

3.5. DIVISIÓN DE LA BASE DE DATOS	15
3.6. SELECCIÓN SUPERVISA DE DESCRIPTORES MOLECULARES	15
3.7. VALIDACIÓN DEL MODELO	16
3.8. DOMINIO DE APLICABILIDAD	18
3.9. MECANISMO DE ACCIÓN	18
CONCLUSIONES.....	19
BIBLIOGRAFIA	20

INDICE DE TABLAS

Tabla 1. Resumen de los mejores modelos QSPR de 1 a 7 descriptores moleculares obtenidos mediante el método de reemplazo.....	15
--	----

INDICE DE FIGURAS

Figura 1. Diagrama de un cromatógrafo de gases.....	5
Figura 2. Diagrama de flujo programado en KNIME para el curado de la base de datos de compuestos orgánicos volátiles del café.....	14
Figura 3. Índices de retención experimentales en función de los predichos por el modelo QSPR.	17
Figura 4. Dispersión de los residuos de los índices de retención predichos por el modelo QSPR	17

Cabrera Eras Cristhian Joel

Trabajo de Titulación

Dr. Cristian Rojas Villa

Dr. Piercosimo Tripaldi

Septiembre, 2018

MODELADO QUIMIOINFORMÁTICO DE ÍNDICES DE RETENCIÓN DE COMPUESTOS ORGÁNICOS VOLÁTILES DEL CAFÉ.

INTRODUCCIÓN

Hablar del aroma del café, desde un punto de vista químico, sería complejo ya que es una mezcla de diversos compuestos orgánicos volátiles (VOCs, volatile organic compounds) en diferentes concentraciones que se encuentran en el grano de café y otros diversos compuestos que se forman durante el tostado debido a la aplicación de calor (González *et al.*, 2011; Puerta Quintero, 2011). La búsqueda de los VOCs responsables del aroma del café data desde 1926, cuando los químicos Tadeus Reichstein y Hermann Staudinger lograron identificar 13 compuestos orgánicos volátiles involucrados en el aroma, esta cifra fue aumentando lentamente al pasar el tiempo, debido a que las sensaciones olfatorias, como la percepción del umbral de olor difieren en cada persona, son efímeras y difíciles de describir o clasificar, debido a que no existe una escala del olor (Grosch, 2001).

Las relaciones cuantitativas estructura-propiedad (QSPRs) pertenecen a los métodos quimioinformáticos y permiten predecir diversas propiedades fisicoquímicas de un compuesto particular. Esto indica que existe una fuerte correlación entre la estructura molecular y los índices de retención (propiedad observada) de los compuestos orgánicos volátiles que constituyen el perfil aromático y que se detectan mediante cromatografía de gases (GC, Gas Chromatography) (Kaliszan, 1977; Kaliszan & Foks, 1977; Kaliszan, 2007). El índice de retención (I) es un parámetro de la cromatografía de gases que permite investigar los mecanismos que se desarrollan para la retención de la fase móvil (Kaliszan, 1977; Kaliszan & Foks, 1977).

En este trabajo se utilizarán métodos QSPRs, para modelar el índice de retención de compuestos orgánicos volátiles que constituyen el perfil aromático del café, utilizando la información experimental obtenida por GC en una columna polar HP-INNOWAX. El aroma como parámetro fisicoquímico, es relevante para diferenciar varios atributos del café, tales como: calidad, defectos de procesamiento y variedades (Mendizábal de Montenegro *et al.*, 2013). Dado que se encuentra poca literatura relacionada al modelado *in silico* para los VOCs del café, se aplicó el modelado QSPR para encontrar la relación entre la estructura química de los compuestos orgánicos volátiles y el índice de retención de los mismos. Por otra parte,

en la Universidad del Azuay no se evidencian trabajos con respecto a esta temática y este es un punto de partida para el desarrollo de futuras investigaciones en la química de los aromas.

- **Objetivo general:**
Desarrollar un modelo predictivo basado en las relaciones cuantitativas estructura-propiedad para los índices de retención de los compuestos orgánicos volátiles del café.

- **Objetivos específicos:**
 - a. Recopilar la base de datos de VOCs del café y curar la información.
 - b. Digitalizar las estructuras químicas de los VOCs y calcular los descriptores moleculares.

CAPÍTULO I

MARCO TEORICO

1.1. EL CAFÉ

El grano de café es una semilla que proviene del arbusto cafeto, nativa de África, de la familia *Rubiaceae* y del género *Coffea*, del cual se derivan cientos de especies; sin embargo existen dos especies con mayor importancia económica en el mundo, puesto que se consideran comerciales para su cultivo: *Coffea arabica* L. y *Coffea canephora*. Comercialmente se los conoce como café arábigo y café robusta respectivamente. Ambas especies se distinguen por sus características botánicas, genéticas, químicas y morfológicas (González *et al.*, 2011). El mayor cultivo de café arábigo se da principalmente en países de Sudamérica como Colombia y Brasil, además en países asiáticos como la India. Por otra parte, la mayor parte de café robusta se cultiva en Brasil, Indonesia y África (Echeverri *et al.*, 2005). En Ecuador el cultivo de café es una de las principales actividades agrícolas y se encuentra dentro de los 10 cultivos con mayores superficies (Santistevan Méndez *et al.*, 2014).

Las características que definen la calidad del café son las cualidades organolépticas, estado sanitario y la aceptación del consumidor. Atributos como el color, sabor, aroma, amargor, cuerpo y acidez comprenden la calidad física y organoléptica del café (Palomares *et al.*, 2013).

El control sanitario implica evitar la presencia de agentes contaminantes microbianos y químicos en el alimento (Puerta Quintero, 2011). Desde el punto de vista del consumidor se puede distinguir la calidad del café con determinantes simples de expresión como “*me gusta*” y “*no me gusta*”, expresiones que distinguen varios niveles de calidad: calidad inducida por marca o procedencia, calidad esperada del alimento y calidad efectiva; estas escalas miden las características del producto y permite aceptarlo o rechazarlo (Puerta Quintero, 2011).

Para producir el café molido listo para preparar, primero se cosecha la semilla del mismo para luego ser clasificada y despulpada; posteriormente, pasa a la etapa de fermentación y lavado; seguidamente, se procede a exponer el grano al sol por varios días hasta conseguir la humedad óptima de alrededor del 12% (Echeverri *et al.*, 2005; González *et al.*, 2011; Puerta Quintero, 2011). Después de estos procesos inicia la expansión del grano y por tanto el desarrollo de las reacciones químicas. Inmediatamente el grano es molido (molino de rodillo) para eliminar parcialmente la cascarilla (endocarpio), para seguidamente someterlo al proceso de tostado (Echeverri *et al.*, 2005). La reacción de Maillard es el proceso en el cual los azúcares reductores reaccionan con los aminoácidos y tienen una influencia directa en el aroma. Durante este proceso de oscurecimiento no enzimático que se somete a calor inicial de 140°C, se desarrollan los melanoídes que dan el color al café (Grosch, 2001; Puerta Quintero, 2011; Mendizábal de Montenegro *et al.*, 2013). Al aumentar la temperatura para el tueste del grano de café, la porción restante del endocarpio es desprendido, los azúcares se

empiezan a caramelizar dando como resultado el color característico del café y se termina de desarrollar la crepitación, es decir, el sonido que se produce al momento de realizar la tostación del grano junto con los aromas y gases (González *et al.*, 2011; Puerta Quintero, 2011; Mendizábal de Montenegro *et al.*, 2013). De esta manera el aroma del café se intensifica y junto con el sabor constituyen los principales atributos que caracterizan a un café de gran calidad. Finalmente, el grano se deja reposar antes de proceder al molido fino y su respectivo empaçado (Palomares *et al.*, 2013).

1.2. COMPUESTOS ORGÁNICOS VOLÁTILES DEL CAFÉ

Durante años se han realizado investigaciones para dar a conocer los compuestos orgánicos volátiles que forman parte del aroma del café (Palomares *et al.*, 2013). Dichos estudios encontraron varias familias de compuestos volátiles tales como aldehídos, ácidos, alcanos, alquenos, ésteres, lactonas, oxazoles, piridinas, furanos, cetonas, pirroles, pirazinas y compuestos azufrados (Grosch, 2001; Puerta Quintero, 2011). Durante el tostado del grano, estos compuestos varían en cantidad, es decir, algunos aumentan mientras otros disminuyen. Estas variaciones se deben a que los VOCs son derivados de dos o más precursores los cuales tienen una degradación heterogénea (González *et al.*, 2011; Puerta Quintero, 2011). Por otra parte, los VOCs poseen propiedades antioxidantes y antimicrobianas (Palomares *et al.*, 2013). Se estima que en 1 kg de café tostado se encuentran 500 mg de compuestos volátiles (Puerta Quintero, 2011) y se pueden encontrar compuestos químicos con distintas clases de aromas, tales como a tostados, caramelo, almendras, cítricos, frutales, etc. La composición final de los compuestos aromáticos volátiles del café está fuertemente relacionada a la especie y variedad de café, almacenamiento, tiempo y temperatura de tostado del grano (Grosch, 2001; Puerta Quintero, 2011). Cabe destacar que casi la mitad de los compuestos orgánicos volátiles producidos durante el tostado del café se pierden al momento de la molienda y almacenamiento (Grosch, 2001).

1.3. CROMATOGRAFÍA DE GASES E ÍNDICES DE RETENCIÓN

La cromatografía de gases (GC, Gas Chromatography) es un método instrumental de análisis que es capaz de producir información sobre la composición cualitativa y cuantitativa de mezclas de compuestos orgánicos volátiles (Schomburg, 1990). Los resultados cuantitativos y cualitativos de cromatografía solo se pueden obtener después de la separación de los componentes de la mezcla. En condiciones óptimas de separación se da un flujo continuo de señales eléctricas que forman el cromatograma (Schomburg, 1990), el cual recopila la información de todos los componentes en los que se puede separar la mezcla y brinda también información cuantitativa de cada componente eluido en el sistema de cromatografía. En la Figura 1 se presenta el diagrama de un equipo de cromatografía de gases.

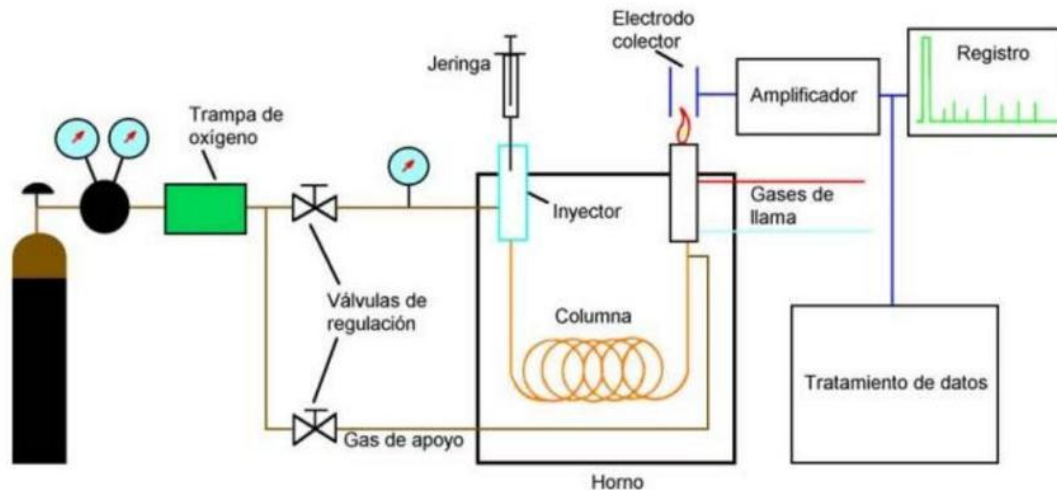


Figura 1. Diagrama de un cromatógrafo de gases (Skoog *et al.*, 2001)

En la cromatografía de gases se usa una fase móvil que es un medio gaseoso y una fase estacionaria en una columna de separación (Schomburg, 1990; Higson, 2007). Para realizar la separación por partición de muestras gaseosas entre el gas de arrastre y la fase estacionaria, primero se debe tener la muestra en fase gaseosa y luego insertarla en la corriente gaseosa de arrastre que se encuentra generalmente a una temperatura de entre 150°C y 250°C, el mismo provoca que los componentes volátiles de la muestra se vaporicen. El componente más rápido es el primero en salir (eluir) de la columna, seguido por el resto de componentes en el orden correspondiente. Para la elución de la muestra se utiliza un gas de arrastre de alta pureza y químicamente inerte. Algunos ejemplos de fase móvil son: nitrógeno, helio, dióxido de carbono o argón, los gases con mayor peso o densidad permiten eluciones más eficaces pero lentas, mientras que gases con menor peso generan eluciones más rápidas pero poco definidas (Higson, 2007).

La cromatografía es por excelencia una técnica analítica para el análisis de compuestos orgánicos volátiles, debido a su precisión y exactitud; sumado a que sus tiempos de análisis son relativamente cortos (Schomburg, 1990). En cromatografía se pretende medir un parámetro analítico denominado índice de retención, el cual fue propuesto por Kovats en 1958 (Rojas-Vargas *et al.*, 2012). Este parámetro se define como datos de retención relativos de compuestos estandarizados con relación a las retenciones de los *n*-alcanos homólogos a una determinada temperatura (Schomburg, 1990). El índice de retención de los *n*-alcanos es igual al producto entre el número de carbonos multiplicado por 100. Por ejemplo, para el *n*-undecano ($n\text{-C}_{11}\text{H}_{24}$) el índice de retención es $I = 1100$. Los índices de retención permiten estandarizar las variables del sistema GC, de tal forma que los datos se pueden comparar en diversos sistemas de cromatografía gaseosa (Schomburg, 1990). Generalmente se trabaja con temperatura fija (isotérmica) o con temperatura programada, en la cual se parte de una temperatura inicial y se eleva linealmente hasta una temperatura máxima. Este segundo método tiene la ventaja de acelerar el análisis y producir particularmente picos más simétricos.

El índice de Kovats se utiliza para el método isotérmico, mientras que para el método de temperatura programable se utiliza una variación del índice de Kovats (Schomburg, 1990).

Para temperatura constante se tiene:

$$I = 100y + 100(z - y) \frac{\log t_{R(x)} - \log t_{R(y)}}{\log t_{R(z)} - \log t_{R(y)}} \quad (\text{Ec. 1})$$

donde,

t_R = tiempo de retención ($t_M - t_S$) (tiempo muerto – tiempo neto de retención)

x = soluto de interés

y = alcano normal con número y de átomos de carbono que eluyen antes del soluto x

z = alcano normal con número z de átomos de carbono que eluyen después del soluto x

$z-y$ = diferencia en número de átomos de carbono existente entre los dos alcanos normales

Para temperatura programable se usa la siguiente expresión:

$$I = 100 \left(\frac{t_{R(x)} - t_{R(y)}}{t_{R(z)} - t_{R(y)}} \right) + y \quad (\text{Ec. 2})$$

Tiempo de retención total (t_R)

Es el tiempo que un soluto tarda en recorrer la columna por completo. Se trata de la suma del tiempo que todas las moléculas pasan por la fase estacionaria y la fase móvil (Razmilic, 1994).

Tiempo muerto (t_M) y Factor de retención (k)

El tiempo de retención (t_M o t_0), también conocido como tiempo muerto, es el tiempo de retención que un compuesto que no es retenido tarda en recorrer toda la columna. Este parámetro se obtiene inyectando un compuesto de este tipo y determinando el tiempo que tarda desde la inyección hasta la elución en el detector (Razmilic, 1994). Por otra parte, el factor de retención (k), también conocido como relación de partición o factor de capacidad, se define como la relación entre la cantidad de sustancia en la fase estacionaria y la cantidad de sustancia en la fase móvil. Se calcula mediante la siguiente ecuación (Razmilic, 1994):

$$k = \frac{t_R - t_M}{t_M} \quad (\text{Ec. 3})$$

1.4. RELACIONES CUANTITATIVAS ESTRUCTURA-PROPIEDAD

Las relaciones cuantitativas estructura-propiedad (QSPRs) son una herramienta útil para cuantificar y sistematizar la relación existente entre la estructura y la propiedad de las sustancias químicas. Las QSPRs son modelos matemáticos asistidos por computadora (*in silico*), los cuales relacionan la propiedad fisicoquímica de los compuestos con su estructura química codificada dentro de variables teóricas denominadas descriptores moleculares (Roy

et al., 2015). Todos los estudios de la teoría QSPR son realizados en base de la correlación entre los valores experimentales y los descriptores moleculares teóricos que se derivan de la estructura química de los compuestos respectivos. Los modelos QSPR encontrados en estudios previos se pueden usar para predecir las propiedades de los compuestos que no se han medido o incluso no han sido aún sintetizados. Por consiguiente, este enfoque QSPR genera un menor gasto de recursos y agiliza el proceso de desarrollo de nuevas moléculas (Katritzky *et al.*, 2000). Para establecer una relación QSPR es imprescindible calibrar el modelo con un conjunto de moléculas para las cuales se conoce los valores experimentales de la propiedad. Este conjunto se denomina de calibración. La función matemática estructural con la que se construye el modelo se escoge empleando el método de reemplazo y el criterio práctico es seleccionar aquella que brinde los mejores resultados.

CAPÍTULO II

METODOLOGIA

2.1. GENERACIÓN DE LA BASE DE DATOS

El primer paso para desarrollar un modelo QSPR es elaborar una base de datos extensa y robusta. Estas características determinan la fiabilidad de la investigación. Para construir la base de datos inicial se recopiló información de moléculas relacionadas con el perfil aromático del café de la literatura especializada:

- Utilización del índice de retención lineal para caracterización de compuestos volátiles con café soluble utilizando GC-MS y columna HP-INNOWAX (Viegas & Bassoli, 2007).
- Manual de aromas (Yeretzian, 2017).
- Química III: Compuestos volátiles (Grosch, 2001)

Además se recopiló para cada molécula su respectivo índice de retención experimental, en una columna polar HP-INNOWAX programada con rampa de temperatura. Se obtuvo el número de registro CAS (*Chemical Abstracts Service*) y la notación lineal de cadena SMILES (*Simplified Molecular Input Line Entry System*) de todos los compuestos a partir de diversas librerías químicas de acceso libre: PubChem (Kim *et al.*, 2015), ChemSpider (Pence & Williams, 2010) y NIST Chemistry WebBook (Linstrom & Mallard, 2001). Esta información será utilizada en el proceso de curado de los datos para eliminar compuestos repetidos y aquellos que tengan misma notación SMILES.

2.2. CURADO DE LA INFORMACIÓN

Para realizar el curado de la información se utilizó el software de minería de datos KNIME (Konstanz Information Miner) (Berthold *et al.*, 2008). Este software brinda acceso a nodos preinstalados y también permite añadir extensiones de nodos de otros programas (Berthold *et al.*, 2008). Se inicia el curado de la información con el acceso a los datos, el cual posteriormente fue filtrado a través de nodos hasta llegar a la visualización y reporte de resultados. Para el filtrado de la base de datos inicial, se definieron dos criterios de agrupamiento: 1) nombre químico y 2) notación lineal de cadena SMILES canónico. Una vez que se identificaron los compuestos duplicados con el mismo código SMILES, se procedió a fusionarlos y usar el índice de retención promedio como propiedad experimental.

2.3. DISEÑO MOLECULAR Y CÁLCULO DE DESCRIPTORES MOLECULARES

La estructura de los compuestos orgánicos volátiles del café se diseñaron en el programa HyperChem (Hypercube Inc.) y sus geometrías fueron optimizadas mediante los campos de

fuerza de la mecánica molecular (MM+), mediante el uso del algoritmo de gradiente conjugado en la versión Polak-Ribiere hasta que la desviación estándar del vector gradiente sea menor a $0.01 \text{kcal} \times (\text{Å} \times \text{mol})^{-1}$. La mecánica molecular está basado en el modelado matemático de una molécula compuesta por átomos que se mantienen unidos por enlaces. Utiliza los parámetros de fuerza de tensión y flexión de enlace, lo cual permite interacciones entre los átomos no enlazados (Valles-Sánchez *et al.*, 2014).

Para el cálculo de los descriptores moleculares y huellas dactilares moleculares se utilizaron los programas PaDEL-Descriptor versión 2.21 (Yap, 2011) y Dragon versión 7 (Kode srl., 2018). El programa PaDEL-Descriptor calcula 1875 descriptores (1444 1D, 2D descriptores y 431 3D descriptores) y 12 tipos de huellas dactilares moleculares (Yap, 2011). Por otra parte, el programa Dragon calcula 5270 descriptores moleculares.

Un descriptor molecular se define como el resultado de un procedimiento lógico que transforma la información química codificada de una estructura molecular en información matemática útil para establecer una relación estructura-propiedad (Todeschini & Consonni, 2009). Las huellas dactilares moleculares (FPs, FingerPrints) son vectores booleanos (cuyos elementos se denominan bits) de dimensión fija que definen un conjunto de patrones de la molécula en un índice. Las FPs se generan de tal manera que permitan capturar las características estructurales locales presentes en las moléculas; particularmente identifican grupos de fragmentos que componen la estructura molecular (Shemetulskis *et al.*, 1996).

Después de calcular los descriptores moleculares y huellas dactilares moleculares que serán usados en la elaboración del modelo, se utilizará el software MatLab (The MathWorks Inc.) para la aplicación de diversas funciones como la reducción no supervisada de descriptores, partición de la base de datos, selección supervisada de descriptores y validación interna del modelo.

2.4. REDUCCIÓN NO SUPERVISADA DE DESCRIPTORES MOLECULARES

Los métodos de reducción de variables no supervisados (no consideran la respuesta experimental) están destinados a reducir la presencia de ruido, redundancia y multicolinealidad en los datos; es decir, selecciona un subconjunto de variables capaces de preservar la información representativa que poseen los descriptores moleculares. El método V-WSP es una modificación del algoritmo propuesto por Wootton, Sergent y Phan-Tan-Luu (WSP) para el diseño de experimentos, que permite la retención de únicamente descriptores relevantes para el desarrollo del modelo (Ballabio *et al.*, 2014). Para el análisis de correlación entre pares de variables se realiza una comparación entre el coeficiente absoluto de correlación de Pearson (R_{ij}) y un valor de umbral de corte preestablecido (thr), por ejemplo 0.95. Si se cumple que $R_{ij} \geq thr$, se eliminará el descriptor que presenta una mayor correlación promedio con las demás variables.

2.5. SELECCIÓN SUPERVISADA DE DESCRIPTORES MOLECULARES

Para desarrollar el modelo QSPR basado en regresión de mínimos cuadrados ordinarios (OLS, ordinary least squares), se utilizó el Método de Reemplazo (RM, Replacement Method) (Duchowicz *et al.*, 2005; Duchowicz *et al.*, 2006) como método de selección supervisada de descriptores. El RM es un algoritmo secuencial de reemplazo de variables de tal forma de minimizar la desviación estándar residual (S). El RM consiste en calcular el error relativo de cada coeficiente de regresión dentro de un modelo de dimensión d (número de variables presentes en el modelo). Posteriormente se reemplazan las variables que presenten el mayor error relativo por aquellas variables presentes en el conjunto total D (número total de descriptores disponibles).

2.6. VALIDACIÓN DEL MODELO

Para la validación externa del modelo QSPR, se dividió la base de datos en tres grupos: calibración (cal), validación (val) y predicción (pred). Esta partición se realizó según el método de subconjuntos balanceados (BSM, Balanced Subsets Method) (Rojas *et al.*, 2015). El método BSM se fundamenta en análisis de conglomerados k -medias (k -MCA, k -Means Cluster Analysis). Por consiguiente, el BSM crea k -conglomerados o grupos de moléculas en términos de una medida de distancia, de tal manera que los compuestos en el mismo conglomerado son muy similares entre sí y los compuestos en diferentes grupos son estructuralmente diversos. La partición BSM considera la propiedad experimental y únicamente los descriptores independientes de la conformación, después de la exclusión de aquellos linealmente dependientes. Los grupos de calibración y validación se usaron para la selección supervisada de descriptores, mientras que el de predicción se utilizó para medir la capacidad predictiva del modelo.

Adicionalmente, se aplicó la validación interna o cruzada de dejar-uno-fuera (loo, leave-one-out) y dejar-varios-fuera (lmo, leave-many-out) (Burden *et al.*, 1997; Baumann & Stiefl, 2004; Arlot & Celisse, 2010). En loo se excluye una molécula del modelo a la vez, luego se recalibra el modelo y se utiliza para predecir su propiedad. De forma análoga, en lmo se excluye un porcentaje definido de moléculas (por ejemplo el 20%) y se usan las restantes para recalibrar el modelo y predecir la propiedad de las moléculas excluidas. Finalmente, se aplicó la aleatorización-Y (Rücker *et al.*, 2007) para descartar la presencia de correlación casual en el modelo. Esta metodología consiste en modificar de forma casual los elementos del vector respuesta, de tal manera que no correspondan a las asignaciones originales.

2.7. ANÁLISIS DEL DOMINIO DE APLICABILIDAD

El dominio de aplicabilidad (AD, Applicability Domain) del modelo se define como un espacio teórico que depende de la naturaleza de los descriptores moleculares y la propiedad experimental (Jaworska *et al.*, 2005). De esta manera, el modelo se encuentra acotado dentro

de un espacio teórico definido por la información química proporcionada por las moléculas del conjunto de calibración. La predicción de dicho modelo para las moléculas del grupo de predicción se restringe únicamente a aquellos compuestos que son estructuralmente similares a los compuestos presentes en el conjunto de calibración.

Una forma de definir el AD del modelo de regresión lineal es el enfoque de influencia (Rencher & Schaalje, 2008), que se basa en el cálculo del valor de influencia (h^*) y el valor de influencia para cada i -ésimo compuesto del grupo de predicción (h_i). El valor de influencia de cada molécula del grupo de predicción mide su distancia con respecto al centro del modelo. La matriz de influencia H se calcula partiendo de la matriz del modelo X .

$$H = X(X^T X)^{-1} X^T \quad (\text{Ec. 4})$$

Los elementos diagonales (h_i) de la matriz H son los valores de influencia de cada compuesto. El valor (h_i) es la contribución de cada i -ésimo elemento en la estimación de su respuesta. Para el conjunto de entrenamiento (calibración, validación) estos valores están acotados entre valores de 0 a 1, mientras que los valores calculados para el conjunto de predicción están acotados con valores inferiores a 0. Se define un valor umbral límite (h^*) para el análisis del AD, el cual se calcula como 3 veces el valor promedio de los valores de influencia del grupo de calibración:

$$h^* = 3 \frac{p}{n} \quad (\text{Ec. 5})$$

donde p es el número de parámetros del modelo y n es el número de compuestos en el grupo de calibración. El valor de influencia para una molécula del grupo de predicción se calcula de la siguiente manera:

$$h_i = x_i (X^T X)^{-1} x_i^T \quad (\text{Ec. 6})$$

De esta forma, solo las moléculas que caen dentro de este espacio teórico ($h_i < h^*$) se consideran predicciones confiables o interpolación modelo; caso contrario, las predicciones se consideran extrapolación del modelo (predicciones poco confiables).

2.8. MECANISMO DE ACCIÓN (INTERPRETACIÓN DE LOS DESCRIPTORES).

La interpretación de un modelo QSPR hace referencia a la posibilidad de establecer una ruta por la cual la propiedad experimental se describe mediante los descriptores moleculares del modelo *in silico*. El grado de contribución de los descriptores se obtiene ordenando los valores absolutos de los coeficientes estandarizados de los descriptores (Draper & Smith, 2014). Cuanto mayor es el valor absoluto del coeficiente para un descriptor definido, mayor es la importancia de tal descriptor en la predicción del índice de retención. Finalmente, se espera realizar una explicación del significado de cada descriptor y cómo se relaciona con la

propiedad experimental (si es posible). Este hecho se debe a la dificultad para explicar el significado de algún descriptor, debido a que provienen de diversas teorías, en algunos casos a nivel abstracto. Esta interpretación del mecanismo de acción contribuye al conocimiento de cómo los descriptores moleculares describen el fenómeno del índice de retención.

CAPÍTULO III

RESULTADOS Y DISCUSIÓN

3.1. DESCRIPCIÓN DE LA BASE DE DATOS

La información experimental se recopiló de la literatura especializada: 88 moléculas procedentes de Viegas y Bassoli (Viegas & Bassoli, 2007), para las cuales se reporta el índice de retención medido en la columna polar HP-INNOWAX. Adicionalmente, se obtuvieron 184 moléculas de Grosch (Grosch, 2001) y 96 moléculas de Yeretizian (Yeretizian, 2017); dando un total de 368 compuestos orgánicos volátiles presentes en el café. Para cada compuesto se verificó el número de registro CAS y notación lineal de cadena SMILES canónico en las bibliotecas químicas PubChem (Kim *et al.*, 2015), ChemSpider (Pence & Williams, 2010) y NIST Chemistry WebBook (Linstrom & Mallard, 2001).

3.2. CURADO DE LA BASE DE DATOS

Para elaborar un modelo que sea independiente de la conformación se deben identificar únicamente aquellos compuestos con una misma notación lineal SMILES canónico. Para esto se utilizó el software de minería de datos KNIME (Berthold *et al.*, 2008). En este programa se aplicó el nodo de agrupamiento para identificar aquellos compuestos repetidos por nombre y particularmente aquellos VOCs que tienen la misma notación lineal de cadena SMILES. Para aquellos compuestos que tienen el mismo código SMILES se usó el índice de retención promedio para el desarrollo del modelo QSPR. De esta manera se redujeron de 368 moléculas a 114 compuestos orgánicos volátiles.

Una vez curada la base de datos se procedió a recopilar los índices de retención experimentales medidos en una columna polar HP-INNOWAX para 26 VOCs a partir de la biblioteca química NIST Chemistry WebBook (Linstrom & Mallard, 2001) y Pherobase (El-Sayed, 2018), obteniendo así una total de 114 compuestos con sus respectivos índices de retención. Cabe destacar que si se obtenía más de un índice de retención para un mismo compuesto, se calculó el índice de retención promedio como propiedad experimental. En la Figura 2 se presenta el diagrama de flujo programado en KNIME.

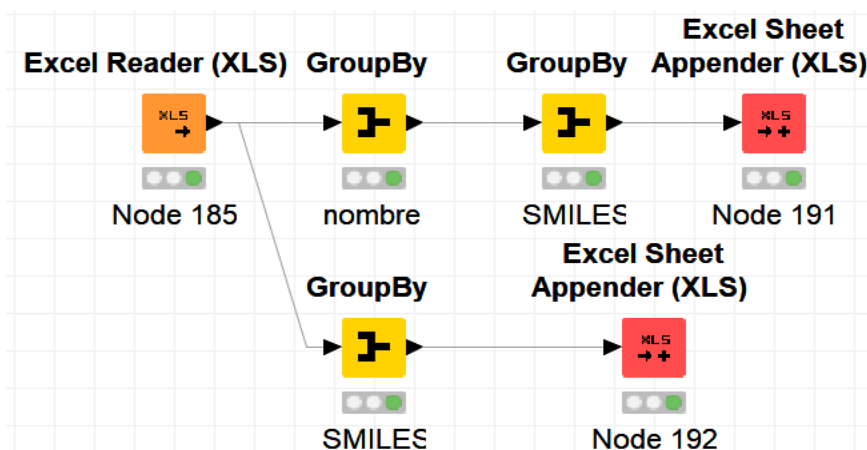


Figura 2. Diagrama de flujo programado en KNIME para el curado de la base de datos de compuestos orgánicos volátiles del café.

3.3. DISEÑO MOLECULAR Y CÁLCULO DE DESCRIPTORES MOLECULARES

Los 114 compuestos obtenidos en el curado de datos, fueron modelados en el programa HyperChem (Hypercube Inc.) y posteriormente, se procedió a calcular los descriptores moleculares y huellas dactilares moleculares usando el programa PaDEL-Descriptor versión 2.21 y Dragon versión 7, de la siguiente manera:

- 3819 descriptores moleculares de Dragon.
- 1444 descriptores moleculares de PaDEL-Descriptor.
- 13020 huellas dactilares moleculares de PaDEL-Descriptor.

De esta manera, se obtienen 18283 descriptores moleculares independientes de la conformación. Para reducir la dimensionalidad de la base de datos se excluyeron descriptores con valores constantes (o casi constantes) y descriptores con al menos un valor faltante; de tal forma que se retuvieron 4029 descriptores moleculares y huellas dactilares moleculares, según el siguiente detalle:

- 1752 Descriptores moleculares de Dragon.
- 1014 Descriptores moleculares de PaDEL-Descriptor.
- 1263 Huellas dactilares moleculares de PaDEL-Descriptor.

3.4. REDUCCIÓN NO SUPERVISADA DE DESCRIPTORES MOLECULARES

Para la reducción no supervisada de descriptores moleculares se usó el algoritmo V-WSP implementado en MatLab el cual reduce la presencia de ruido, redundancia y multicolinealidad en los datos. Este algoritmo matemático selecciona una variable inicial d (por ejemplo 1) y un valor umbral de correlación predefinido ($thr = 0.95$). Para el análisis de correlación entre pares

de variables se realizó una comparación entre el coeficiente absoluto de correlación de Pearson (R_{ij}) y el valor de umbral pre-establecido ($thr = 0.95$). Si se cumple que $R_{ij} \geq thr$, se eliminará el descriptor que presenta una mayor correlación promedio con las demás variables. De esta manera se excluyeron 2476 descriptores moleculares, para obtener 1553 descriptores moleculares para la etapa posterior de desarrollo del modelo QSPR.

3.5. DIVISIÓN DE LA BASE DE DATOS

Para la división de la base de datos por el método de subconjuntos balanceados (BSM) se utilizó la matriz inicial de datos de 114 VOCs y 4029 descriptores moleculares más el vector de índices de retención. El algoritmo del BSM comienza con la exclusión de descriptores linealmente dependientes, de tal forma que se retienen 113 variables. Posteriormente, el método identifica los valores máximo y mínimo de la propiedad experimental y las coloca directamente en el grupo de calibración (de esta manera se garantiza interpolación del modelo). Entonces al algoritmo divide la base de datos en grupos de calibración, validación y predicción de tal forma que se mantenga una relación estructura-propiedad en los tres grupos. Esta partición se logra luego de 5000 interacciones y coloca 38 VOCs en cada uno de los tres grupos.

Tabla 1. Resumen de los mejores modelos QSPR de 1 a 7 descriptores moleculares obtenidos mediante el método de reemplazo.

d	R_{cat}^2	S_{cat}	R_{val}^2	S_{val}	$R_{ij\ max}^2$	Descriptores
1	0.604	281,6	0,491	259,2	0	SM1_B(s)
2	0,818	193,7	0,694	202,7	0,04	SM2_B(m), P_VSA_v_2
3	0,909	138,9	0,854	143,6	0,09	maxHBd, nAtomP, MLFER_L
4	0,949	105,7	0,769	193,4	0,05	SM1_B(p), P_VSA_v_2, KRFP4413, MLFER_E
5	0,979	68,0	0,866	147,5	0,08	Chi0_EA(ed), DELS, KRFP4413, MLFER_A, MLFER_E
6	0,979	69,7	0,820	169,1	0,34	Chi0_EA(ed), P_VSA_v_2, MLFER_A, MLFER_E, ATSC6e, Eig06_EA(dm)
7	0,988	52,8	0,877	140,8	0,29	KRFP4413, MLFER_A, SM1_B(s), Kier2, nHBAcc3, MLFER_E, ATSC8e

d: número de descriptores; R^2 : coeficiente de determinación; $R_{ij\ max}^2$: coeficiente de determinación máximo entre descriptores; S: desviación estándar residual.

3.6. SELECCIÓN SUPERVISA DE DESCRIPTORES MOLECULARES

Los grupos de calibración y validación se usaron para la selección supervisada mediante el método de reemplazo. El grupo de calibración permite realizar los ajustes por mínimos cuadrados ordinarios, mientras que el grupo de validación evita la presencia de sobreajuste

en el modelo. Para la selección por el RM, se usaron los 1553 descriptores obtenidos luego de aplicar el algoritmo V-WSP. De esta manera se generaron modelos de 1 a 7 descriptores (Tabla 1).

La selección del modelo óptimo se realizó considerando los parámetros de calidad del grupo de calibración y validación, es decir la menor desviación estándar residual (S) y el mayor coeficiente de correlación (R^2) en validación; para así evitar la presencia de sobreajuste del modelo. De la misma manera, se buscó que el coeficiente de correlación máximo entre descriptores ($R_{ij\ max}^2$) sea el más bajo posible. También se utilizó el principio de parsimonia de Ockham (Hoffmann *et al.*, 1996), el cual sugiere que en igualdad de calidad de diversos modelos se elija el más sencillo. Así, se procedió a elegir el mejor modelo de 3 descriptores (resaltado en negrita en la Tabla 1) independientes de la conformación para un análisis más detallado:

$$I = 376.4 + 613.1 \max HBd + 63.6 nAtomP + 177.8 MLFER_L \quad (\text{Ec. 7})$$

$$N_{cal} = 38, R_{cal}^2 = 0.909, S_{cal} = 138.9$$

$$N_{val} = 38, R_{val}^2 = 0.854, S_{val} = 143.6$$

$$N_{pred} = 38, R_{pred}^2 = 0.845, S_{pred} = 151.7$$

$$R_{loo}^2 = 0.875, S_{loo} = 144.2, R_{lmo}^2 = 0.846, S_{lmo} = 160.4$$

$$R_{rand}^2 = 0.248, S_{rand} = 353.6, o(3S) = 1, R_{ij\ max}^2 = 0.09$$

3.7. VALIDACIÓN DEL MODELO

El modelo desarrollado se sometió a validación interna o cruzada, particularmente con la técnica de dejar-uno-fuera (*loo*) y dejar-varios-fuera (*lmo*). En el procedimiento *loo* se excluyó una molécula a la vez y las restantes (75 moléculas) se usaron para recalibrar el modelo y predecir la propiedad del compuesto excluido. En la técnica *lmo* se excluyó de forma aleatoria el 20% de moléculas (15 moléculas) y se utilizó las restantes para recalibrar el modelo y predecir la propiedad de las moléculas excluidas. El número de casos analizados para la exclusión aleatoria de datos es de 5000 iteraciones. Por otra parte, la aleatorización-Y permite verificar la estabilidad del modelo luego de 10000 iteraciones. Se observa que la S_{cal} (138,9) < S_{rand} (353,6), lo que confirma la ausencia de correlación casual en el modelo de la Ec. 7. En la Figura 3 se muestran los valores de índices de retención experimentales en función de los

predichos por el modelo QSPR. Se observa la tendencia lineal alrededor de la recta de ajuste perfecto.

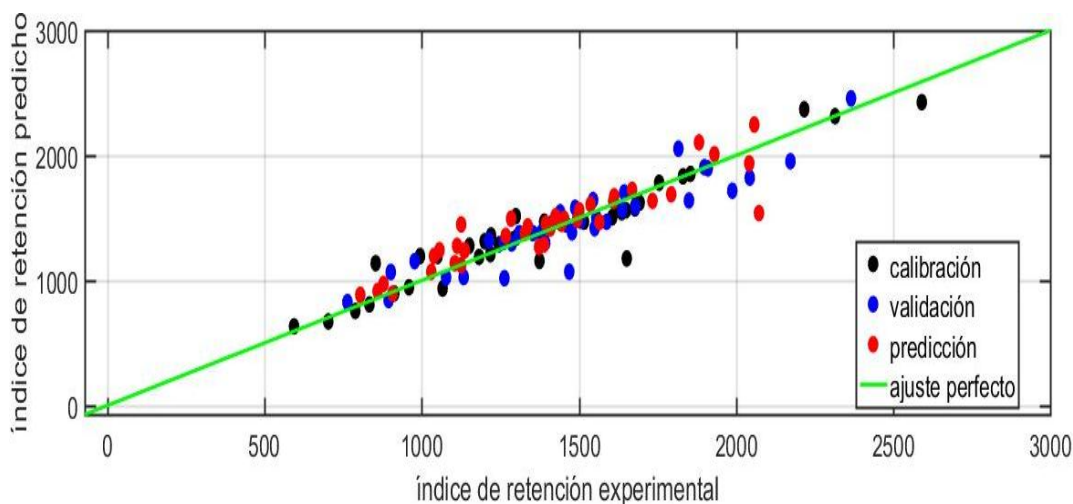


Figura 3. Índices de retención experimentales en función de los predichos por el modelo QSPR.

Por otra parte, en la Figura 4 se observan los índices de retención predichos en función de los residuos. Aquí se observa que los residuos siguen un patrón aleatorio alrededor de la línea cero, indicando de esta manera que el modelo de mínimos cuadrados ordinarios es apropiado para modelar los índices de retención. El único compuesto que tiene valor de residuo mayor a 3 veces la desviación estándar es el *Butyrolactone* (número CAS 96-48-0). Para este compuesto se verificó que su fórmula química y su índice de retención experimental sean correctos en las fuentes bibliográficas. Se asume que su comportamiento atípico se deba a la diversidad química de los VOCs considerados en esta base de datos, así como a otros aspectos inherentes a la medición de la propiedad experimental en GC (Rojas *et al.*, 2015).

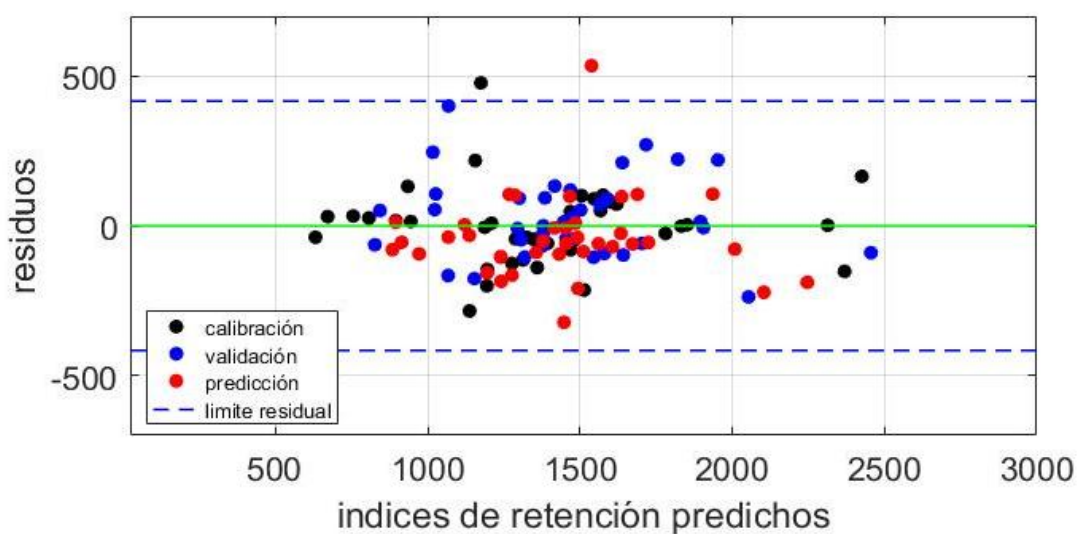


Figura 4. Dispersión de los residuos de los índices de retención predichos por el modelo QSPR

3.8. DOMINIO DE APLICABILIDAD

En el análisis del AD se estableció un valor crítico $h^* = 0.1579$. Al comparar el valor de influencia de las 38 moléculas del grupo de predicción con respecto a h^* , se observa que no existen VOCs que superen dicho umbral, por lo que sus predicciones son confiables.

3.9. MECANISMO DE ACCIÓN

Los descriptores moleculares que mejores condiciones desarrollaron para la construcción del modelo fueron *maxHBd*, *nAtomP* y *MLFER_L* que pertenecen al grupo de descriptores 2D o topológicos, los descriptores *maxHBd* y *nAtomP* presentan la máxima correlación ($R_{ij}^2_{max} = 0,09$) indicando que no existe multicolinealidad entre los descriptores del modelo. Por consiguiente, cada descriptor explica aspectos particulares en el mecanismo de retención de los compuestos orgánicos volátiles en una columna polar HP-INNOWAX.

El índice de estados E máximos para donantes (fuertes) de enlaces de hidrógeno (*maxHBd*), está relacionado con la electronegatividad efectiva de un átomo, la cual depende de la influencia de átomos que existen alrededor del mismo en la molécula (Hall & Kier, 1995). Por otra parte, el índice de coeficiente de partición soluto gas-hexadecano (*MLFER_L*) está relacionado con la energía libre de solvatación calculada según el modelo aditivo propuesto por Platts y colaboradores (Platts *et al.*, 1999), en el cual se calcula el logaritmo de la constante de distribución vapor-líquido como suma de diferentes aportes de la estructura de características moleculares de la molécula en fase de vapor. Entre estas características moleculares se tiene en cuenta la polarizabilidad de la molécula. Finalmente, el índice del número de átomos en el sistema pi (π) más grande (*nAtomP*) está relacionado con el número de átomos con más enlaces (π), es decir, con más conjugación o polarizabilidad de los electrones en la molécula del compuesto orgánico volátil. Se puede ver que al aumentar el valor de estos tres descriptores aumenta el índice de retención, en cuanto a su signo en el modelo es siempre positivo, por lo tanto aumentando la polaridad de la molécula en fase de vapor (soluto) aumenta también su solubilidad en la fase estacionaria polar, por esta razón es retenida más que una molécula menos polar.

CONCLUSIONES

Los índices de retención de los compuestos volátiles del café recopilados de literatura especializada fueron descritos y predichos en un modelado QSPR de tres descriptores independientes de la conformación. Este estudio confirma los resultados obtenidos para los índices de retención medidos en las columnas HP-INNOWAX; es decir, los modelos independientes de la conformación son apropiados para modelar y predecir esta propiedad. El modelo desarrollado en este trabajo de tesis es una herramienta útil para dar un punto de partida a futuras investigaciones en la química de los aromas, además de aportar una base de datos sólida y robusta que dé a conocer los beneficios en el reconocimiento de la relación entre la estructura química de los compuestos orgánicos volátiles y el índice de retención de los mismos.

BIBLIOGRAFIA

- Arlot, S., & Celisse, A. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*, 4, 40-79.
- Ballabio, D., Consonni, V., Mauri, A., Claey's-Bruno, M., Sergent, M., & Todeschini, R. (2014). A Novel Variable Reduction Method Adapted from Space-Filling Designs. *Chemometrics and Intelligent Laboratory Systems*, 136, 147-154.
- Baumann, K., & Stiefl, N. (2004). Validation Tools for Variable Subset Regression. *Journal of Computer-Aided Molecular Design*, 18(7), 549-562.
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications*, (pp. 319-326): Springer Berlin Heidelberg.
- Burden, F. R., Brereton, R. G., & Walsh, P. T. (1997). Cross-Validatory Selection of Test and Validation Sets in Multivariate Calibration and Neural Networks as Applied to Spectroscopy. *Analyst*, 122(10), 1015-1022.
- Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (Vol. 326): John Wiley & Sons.
- Duchowicz, P. R., Castro, E. A., Fernández, F. M., & Gonzalez, M. P. (2005). A New Search Algorithm for QSPR/QSAR Theories: Normal Boiling Points of Some Organic Molecules. *Chemical Physics Letters*, 412(4), 376-380.
- Duchowicz, P. R., Castro, E. A., & Fernández, F. M. (2006). Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies. *MATCH Communications in Mathematical and in Computer Chemistry*, 55, 179-192.
- Echeverri, D., Buitrago, L., Montes, F., Mejía, I., & González, M. d. P. (2005). Coffee for cardiologists. *Revista Colombiana de Cardiología*, 11(8), 357-365.
- El-Sayed, A. M. (2018). The Pherobase: Database of Pheromones and Semiochemicals, <http://www.pherobase.com>.
- González, H., González, P., & Rosales, R. (2011). Café (*Coffea arabica* L.): Compuestos volátiles relacionados con el aroma y sabor. *Unacar Tecnociencia*, 5, 35-45.
- Grosch, W. (2001). Chemistry III: volatile compounds. In R. J. Clarke & O. G. Vitzthum (Eds.), *Coffee: recent developments*, (pp. 68-89): Blackwell Science Ltd.
- Hall, L. H., & Kier, L. B. (1995). Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *Journal of chemical information and computer sciences*, 35(6), 1039-1045.
- Higson, S. (2007). *Química analítica*: McGraw-Hill.
- Hoffmann, R., Minkin, V. I., & Carpenter, B. K. (1996). Ockham's Razor and Chemistry. *Bulletin de la Société chimique de France*, 133(2), 117-130.
- Hypercube Inc. HyperChem. <http://www.hyper.com>.
- Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR Applicability Domain Estimation by Projection of the Training Set Descriptor Space: A Review. *ATLA*, 33(5), 445-459.

- Kaliszan, R. (1977). Correlation between the Retention Indices and the Connectivity Indices of Alcohols and Methyl Esters with Complex Cyclic Structure. *Chromatographia*, 10(9), 529-531.
- Kaliszan, R., & Foks, H. (1977). The Relationship between the R_M Values and the Connectivity Indices for Pyrazine Carbothioamide Derivatives. *Chromatographia*, 10(7), 346-349.
- Kaliszan, R. (2007). QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. *Chemical reviews*, 107, 3212-3246.
- Katritzky, A. R., Chen, K., Maran, U., & Carlson, D. A. (2000). QSPR Correlation and Predictions of GC Retention Indexes for Methyl-Branched Hydrocarbons Produced by Insects. *Analytical Chemistry*, 72(1), 101-109.
- Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2015). PubChem Substance and Compound databases. *Nucleic acids research*, 44(D1), D1202-D1213.
- Kode srl. (2018). Dragon version 7. Software for Molecular Descriptor Calculation. <http://chm.kode-solutions.net/>.
- Linstrom, P. J., & Mallard, W. G. (2001). NIST Chemistry WebBook, NIST Standard Reference Database Number 69. In). Gaithersburg MD: National Institute of Standards and Technology.
- Mendizábal de Montenegro, A. L., Samayoa, C., & Rolz, C. (2013). Diferenciación del Café de Guatemala por Medio de la Composición Química del Aroma. *Revista de la Universidad Del Valle de Guatemala*, 25, 7-28.
- Palomares, S. G., Reyes, T. R., Humberto, L., Cambero, R., Sánchez, H. M. G., Sobel, E., & de Ciencia, V. G.-C. E. (2013). Caracterización de compuestos volátiles de L. variedad Borbón. *Universo de la Tecnológica*, 14, 3-6.
- Pence, H. E., & Williams, A. (2010). ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*, 87(11), 1123-1124.
- Platts, J. A., Butina, D., Abraham, M. H., & Hersey, A. (1999). Estimation of molecular linear free energy relation descriptors using a group contribution approach. *Journal of chemical information and computer sciences*, 39(5), 835-845.
- Puerta Quintero, G. I. (2011). Composición química de una taza de café. *Avances Técnicos Cenicafé*, 1-12.
- Razmilic, B. (1994). Principios básicos de Cromatografía. *Control de calidad de insumos y dietas acuícolas. México Merck Química Chilena*.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear Models in Statistics*: John Wiley & Sons.
- Rojas-Vargas, J. A., Speck-Planche, A., García-López, A., & Acevedo-Martínez, J. (2012). Modelación del índice de retención en iminas usando descriptores tops-mode. *Revista Cubana de Química*, 24(3).
- Rojas, C., Duchowicz, P. R., Tripaldi, P., & Pis Diez, R. (2015). QSPR Analysis for the Retention Index of Flavors and Fragrances on a OV-101 Column. *Chemometrics and Intelligent Laboratory Systems*, 140, 126-132.
- Roy, K., Kar, S., & Das, R. N. (2015). *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*: Springer.

- Rücker, C., Rücker, G., & Meringer, M. (2007). Y-Randomization and its variants in QSPR/QSAR. *Journal of chemical information and modeling*, 47(6), 2345-2357.
- Santistevan Méndez, M., Julca Otiniano, A., Borjas Ventura, R., & Tuesta Hidalgo, O. (2014). Caracterización de fincas cafetaleras en la localidad de Jipijapa (Manabí, Ecuador). *Ecología Aplicada*, 13, 187-192.
- Schomburg, G. (1990). *Gas chromatography: A practical course*: VCH.
- Shemetulskis, N. E., Weininger, D., Blankley, C. J., Yang, J., & Humblet, C. (1996). Stigmata: An Algorithm to Determine Structural Commonalities in Diverse Datasets. *Journal of Chemical Information and Computer Sciences*, 36(4), 862-871.
- Skoog, D. A., Holler, F. J., & Nieman, T. A. (2001). *Principios de análisis instrumental*: McGraw-Hill Interamericana de España.
- The MathWorks Inc. MatLab. <http://www.mathwoks.com>.
- Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*: Wiley-VCH.
- Valles-Sánchez, A., Rosales-Marines, L., Eugenia, L. E., & Farías-Cepeda, L. (2014). Métodos y Usos de la Química Computacional. *Revista Científica de la Universidad Autónoma de Coahuila*, 6(11), 16-21.
- Viegas, M. C., & Bassoli, D. G. (2007). Utilização do índice de retenção linear para caracterização de compostos voláteis em café solúvel utilizando GC-MS e coluna HP-Innowax. *Química nova*, 30, 2031-2034.
- Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466-1474.
- Yeretzian, C. (2017). Coffee. In A. Buettner (Ed.), *Springer handbook of odor*, (pp. 107-127): Springer.