Universidad Autónoma de Zacatecas
"Francisco García Salinas"
Unidad Académica de Ingeniería Eléctrica
Doctorado en Ciencias de la Ingeniería

Análisis de patrones acústicos para determinar estado de colonias de abejas mediante Regresión Logística Lasso y Descomposición en Valores Singulares

Que como parte de los requisitos

para obtener el grado de:

Doctor en Ciencias de la Ingeniería

Presenta:

Antonio Robles Guerrero

Asesor:

Dr. Efrén González Ramírez

Co-Asesor:

Dr. Tonatiuh Saucedo Anaya

Zacatecas, Zac. Marzo 2019

Resumen

El trabajo de tesis presenta un análisis de un problema de clasificación multiclase para identificar colonias saludables, sin reina y de baja población, mediante el análisis de las señales acústicas. Se estudiaron cinco colonias de abeja carniola en distintas condiciones biológicas. Para grabar el sonido de las colonias de abejas se implementó un sistema de monitoreo basado en una Raspberry Pi 2 y micrófonos omnidireccionales. La metodología de análisis está basada en MFCC y un modelo de clasificación penalizado mediante Regresión Logística Lasso. Los resultados demuestran que es posible identificar cada una de las condiciones bajo estudio con una alta tasa de correcta clasificación. Sin embargo, la parte más interesante del estudio se centra en el análisis de los conglomerados mediante gráficas de dispersión utilizando descomposición en valores singulares. El análisis permitió observar que las muestras de colonias que representan la misma condición se agrupan, además, las colonias pueden generar ligeras variaciones en el patrón y puede estar relacionado con su estado de salud, volumen poblacional y actividad de pecoreo. Finalmente, se calcularon los parámetros de monitoreo para determinar la cantidad de muestras y la longitud que brinden un buen rendimiento y al mismo tiempo el espacio de almacenamiento en memoria sea mínimo.

Abstract

The thesis work presents an analysis of a multiclass classification problem through the analysis of acoustic signals to identify healthy bee colonies, queenless colonies and colonies of low population. Five colonies of Carniola honey bee were studied under different biological conditions. A monitor system based on a Raspberry Pi 2 and omnidirectional microphones was implemented to study the honey bee colonies. The analysis methodology is based on MFCC's and a penalized Lasso logistic regression model. The results show that is possible to identify every condition in the model with a high classification rate. However, the most interesting results came from the cluster analysis of the data through scatters plots of singular value decomposition. The results show that samples that represents the same condition tend to group, also, the colonies can generate slightly different patterns and the data clusters of the same condition tend to be close, the clusters can be affected by healthy conditions, weather, number of bees and ecological factors. Finally, an analysis of the monitoring parameters to determinate the data size and number of samples that offer a good performance and with a minimum memory space requirement was carried out.

Dedicatorias

Dedicado a mi esposa Anita y mi hija Reginita
por quienes me esforzaré todos los días de mi vida.

Agradecimientos

A mi esposa Anita y a mi pequeña Regina por apoyarme, su paciencia, su amor incondicional y por creer en mi

A mi familia, mi abuela Ma. Jesús y mis hermanos Saul y Carolina

A mis asesores por apoyarme en el desarrollo del trabajo

A mis amigos del posgrado Luis Pérez, David Zapata, Sahrai Castro, Karlita Jiménez, Iván Zamora, Manuel, Angél Osiris por su compañía y los buenos momentos

Al programa de Doctorado en Ciencias de la Ingeniería por proveerme de las herramientas para realizar el tema de investigación

A mi amiga Esmeralda Ibarra por ayudarme en la corrección de la segunda publicación

A Eduardo Santos por su ayuda y consejos en la electrónica del sistema de monitoreo

A CONACYT por la beca otorgada.

Índice general

Resumen III

Lista de figuras X

Lista de tablas XII

Introducción **1**

Capítulo 1: Antecedentes y Estado del arte **4**

1.1 Introducción 4

1.2 Sonidos de las abejas 4

1.2.1 Sonido de colonias huérfanas 6

1.2.2 Sonidos en respuesta a la defensa de la colonia 6

1.2.3 Sonido en respuesta al periodo de pre-enjambrazón 6

1.2.4 Sonido en respuesta en enfermedades 7

1.2.5 Otros sonidos dentro de la colmena 7

1.3 Apicultura de precisión 7

1.3.1 Esquemas para el monitoreo de colonias de abejas 8

1.3.2 Medición de temperatura y humedad 8

1.3.3 Estudio del sonido en colonias de abejas 9

1.3.4 Otros parámetros de interés en la colonia 10

Capítulo 2: Marco Teórico **11**

2.1 Introducción 11

2.2 Transformada de Fourier de corta duración 11

2.3 Extracción de características y preprocesamiento 11

2.3.1 Coeficientes cepstrales en la frecuencia de Mel 11

2.3.2 Normalización Z 14
2.4 Regresión Logística 14
2.5 Función de verosimilitud para regresión logística 14
2.6 Regresión logística Lasso 15
2.7 Diagnóstico de sesgo y varianza 16
2.7.1 Sesgo y varianza 16
2.7.2 Curvas de aprendizaje 16
2.8 Desempeño de un algoritmo 18
2.8.1 Sensibilidad y Especificidad 18
2.8.2 Análisis ROC 19
2.9 Descomposición en Valores Singulares 20
Capítulo 3: Experimentación **22**
3.1 Introducción 22
3.2 Colonias de abejas 22
3.3 Condiciones en las colonias de abejas 23
3.4 Sistema de monitoreo 24
3.4.1 Instalación de los micrófonos 24
3.5 Metodología de análisis 25
3.6 Parámetros de monitoreo 27
Capítulo 4: Resultados y Discusión **29**
4.1 Análisis espectral de la señales 29
4.2 Extracción de características 29
4.3 Selección de características, regularización y clasificación del modelo multiclase 32
4.4 Evaluación de los modelos 32
4.5 Análisis de conglomerado 33
4.5.1 Análisis de conglomerados por días 33
4.6 Evaluación de los parámetros de monitoreo 36

Conclusiones **39**

Referencias 40

Bibliografía 45

Anexo 1: Publicaciones 46

Índice de figuras

1 Etapas en la identificación y clasificación de sonidos de abejas. 2

1.1 Clasificación de los sonidos emitidos por abejas. 5

2.1 Espectrograma del sonido producido por una colonia de abejas, cada columna de la imagen representa la amplitud del espectro en un mapa de colores, en [1, Fig. 2]. 12
2.2 Metodología para la extracción de características mediante MFCC. 12
2.3 Banco de filtros triangulares en la escala de Mel. 13
2.4 Problema de sobreajuste y ajuste pobre. 17
2.5 Curvas de aprendizaje para la evaluación del ajuste del modelo con base al número de características. 17
2.6 Curvas de aprendizaje para la evaluación del ajuste del modelo con base al número de muestras u observaciones. 18
2.7 Matriz de confusión. 19
2.8 Clasificadores discretos en el espacio ROC. 19
2.9 Curva en el espacio ROC. 20

3.1 Apiario en San José de la Era. 23
3.2 Configuración del sistema de monitoreo. 25
3.3 Instalación de los micrófonos en el interior de las tapas de las colmenas (a) para el primer experimento y en el marco de madera (b) para el segundo experimento. 25
3.4 Instalación de los micrófonos en la camara de la cria en el primer experimento, y en un marco entre la cámara de la cría y la alza melaria en el segundo experimento. 26
3.5 Etapas en la identificación, clasificación y análisis de sonidos de abejas. 27

4.1 Espectrogramas de las colonias mostrando bandas de mayor energía entre los 250 a 300 Hz exceptuando por la colonia 3 que tiene una reducida población y no tiene reina. 30
4.2 Espectrogramas posteriores a la remoción de la reina en las colmenas 2 y 4. En comparación con los espectrogramas anteriores las bandas de mayor energía parecen haberse reducido. 31
4.3 Descomposición en valores singulares, CR - con reina y SR - sin reina, toda las colonias con reina excepto por la colonia 3. 34
4.4 Descomposición en valores singulares, colonia 1 y 5 con reina, colonia 2 y 4 reina removida y colonia 3 inicialmente sin reina y baja población. 34
4.5 Análisis SVD de las colonias consigo mismas a través de distintos días, azul primer día, verde cuarto día, naranja sexto día y azul marino noveno día. . . . 35

4.6 Comparativa de las colonias en periodos de un 36 dias de diferencia. En la colmena 1 y 5 las muestras en azul y verde corresponden a la colonia con reina. En la colmena 2 y 4 las muestras en azul corresponden a la colonia con reina y en verde posterior a la remoción de la misma. Finalmente en la colmena 3 ambos grupos de muestras son de la colonia sin reina. 36
4.7 Comparativa del error de predicción contra número de muestras de conjuntos de datos con distintas duraciones de muestras. 37
4.8 Evaluación de los modelos según la longitud de las muestras y el número de muestras. 37
4.9 Espacio requerido por los conjuntos de datos de acuerdo a la longitud y número de muestras. 38

Índice de tablas

1.1 Sonidos de abejas. 5

3.1 Experimentos realizados en colonias de abejas, CR = con reina, SR = sin reina. 26
3.2 Ejemplos de parámetros de muestreo en la literatura. 28

4.1 Selección de características para cada clase del modelo multivariado. 32
4.2 Evaluación del rendimiento del modelo multiclase. 32
4.3 Espacio requerido por las muestras según su duración. 38

Introducción

Entre las 20,000 especies de abejas conocidas alrededor del mundo, las abejas (*Apis mellifera*) son las más comunes y útiles para el ser humano. Nativa de Europa, Asia y África, su valor va desde la producción de miel, cera, propóleos y jalea real, hasta la polinización eficiente de cultivos. Es de gran importancia para los ecosistemas naturales, de su actividad dependen una gran variedad de cultivos de frutos, vegetales, legumbres y semillas. Por lo que es conocida como el polinizador más valioso y económico alrededor del mundo [2].

Hasta hace algunos años, México se había posicionado como el sexto productor mundial de miel, 40,000 apicultores y sus familias dependían de dicha producción [3]. Sin embargo, la actividad apícola en el país se ha visto afectada. Una colonia de abejas está expuesta a un sin número de factores que dificultan su supervivencia y que lo hacen un superorganismo muy susceptible, entre los que se puede mencionar, depredadores, inclemencias del clima, a la reducción de alimentos beneficiosos para el desarrollo de la abeja, al aumento constante de la mancha poblacional, al uso de pesticidas en cultivos agrícolas, por mencionar algunos [4, 5]. Además de lo anterior, dentro de las colonias de abejas se pueden presentarse diversas condiciones como son: la falta de reina, reducido número de abejas, enfermedades, el fenómeno de enjambrazón. Que si no son detectados a tiempo pueden conducir a la muerte de la colonia.

El sonido de una colonia de abejas lleva consigo información muy relevante. Los apicultores más experimentados pueden conocer el estado del apiario escuchando los sonidos producidos, pueden determinar si la colonia está huérfana, si está enferma, estimar el tamaño de la colonia, si se encuentra en periodo de pre-enjambrazón, si no tiene alimento o si cosechó grandes cantidades durante el día [6]. Generalmente los apiarios se encuentran localizados en zonas alejadas de las ciudades y las visitas de inspecciones frecuentes resultan costosas, además, de que producen estrés en las colonias y que repercute de manera directa en una disminución de la productividad.

Investigadores alrededor del mundo han generado propuestas, desde enfoques muy diversos, encaminadas a resolver la problemática en la actividad apícola. De manera general las propuestas están basadas en el uso de tecnologías que permiten el monitoreo continuo de aspectos físicos de las colonias. Los parámetros de mayor interés para los investigadores son: temperatura, humedad, sonido y vibraciones, los cuales, son estudiados para desarrollar herramientas de apoyo en la producción precisa y manejo apícola. Estas mejoras ayudarán al apicultor tener un mejor control de los apiarios, minimizar los gastos de manutención, aumentar la productividad, reducir las inspecciones invasivas y finalmente llevar a cabo una mejor toma de decisiones para el correcto control de las colonias de abejas.

Este trabajo se suma a los esfuerzos en la mejora de la actividad apícola tradicional con la propuesta de una metodología basada en el análisis de las señales de sonido mediante técnicas de aprendizaje automático para la identificación, clasificación y reconocimiento de patrones en el sonido de colonias de abejas. El proceso de estudio de los sonidos de abejas está divido en varias etapas mostradas en la Figura 1. De manera específica se estudiaron dos

casos en colonias de abejas, cuando se encuentra huérfana y cuando su población es reducida. Los patrones se compararon con colonias de abejas sanas como referencia. Se estudiaron cinco colonias de abeja carniola en distintas condiciones biológicas, dos colonias fuertes y saludables, dos colonias más con reina, pero de menor población que las anteriores y una última colonia que perdió su reina de manera natural y como consecuencia su población se vio disminuida. Para simular colonias huérfanas se removió la reina a dos colonias. Los sonidos fueron adquiridos mediante un sistema de monitoreo basado en una Raspberry Pi 2 y micrófonos omnidireccionales. De las señales de audio se extrajeron características mediante Coeficientes Cepstrales en la Frecuencias de Mel (MFCC), así mismo de cada coeficiente de Mel se calcularon descriptores estadísticos: media, mediana, varianza, desviación estándar, curtosis y sesgo. Para la regularización y selección de características se empleó un modelo de regresión logística penalizada Lasso junto a una técnica de uno contra todos, creando un modelo de clasificación binaria para cada condición en la colonia. Finalmente, el modelo fue validado mediante el cálculo del área bajo la curva ROC. La metodología se puede extender para el caso de estudio de más condiciones en colonias, es decir, de su estado de salud, existencia de patologías o detección temprana de enjambrazón. La parte más fuerte del estudio se centró en diversos análisis de conglomerados mediante descomposición en valores singulares (SVD) y gráficas de dispersión para analizar de manera visual el comportamiento de las muestras y la forma en que se agrupan. Este análisis permitió observar que los patrones de las colonias pueden estar influenciados por las condiciones de las colonias, es decir, estado de salud, volumen poblacional, entre otros. Generando patrones ligeramente distintos incluso en colonias de características similares. Además, se demostró que para que un modelo de clasificación pueda identificar de manera correcta condiciones en colonias es necesaria incluir información de colonias con distintas características. Para complementar el trabajo de tesis se hizo un cálculo de los parámetros de monitoreo, es decir, número y longitud de muestras, para garantizar que la base de datos logre una alta tasa de clasificación y al mismo tiempo reducir los requerimientos de espacio de almacenamiento y reducir los tiempos de procesamiento de los datos. El trabajo en conjunto permite establecer pautas para el monitoreo de colonias de abejas y el estudio de los patrones en el sonido.

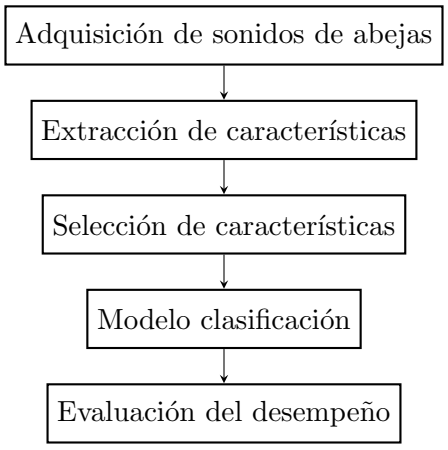


Figura 1: Etapas en la identificación y clasificación de sonidos de abejas.

Hipótesis

Es posible identificar patrones en el zumbido producido por colonias de abejas bajo distintas condiciones y aplicar algoritmos de clasificación para determinar el estado de salud de colonias de abejas

Objetivos

Objetivo general

Identificar las características de los patrones de abejas sanas, sin reina y de reducida población mediante el análisis de señales acústicas.

Objetivos específicos

1. Inferir los tipos de sonido que se producen en las colonias de abejas.
2. Crear un sistema de monitoreo para grabar sonidos de abejas.
3. Caracterizar los patrones de sonido de abejas bajo distintos estados de salud de interés.
4. Proponer una metodología para el reconocimiento de los patrones en el sonido de las abejas.

El trabajo de tesis está organizado como sigue: en el Capítulo 1, se describen distintos sonidos que se pueden presentar en colonias de abejas y el estado del arte del estudio apícola específicamente de la apicultura de precisión, en el Capítulo 2 se detalla la metodología para la extracción de características, el modelo de clasificación y validación, y se hace una breve descripción del sistema de monitoreo, en el Capítulo 3 se detallan las características y condiciones de las colonias de abejas y de los experimentos realizados, además, de manera breve se describe el sistema de monitoreo, finalmente en el Capítulo 4 se presentan los resultados y discusión.

Capítulo 1

Antecedentes y Estado del arte

1.1 Introducción

La producción de sonidos en abejas tiene características muy peculiares, es decir, desde la forma en que las abejas producen el sonido, hasta la forma en que el sonido de una colonia de abejas sana se ve afectado por diversos factores. Para la mayoría el sonido de abejas podría ser simplemente de naturaleza aleatoria resultado de la actividad y sin ningún patrón definido, pero en realidad el sonido puede contener una gran cantidad de información. El sonido de una colonia puede ser indicativo de su estado de salud, de un estado de alerta sobre posibles peligros como depredadores, o resultado de la intensa actividad de las trabajadoras. En torno al tema, se han realizado diversas investigaciones con el objetivo de encontrar patrones característicos de cada condición, para finalmente desarrollar sistemas capaces de identificarlos. Además del sonido, se han propuesto el monitoreo de distintos parámetros, como la temperatura que está directamente relacionada con la enjambrazón, y que podrían ayudar a facilitar la identificación de los patrones. Adicionalmente, se han desarrollado numerosas propuestas de sistemas basadas en distintas tecnologías que permiten el monitoreo remoto de las colonias, todas con ventajas e inconvenientes.

1.2 Sonidos de las abejas

Una variedad de diferentes sonidos han sido estudiados en las abejas. La mayoría de ellos son caracterizados por una baja frecuencia fundamental (300 a 600 Hz) y sus armónicos, tabla 1.1. En comparación con otros insectos como cigarras o grillos, las abejas no estas dotadas con estructuras diseñadas específicamente para la producción de señales acústicas. En cambio, los sonidos que producen las abejas son generados por *oscilaciones rítmicas torácicas*, es decir, generan sonido a partir de movimientos del cuerpo. Las alas de la abeja no son esenciales para la producción de sonidos, sin embargo, son importantes para la transmisión de los sonidos a través de aire.

La frecuencia fundamental del sonido puede ser modulado abriendo o cerrando las alas. Las frecuencias más altas de las vibraciones torácicas aparecen cuando las alas están completamente cerradas. Estas señales son transmitidas directamente al panal a través de un contacto directo con el tórax. Las patas crean un enlace entre el tórax y el panal. Para percibir las vibraciones trasmitidas por el panal se atribuye que se hace a través de las patas. Por otro lado, para la percepción de las vibraciones transmitidas a través del aire las abejas utilizan las antenas. El órgano de Johnston, ubicado en las antenas de la abeja, funciona como detector

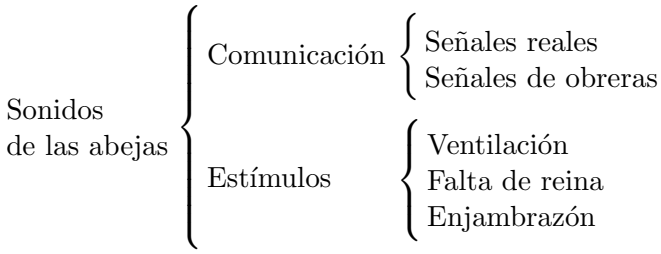


Figura 1.1: Clasificación de los sonidos emitidos por abejas.

de velocidad de partículas, la deflexión de las antenas se convierte en impulsos eléctricos que posteriormente son transmitidos al cerebro.

Los sonidos emitidos por las abejas pueden ser clasificados en dos grupos, Figura 1.1. En el primero, los utilizados para la comunicación entre abejas, como el *tooting* o el *quaking* que son emitidos por abejas reinas en distintos contextos, con estructuras bien definidas de secuencias de pulsos de diferentes duraciones y frecuencias fundamentales, tabla 1.1. Y en el segundo grupo, los sonidos que no son productos de la comunicación, sino que son emitidos por la colmena como un organismo, es decir por el conjunto de individuos de la colonia, y dependerá de distintos factores entre los que se encuentran los climáticos, biológicos, entre otros, este grupo de sonidos son los de mayor relevancia para el presente trabajo de investigación.

Tabla 1.1: Sonidos de abejas.

Señal	Especie	Frecuencia principal (Hz)	Patrón de señal	Emisor	Posible significado
Señales de la reina	**Apini**				
Tooting	Apis melífera	350 a 500	Secuencia de pulsos	Reina virgen emergida	Evita la salida de nuevas reinas. Incita que las reinas no nacidas respondan con Quacking.
Quaking	Apis melífera	300 a 350	Secuencia de pulsos	Reina virgen no emergida	Informa a la reina y trabajadoras de la presencia y de la viabilidad las reinas confinadas.
Señales de defensa	**Bombini**				
Siseo (Hissing)	Bombus agrorum	-	Pulso único	Colonia	Señal aposematica dirigida a los posibles predadores.
	Apini				
	Apis florea	Banda ancha: 90 % de espectro de energía: 500 a 5000	Pulso único	Colonia	Señal aposematica de advertencia.
Pitidos de obreras	**Apini**				
	Apis florea	300 a 500	Secuencia de pulsos	Pecoreadoras	Activar el siseo de la colonia.
	Apis melífera	300 a 550	Pulso único	Danza de trémulo	Detener la danza oscilante, facilitar el reclutamiento de recibidores de néctar.
Señales de Reclutamiento	**Apini**				
Danza de pecoreadoras	Apis melífera	200 a 350	Secuencia de pulsos	Pecoreadoras	Incrementar la atención de los receptores de alimentos, información acerca de la rentabilidad y localización de los alimentos.

1.2.1 Sonido de colonias huérfanas

La reina tiene un rol vital para una colonia de abejas, es la única hembra con la capacidad de reproducirse, por lo tanto, es la encargada de poner huevos de manera continua, aproximadamente 2000 huevos por día. La abeja reina mantiene el orden de la colonia mediante la liberación de feromonas, el cual es un mecanismo químico de comunicación entre los miembros de la colonia. Dentro de un apiario es común encontrar colmenas sin reina, aproximadamente 25 % de las colonias [7]. La abeja reina puede morir por enfermedad, de manera accidental, por ataque de predadores o por error del apicultor. Una colonia sin reina no sobrevivirá a un periodo prolongado. Su ausencia afecta el comportamiento completo de la colonia, cambia de un estado de actividad organizada a uno desorganizado, se vuelve estresada y vulnerable a ataques, plagas o enfermedades. La desorganización en la colonia y estrés de la abeja generan un cambio en el sonido de una colonia. La forma de evitar la muerte de la colonia para un apicultor es la temprana identificación de la ausencia de la abeja reina.

1.2.2 Sonidos en respuesta a la defensa de la colonia

La defensa del nido, de las crías y de los recursos almacenados, es vital para las colonias de abejas sociales y requiere la acción simultánea de muchos individuos. Varios mecanismos, de los cuales la mayoría son altamente sofisticados han evolucionado para proteger el nido contra intrusos. En respuesta a los disturbios como una sacudida mecánica o de un potencial predador, las colonias de varias especies de abejas emiten sonidos llamados *siseo*, que son generados por el movimiento de las alas [8]. El comportamiento del siseo de las abejas ha sido frecuentemente llamado trémulo, que describe la impresión óptica del movimiento de alas coordinado y simultanea de muchos individuos. Los sonidos de siseo en la naturaleza son de banda ancha, principalmente de altas frecuencias. Se ha sugerido que han evolucionado para oídos vertebrados y para representar señales de advertencia emitidas para disuadir o confundir posibles predadores. En las especias *A. cerana* y *A. florea*, la supresión general temporal de la actividad de la colonia frecuentemente seguida del siseo, ha sido propuesto como una conducta de la colonia para hacerla menos sospechosa a los predadores. Para activar y coordinar el siseo simultáneo de muchos individuos un activador o gatillo es requerido. Ha sido demostrado que las abejas pecoreadoras de *A. florea*, después de su retorno a la colonia emiten un pitido de advertencia (duración: 0.2 a 2.0 s; frecuencia fundamental: 300 a 480 Hz), después de haber percibido la presencia de un potencial peligro para la colonia. Inmediatamente después del siseo, los individuos cercanos a la abeja que comenzó el siseo comienzan a silbar. La acción del siseo se esparce rápidamente a los vecinos y finalmente involucra a la mayoría de la colonia.

1.2.3 Sonido en respuesta al periodo de pre-enjambrazón

La enjambrazón es la forma natural de propagación de las colonias de abejas. El fenómeno se produce principalmente en el periodo de más abundancia de flores y polen, cuando el desarrollo de una colonia está en su más alto nivel [9]. Si el espacio dentro de la colonia no es suficiente para albergar el creciente número de abejas, una reina fertilizada o no fertilizada junto con un grupo de trabajadoras (aproximadamente la mitad del enjambre) abandonarán la colmena para establecer una nueva colonia en un lugar alejado del actual. La preparación para la enjambrazón puede tomar de varios días a varias semanas. Es importante para los apicultores poder detectar este periodo a tiempo, ya que puede afectar en la reducción de

la población de abejas, descenso en la producción de crías y la pérdida de la colonia. Es sabido que ocurren cambios en el sonido cuando la colonia se encuentra en preparación para la enjambrazón, pero existen la inconsistencia de los resultados. Por ejemplo, de acuerdo con Eskov [10] citado por [11], la preparación de las abejas para la enjambrazón está marcado por un incremento en las componentes del sonido en un rango de 210 – 240 Hz. Un rango más amplio de frecuencias asociado con la enjambrazón (210–240, 300–330, 390–420, 420–450 Hz), fue reportado por Rybochkin [12] citado por [11]. Otros investigadores reportaron un incremento inicial de la energía acústica relativa de alrededor de 110 Hz, el desplazamiento del pico de 500 – 600 Hz cuando el enjambre esta por abandonar [13].

1.2.4 Sonido en respuesta en enfermedades

La varroasis es una enfermedad causada por el ácaro *Varroa Destructor*, considerada la peor plaga de abejas a nivel mundial, puede provocar la muerte de la colonia si no es tratado oportunamente. El ácaro varroa es de color rojizo del tamaño de la cabeza de un alfiler, se adhiere a las placas ventrales de la abeja succionando la sangre de larvas y abejas adultas. El ácaro puede diseminar enfermedades virales y bacterias. Cuando las abejas portan parásitos se sacuden para tratar de desprenderse de los ácaros, lo que produce sonidos audibles. Sin embargo, solo se presenta cuando el nivel de infestación es muy elevado.

1.2.5 Otros sonidos dentro de la colmena

Existen diversos sonidos dentro de la colmena como el producido por la ventilación, que es descrito como proceso de intercambio de aire en el interior de la colmena por aire fuera de la colmena [14]. Para ventilar la colmena las abejas se posicionan en la piquera o entrada de la colonia y agitan sus alas para introducir aire. La ventilación de la colmena puede activarse debido a altos niveles de dióxido de carbono o debido a la alta temperatura ambiental. Durante la fase activa de la colonia (actividad en primavera), sonidos de alta intensidad son producto de la ventilación de la colmena [11]. Cuando la temperatura ambiente se eleva al sobrecalentamiento del nido, las componentes del ruido se elevan de 75 – 160 Hz y la energía máxima cambia al rango de las altas frecuencias [11]. Existen otros sonidos producidos por acciones mecánicas como: limpieza de las celdas por las trabajadoras, el roer de las jóvenes adultas que emergen, así como de la fricción de las adultas contra otras [11].

1.3 Apicultura de precisión

Para atender la problemática en torno al manejo de apiarios se ha venido desarrollando una rama relativamente nueva de la agricultura de precisión, llamada *apicultura de precisión* que hasta ahora se ha definido como: el manejo estratégico de un apiario basado en el monitoreo de colonias de abejas para minimizar el consumo de recursos y maximizar la producción de abejas. Aunque hay suficiente medios técnicos e industriales para la ejecución práctica de la apicultura de precisión, el proceso es lento debido a las diferentes fases de desarrollo: recolección de datos, análisis de datos y aplicación [15, 16, 17].

Recolección de datos. En la etapa de recolección tiene como objetivo el desarrollo de herramientas para el monitoreo continuo de apiarios, los parámetros de mayor importancia

dentro de una colonia son: temperatura, humedad, contenido de gas, sonido, vibraciones y peso. Todos ellos brindan información importante acerca del estado de la colonia.

Análisis de datos. Solo después de que es entendido el significado de los datos, es posible automatizar análisis y desarrollar algoritmos para la interpretación de datos, los cuales pueden ser usados para informar al apicultor el estado de la colonia.

Fase de aplicación. La tercera fase de la apicultura de precisión es la aplicación. Y que es la meta final de la apicultura de precisión, el desarrollo de sistemas de soporte a la decisión (Design Suport Systems DSS) que tengan la posibilidad recabar información, analizar los datos y tomar una decisión automáticamente [18].

1.3.1 Esquemas para el monitoreo de colonias de abejas

Según Kviesis [19], existen diversos esquemas de monitoreo que pueden ser aplicados para aprovechar esta valiosa herramienta, de acuerdo con las necesidades del investigador, todas con ventajas e inconvenientes, por ejemplo, algunos esquemas emplean computadoras en el lugar del apiario, otros esquemas utilizan dispositivos más compactos como son microcontroladores ATmega [20, 21], microcontroladores como el Freescale [22, 23], FPGAs [24], microcomputadoras como la Rasberry Pi [25, 26] o la Beagleboard [27].

Con el auge de nuevas tecnologías como el WSN (Wireless Sensor Network), que surge a partir del Internet de las Cosas (IoT), involucra tecnologías de comunicación inalámbrica, monitoreo embebido y computo. Un ejemplo de la aplicación de este tipo de tecnologías al monitoreo de colonias de abejas es el estudio hecho por Murphy [20], donde desarrollan un sistema de bajo consumo energético alimentado por paneles solares y donde también incorporan comunicación 3G para informar al apicultor. Otro ejemplo de estos sistemas es el implementado por Grammarini [22] donde proponen un sistema basado en microcontrolador Freescale que incorpora micrófonos, acelerómetros, celdas de carga y equipada con GSM para comunicaciones. Como ejemplo de estos esquemas Zacepins [28], propone un esquema para el monitoreo de temperatura en colonias de abejas, el mismo sistema se utiliza para detectar el periodo de enjambrazón de una colonia [29].

1.3.2 Medición de temperatura y humedad

Las abejas tienen la capacidad de regular su temperatura, en invierno se agrupan en un cuerpo compacto y agitan su cuerpo para generar calor y evitar que las crías mueran. En primavera, las abejas se posicionan a la entrada de colmena y agitan sus alas para hacer circular aire a través de la colonia y reducir la temperatura [30]. El monitoreo de la temperatura y humedad permite relacionar eventos con cambios en el sonido. Basado en las mediciones de temperatura es posible determinar si la colonia ha incrementado su consumo de alimentos, si ha iniciado la producción de crías, es posible reconocer la fase de previa a la enjambrazón o la muerte de la colonia [31]. La medición de la temperatura y la humedad son las formas más sencillas de obtener información con relativamente fácil implementación y de forma económica [31]. La medición de la temperatura puede realizarse de distintas formas. Meitalovs [32] desarrolló un sistema de monitoreo para determinar la etapa previa a la enjambrazón mediante sensores de temperatura y humedad posicionados dentro la colmena, se encontró que las abejas mantienen un control de la temperatura y la humedad con cambios relativamente lentos, por lo que es posible determinar los cambios cuando ocurre la fase previa a la enjambrazón,

sin embargo, la detección de este estado por un solo parámetro y que además depende de diversos factores biológicos y climáticos es cuestionable.

Otros experimentos se han llevado a cabo utilizando luz infrarroja a través de la colmena, para poder realizarlo se reemplazaron los costados de una colmena con láminas de poliuretano translucido, lo anterior para estudiar el impacto de la temperatura externa con la termorregulación de las abejas [33]. La medición de la temperatura como en el caso anterior también se hace mediante radiación infrarroja de onda larga, en el caso de Shaw [34] la medición fue utilizada para determinar el tamaño de la población, es un método no invasivo que no requiere abrir la colmena y no pone en peligro la reina.

El enfoque de Reyes [23] donde se ha diseñado un módulo similar al bastidor de una colmena Langstroth, incorpora sensores de humedad, temperatura y movimiento, el módulo se inserta en la parte media de la colmena para monitorear las condiciones internas y según los investigadores no genera estrés en la colonia.

Se han estudiado los cambios en la temperatura en la parte superior de la colmena para correlacionar los datos de la energía liberada por la colonia y su estado anual, los estados de importancia son: 1) crianza en invierno, 2) crianza en primavera, 3) crianza en verano, 4) crianza en otoño y 5) periodo de no crianza en otoño, esto es de ayuda para organizar las actividades de los apicultores [35].

1.3.3 Estudio del sonido en colonias de abejas

Por otro lado, el monitoreo del sonido es donde se han llevado a cabo recientes investigaciones. Se han propuesto métodos para determinar el periodo de enjambrazón basado en el etiquetado de sonidos, cuando las abejas se encuentran en periodo de pre-enjambrazon el sonido cambia en amplitud y frecuencia, Ferrari [13] analizó 3 colmenas de abejas y fueron detectadas nueve actividades de pre-enjambrazón que resultaron en un aumento en la frecuencia, la detección se realizó mediante inspección visual de los espectrogramas de las señales, los cambios en la frecuencia van de 100-300 Hz (considerada como actividad normal) hasta los 500-600 Hz cuando se presentan los sonidos de pre-enjambrazón.

Otros sistemas emplean enfoques distintos, por ejemplo, Murphy [20] diseña un sistema que de la misma manera detecta el periodo de pre-enjambrazón, pero únicamente graba sonidos cuando existen frecuencias que caracterizan el evento, es decir, cuando las frecuencias del sonido superan un límite preestablecido, los audios son correlacionados con mediciones de humedad y temperatura, sin embargo, el sistema se enfoca en el diseño electrónico más que en el análisis de los datos. También se han desarrollado plataformas para monitorear el sonido en colmenas con el objetivo de caracterizar la actividad del enjambre [36]. Se ha analizado el zumbido de la abeja cerca de la entrada o piquera para evaluar la productividad de la colmena [37].

Existen también, dispositivos patentados que aclaman poder detectar el estado de salud de la colonia a través del monitoreo del sonido, el dispositivo puede detectar variaciones en el zumbido de las abejas que se producen al someter a las abejas en ambientes contaminados, estos sonidos son comparados con una base de datos [38]. Otra propuesta similar es un sistema para detectar el nivel de infestación por ácaro varroa de una colonia de abejas a través del análisis de señales acústicas [27]. La metodología propuesta funciona comparando la huella acústica de una colonia infestada con una huella de un estatus conocido, para llevar a cabo el reconocimiento usaron técnicas de Aprendizaje Automático como Maquinas de Soporte de

Vectores y Análisis Discriminante Lineal, sin embargo, los resultados son poco favorables, los sonidos de las colonias infectadas no provienen del mismo apiario y son pocos lo audios recabados, lo que resulta en un conjunto de datos no balanceado, con más muestras de colonias sanas que podría no ser suficiente para brindar experiencia a los modelos de reconocimiento.

La detección de estado de orfandad de colonias también ha sido de interés para los investigadores, Howard [39] analiza los sonidos de colonias con y sin reina para encontrar un patrón que identifique cada condición. En el estudio se discriminaron las señales realizando un análisis en frecuencia utilizando la transformada S. El experimento se llevó a cabo empleando cuatro colonias de abejas con reina, posteriormente en dos colonias se removió la reina y se capturaron los sonidos por siete días. Los audios fueron divididos en 24 cuartos de tonos y se extrajeron características, para la clasificación utilizaron un método llamado SOM (del inglés, Self Organis Maps). Sin embargo, los resultados obtenidos del clasificado no fueron satisfactorios. Otro estudio de orfandad en colonias de abejas fue llevado a cabo por Cejrowski [40] donde desarrollan un bastidor especial con micrófonos y sensores de temperatura y humedad que se introduce en la colmena, de los sonidos capturados se extrajeron características mediante Codificación Lineal Predictiva (LPC, por sus siglas en inglés), para la clasificación de los datos se utilizaron Máquinas de soporte de vectores (SVM, por sus siglas en inglés). El experimento fue remover la reina de colonias sanas, dejar las colonias sin reina y en un último paso introducir reinas nuevas. Se capturaron los sonidos con reina y sin reina, y posterior a la introducción de nuevas reinas. El experimentó arrojó resultados favorables durante la primera etapa, antes y después de la remoción de la reina, con un alto porcentaje de clasificación, sin embargo, después de introducir las nuevas reinas las huellas tenían distintas características.

1.3.4 Otros parámetros de interés en la colonia

El monitoreo por video permite observar la actividad en la entrada u otras actividades en la colmena, los cuales son indicadores importantes del estado de la colonia. Posibilita captar actividad inusual o inactividad o detectar enfermedades [41]. Junto al monitoreo de temperatura, el monitoreo por video permite detectar el estado de pre-enjambrazón, descartando la posibilidad de que el aumento de temperatura experimentado dentro de la colmena se deba a esta condición y no a factores climáticos [32]. Existen también sistemas de bajo costo basados en video que permiten el conteo de abejas, identificar su posición y medir su actividad en entrada y salida [26]. La detección del periodo de enjambrazón no solo se ha hecho a través del monitoreo de temperatura o de sonido, sino también a través de la medición de vibraciones que son producidas por la danza de las abejas y trasmitidas a través de la colmena [42]. Recientes investigaciones se han enfocado en caracterizar la concentración de los gases dentro de la colmena, se ha documentado que existen cambio en el contenido de dióxido de carbono (CO_2) y oxigeno (O_2) como respuesta a eventos como pueden ser: número de abejas dentro de la colmena, salud de la colonia y el clima en el exterior de la colmena [21].

Las básculas electrónicas pueden dar una buena valoración del flujo de néctar e indican cuando es necesario proveer a la colonia de alzas (caja superior de la colmena donde las abejas almacenan la miel) [43]. Esto se vuelve muy importante cuando las colonias se encuentran en lugares alejados y las visitas frecuentes son costosas. Además, el peso puede ser indicativo de actividad anormal, si éste permanece constante podría indicar muerte de la colonia [44].

Capítulo 2

Marco Teórico

2.1 Introducción

En el presente capítulo se describen las distintas herramientas utilizadas para el análisis y preprocesamiento de las señales de audio. Se describe la metodología de Coeficientes Cepstrales en las frecuencias de Mel para la extracción de características. Así mismo, el modelo de Regresión Logística Lasso para la selección de características y regularización del modelo, las formas de detectar problemas en el desempeño de un clasificador y la forma de evaluar el mismo. Y finalmente, se presenta la descomposición en valores para la creación de las gráficas de dispersión para el análisis de la agrupación de los datos.

2.2 Transformada de Fourier de corta duración

La trasformada de Fourier de tiempo corto (Short Timer Fourier Transform o STFT), permite analizar señales en ventanas de corta duración, reduciendo así los efectos de variaciones temporales y limita el efecto de no estacionariedad inherente en muchas señales del mundo real. La STFT está basada en el análisis de Fourier y la idea de esta herramienta es la de realizar un análisis simultaneo de una señal en tiempo y frecuencia. La STFT se define en tiempo discreto como:

$$X(n,w) = \sum_{m=-\infty}^{\infty} x(n+m)W(m)e^{-Jwm}, \tag{2.1}$$

donde $W(m)$ es la secuencia de la ventana, $x(n)$, es la secuencia de longitud finita, n indica la variable temporal discreta y w corresponde a la variable continua en frecuencia.

La STFT es la base de construcción de los espectrogramas, los cuales son representaciones visuales de la variación en el tiempo del espectro de frecuencias de una señal, Figura 2.1.

2.3 Extracción de características y preprocesamiento

2.3.1 Coeficientes cepstrales en la frecuencia de Mel

La extracción de características es el proceso de transformar un conjunto de datos, en una representación reducida de características. Esta transformación tiene ciertas ventajas, como:

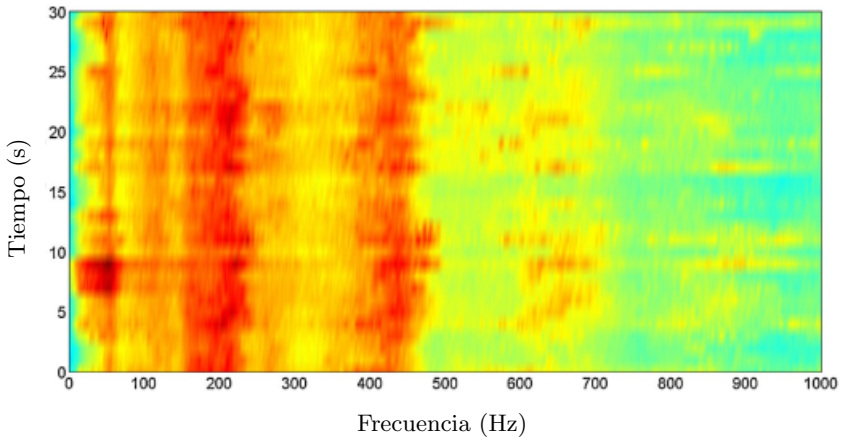


Figura 2.1: Espectrograma del sonido producido por una colonia de abejas, cada columna de la imagen representa la amplitud del espectro en un mapa de colores, en [1, Fig. 2].

mejora en la eficiencia el modelo, reduce el espacio de almacenamiento y mejora el procesamiento de estos. Dentro del reconocimiento de voz existen distintas técnicas para la extracción de características, entre las más populares se encuentran: Codificación Lineal Predictiva (LPC, por sus siglas en inglés), Coeficientes Cepstrales en las Frecuencias de Mel (MFCC, por sus siglas en inglés), Predicción Lineal Perceptual (PLP), entre otras. Sin embargo, la mayoría de los sistemas de reconocimiento automáticos de voz en la actualidad están basados en algún tipo de MFCC, que han probado ser eficientes y robustos bajo diferentes condiciones [45]. Los MFCC no solo han sido usados en reconocimiento de voz, sino además han sido empleados en la clasificación de música basado en el género musical [46], y en la clasificación de sonido ambiental [47] . Los MFCC se basan en la forma en que el sistema auditivo humano percibe las señales de voz, poniendo más énfasis en las frecuencias bajas. La metodología clásica para el cálculo de los MFCC se muestra en la Figura 2.2.

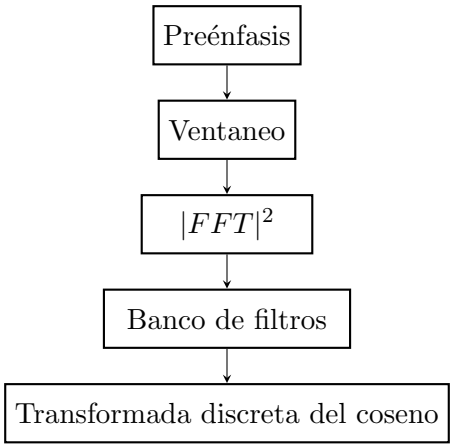


Figura 2.2: Metodología para la extracción de características mediante MFCC.

- **Preénfasis**. De acuerdo con el diagrama mostrado en la Figura 2.2 el primer paso es realizar un preénfasis, en esta etapa se busca aumentar la magnitud de las formantes de

alta frecuencia que usualmente son más débiles que las de baja frecuencia. Se utiliza un filtro digital descrito por la siguiente ecuación:

$$\hat{s}(n) = s(n) - ks(n-1), \tag{2.2}$$

donde k controla el grado de preénfasis, un valor clásico utilizado es 0.94.

- **Ventaneo**. En este bloque se lleva a cabo una segmentación de la señal en tramas de corta duración, de alrededor de 20ms, con traslape de entre 30 % y 75 % de duración de la trama, el traslape se utiliza para evitar los efectos de truncamiento. Durante este intervalo se considera que la señal es cuasi estacionaria. El análisis permite reducir los efectos de las variaciones temporales y limita el efecto de no estacionariedad de la señal. Una función ventana es aplicada para suavizar la señal, entre las más comunes se encuentra la ventana Hanning y la Hamming.

- **Transformada rápida de Fourier**. Después del ventaneo se aplica la transformada rápida de Fourier (FFT, por sus siglas en inglés) en cada ventana para extraer los componentes en frecuencia de la señal en el domino del tiempo.

- **Banco de Filtros**. A la señal se le aplica un banco de filtros triangulares linealmente espaciados con la escala de Mel, Figura 2.3. La escala de Mel es aproximadamente lineal hasta frecuencias de 1 kHz y después se vuelve logarítmica para altas frecuencias, para cada tono con una frecuencia medida en Hz, uno tono en la escala de Mel es calculado mediante:

$$f_{mel} = 2595 log_{10}\left(1 + \frac{f}{700}\right). \tag{2.3}$$

Los filtros triangulares tratan de emular la percepción de las señales de voz en el oído humano, que no sigue una escala lineal, es decir, que es menos sensible a altas frecuencias.

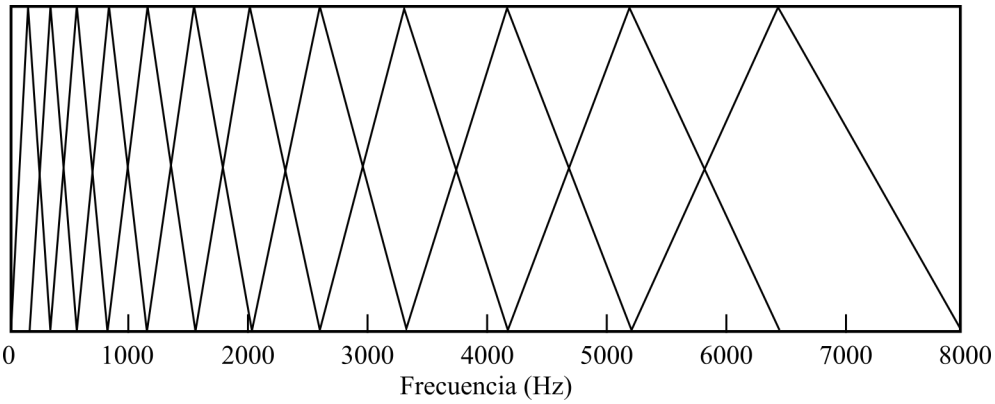


Figura 2.3: Banco de filtros triangulares en la escala de Mel.

- **Transformada Discreta del Coseno**. El último paso es el cálculo de la transformada discreta del coseno (DCT, por sus siglas en inglés) al logaritmo de la energía correspondiente a cada uno de los filtros. La DCT transforma del dominio de la frecuencia en un dominio de la quefrecuencias convirtiéndolos en coeficientes cepstrales (MFCC).

2.3.2 Normalización Z

Cuando los valores de los datos en crudos tienen rangos de magnitudes muy distintas, es imprescindible realizar un proceso de estandarización de los datos antes de realizar cualquier análisis. Una de las técnicas más empleadas para conseguirlo es mediante la medida z (z-score):

$$z = \frac{x - \mu}{\sigma}, \tag{2.4}$$

donde x es el dato a normalizar, μ es la media, σ es la desviación estándar. Este proceso es esencial cuando se aplica un descenso de gradiente, como en el caso de la regresión logística, ya que el modelo convergerá mas rápido en comparación con los datos sin normalizar.

2.4 Regresión Logística

La regresión logística es uno de los modelos de clasificación más populares y comunes en aprendizaje automático, surge de la necesidad de modelar la probabilidad condicional de una variable de salida $Y = 1$ como una función de (predictores) x. La probabilidad condicional se representa como $P(Y = 1|X = x) = p(x)$, y es llamada función hipótesis. Un modelo de regresión logística tiene un intercepto, además de parámetros de pendiente para cada término del modelo. Dado que se requiere que la probabilidad del evento este entre 0 y 1, no se puede garantizar que un modelo de intercepto y pendiente restringirá los valores dentro de este rango. Si $p(x)$ es una función lineal que representa la probabilidad de un evento, por lo tanto la probabilidad del evento es $p(x)/(1 - p(x))$. La regresión logística modela el logaritmo de la probabilidad de un evento como una función lineal:

$$log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n, \tag{2.5}$$

donde, n es el número de predictores, β son los pesos asociados a x_i que son las variables independientes que describen el evento a ser clasificado. El lado derecho de la función es la ecuación conocida de manera habitual como predictor lineal. Resolviendo para $p(x)$, se obtiene nuevamente la función de probabilidad de un evento:

$$p(x) = \frac{1}{1 + exp[-(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)]}. \tag{2.6}$$

Esta función no lineal está basada en una función sigmoidea. El modelo produce fronteras de clase lineal, a menos que los predictores usados en el modelo sean versiones no lineales de los datos.

2.5 Función de verosimilitud para regresión logística

La forma de calcular los parámetros β que mejor se ajusten a los datos es por medio de la función de máxima verosimilitud, para cada vector de características, x_i, hay un vector de clases o salidas, y_i. La probabilidad de la clase es, ya sea p, si $y_i = 1$, o $1 - p$ si $y_i = 0$. La función de verosimilitud es:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_j))^{(1-y_i)}, \tag{2.7}$$

el logaritmo de la función de verosimilitud transforma los productos en sumas.

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{m} y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)), \\ &= \sum_{i=1}^{m} \log(1 - p(x)) + \sum_{i=1}^{m} y_i \log \frac{p(x)}{1 - p(x)}, \\ &= \sum_{i=1}^{m} \log(1 - p(x)) + \sum_{i=1}^{m} y_i p(x), \\ &= \sum_{i=1}^{m} (y_i p(x) - \log(1 + e^{(p(x))}). \end{aligned} \tag{2.8}$$

Para encontrar las estimaciones de máxima verosimilitud se obtiene la derivada de la verosimilitud con respecto a los parámetros, se establecen las derivadas a cero y se resuelve:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{m} x_i (y_i - p(x)) = 0. \tag{2.9}$$

Sin embargo, no será posible establecer las derivadas en cero y resolverlas de manera exacta. Pero es posible utilizar métodos numéricos para obtener una solución aproximada.

2.6 Regresión logística Lasso

La regresión logística Lasso (least absolute shrinkage and selection operator, por sus siglas en inglés) agrega un término de penalización para estimar los coeficientes de regresión a la función de máxima verosimilitud. La versión penalizada del logaritmo de la función de verosimilitud a ser minimizada toma la siguiente forma:

$$l(\beta) = \sum_{i=1}^{m} (y_i p(x) - \log(1 + e^{(p(x))}) - \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.10}$$

sujeto a la restricción:

$$\sum_{j=1}^{p} |\beta_j| \leq t, \tag{2.11}$$

donde λ es un parámetro de complejidad que controla la cantidad de contracción: cuanto mayor sea el valor, mayor será la cantidad de contracción, β son los coeficientes de regresión, y t es una constante positiva que determina la cantidad de regularización.

Una propiedad interesante de Lasso es que el término de penalización funciona de selección de características y de regularización, cuando λ es suficientemente grande los coeficientes estimados de la regresión son escasos lo que significa que muchos de los componentes son

exactamente igual a cero. El modelo final involucra un subconjunto de los parámetros que son más predictivos.

La selección de características relevantes y eliminación de irrelevantes es una tarea importante dentro de Aprendizaje Automático (Machine Learning). En una base de datos no toda la información es necesaria para alcanzar un buen rendimiento. Es común encontrar bases de datos con cientos de características y que solo una pequeña fracción sean relevantes. Existen beneficios al realizar el proceso de selección de características: facilita la visualización de datos, reduce los requerimientos de almacenamiento y los tiempos de entrenamiento, y mejora el desempeño del clasificador.

2.7 Diagnóstico de sesgo y varianza

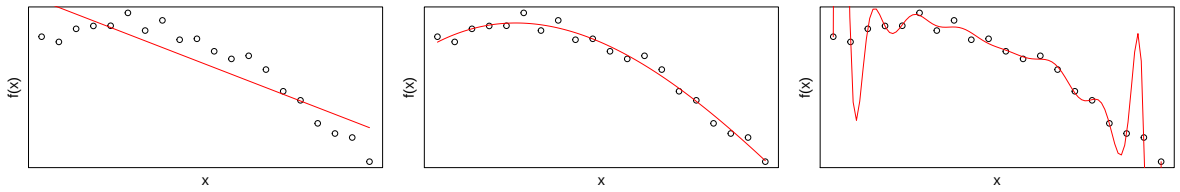
2.7.1 Sesgo y varianza

La causa de bajo rendimiento en aprendizaje automático se debe a problemas de alta varianza también conocido como sobreajuste (**overfitting** término en inglés), o debido a alto sesgo o ajuste pobre (**underfitting** término en inglés). La *varianza* se refiere a que tan sensible es un modelo a los cambios en los datos que lo construyen. La alta varianza causa que el modelo realice predicciones o clasificación de manera perfecta en los datos de entrenamiento. Sucede cuando un modelo aprende el detalle y ruido de los datos de entrenamiento, de manera que afecta en forma negativa el rendimiento del modelo cuando predice datos con los que no fue entrenado. El ruido y las fluctuaciones aleatorias en los datos de entrenamiento es recogido y aprendido como conceptos por el modelo. El problema surge cuando esos conceptos no aplican a nuevos datos y afectan de manera negativa la habilidad de un modelo para generalizar. En aprendizaje automático, el *sesgo* de un modelo se refiere al error introducido al tratar un modelo complicado de la vida real con una aproximación. Un modelo con alto sesgo no podrá modelar los datos de entrenamiento y, por lo tanto, no podrá generalizar nuevos datos. Evidentemente un modelo con estas características tendrá un rendimiento pobre. En la Figura 2.4 se ejemplifica de manera gráfica los problemas antes descritos. En las gráficas se muestran datos en dos dimensiones que podrían representar tamaño contra precio de una vivienda. Cuando se usa una hipótesis con un polinomio de bajo grado, sucede el caso mostrado en la Figura 2.4(a), el modelo queda con un ajuste pobre, es incapaz de generalizar[1]los datos con los que fue entrenado y, por lo tanto, será incapaz de generalizar nuevos datos o realizar predicciones correctas. En el ejemplo de la Figura 2.4 sucede el caso contrario, el modelo está tan ajustado a los datos que realizará predicciones de manera perfecta en los datos de entrenamiento, pero fallará al predecir o clasificar nuevos datos. De manera ideal, se busca que un modelo tenga un buen ajuste, es decir, en un punto entre el sobreajuste y un ajuste pobre, como el mostrado en la Figura 2.4(b). Esto permite a un modelo la capacidad de generalizar, es decir, realizar perdiciones en datos que no ha visto.

2.7.2 Curvas de aprendizaje

La forma de detectar problemas de sesgo alto o varianza alta es mediante curvas de aprendizaje, Figura 2.5 y 2.6. Las curvas de aprendizaje son gráficas de error de entrenamiento,

[1]Generalizar se refiere a que tan bien la hipótesis se aplica a nuevos datos.



(a) Ajuste pobre (alto sesgo). (b) Buen ajuste. (c) Sobreajuste (alta varianza).

Figura 2.4: Problema de sobreajuste y ajuste pobre.

ecuación 2.12 y error de prueba o validación cruzada, ecuación 2.13 contra el número de datos de entrenamiento o el número de características del modelo. Para la evaluación del desempeño de un algoritmo una forma común de llevarlo a cabo es dividir el conjunto de datos en dos, uno para entrenamiento del modelo que generalmente corresponde al 70 % y otro conjunto para prueba ciega que corresponde al 30 %. Las funciones de costo para error de entrenamiento y error de prueba son:

$$J_{ent}(\beta) = \frac{1}{2m_{ent}} \sum_{i=1}^{m_{ent}} (h_\beta(x_{ent}^{(i)}) - y_{ent}^{(i)^2}), \tag{2.12}$$

$$J_{vc}(\beta) = \frac{1}{2m_{vc}} \sum_{i=1}^{m_{vc}} (h_\beta(x_{vc}^{(i)}) - y_{vc}^{(i)})^2, \tag{2.13}$$

en donde m es el número de instancias u observaciones, y son las clases o salidas del modelo, h_β es la hipótesis y x son las características del modelo o datos de entrada.

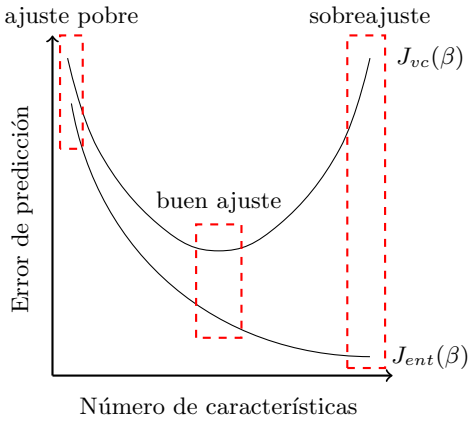


Figura 2.5: Curvas de aprendizaje para la evaluación del ajuste del modelo con base al número de características.

Una gráfica de error de predicción contra número de características produce curvas como las mostradas en la Figura 2.5. Cuando el número de características introducidas al modelo es bajo, el error de predicción tanto en el conjunto de datos de entrenamiento con el conjunto de datos de prueba será alto, es decir tendrá un ajuste pobre. Al introducir más características a modelo, tendrá un mejor desempeño. Sin embargo, el seguir introduciendo puede suceder que el modelo tengo un sobre ajuste y que haga predicciones de manera perfecta en los datos

de entrenamiento, pero que a su vez tendrá un bajo rendimiento al tratar de predecir nuevos datos. En todo caso se busca que el ajuste sea en un punto intermedio entre un ajuste pobre y un sobreajuste. Gráficas similares se obtiene al variar el número de muestras u observaciones, Figura 2.6. Un modelo entrenado con un número insuficiente de muestras genera alto error de predicción tanto en el conjunto de entrenamiento como en el conjunto de prueba. Al agregar más muestras al modelo pueden suceder dos casos, ya sea que el modelo no obtenga experiencia de las muestras, Figura 2.6(a), o el segundo caso que al introducir más muestras al modelo éste se sobreajusta, Figura 2.6(b).

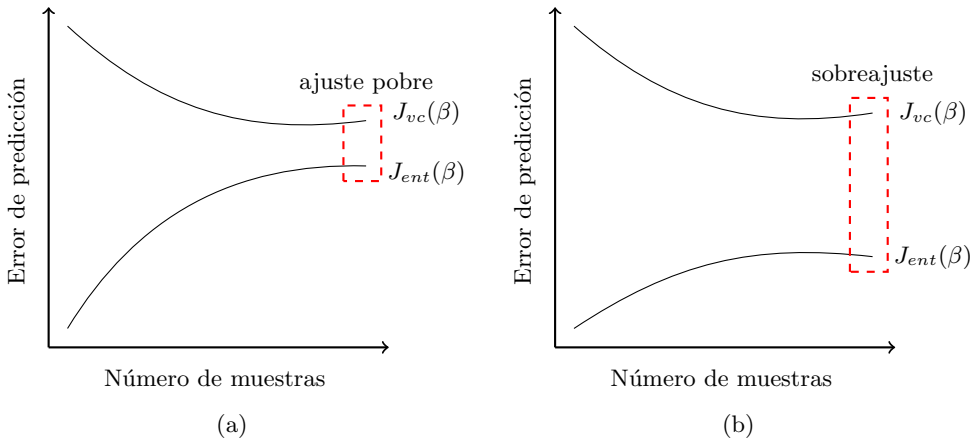


Figura 2.6: Curvas de aprendizaje para la evaluación del ajuste del modelo con base al número de muestras u observaciones.

2.8 Desempeño de un algoritmo

2.8.1 Sensibilidad y Especificidad

La decisión tomada por un clasificador puede ser representada en una estructura conocida como matriz de confusión. La matriz de confusión tiene cuatro categorías: Verdaderos Positivos (VP) que se refiere a muestras correctamente clasificadas como positivas, Verdaderos Negativos (VN) que corresponde a muestras negativas correctamente clasificadas como negativas, Falsos Positivos (FP) que son muestras negativas incorrectamente clasificadas como positivas y Falsos Negativos (FN) que son muestras positivas incorrectamente clasificadas como negativas, la matriz de confusión se muestra en la Figura 2.7.

Dada la matriz de confusión pueden definirse distintas métricas, como la Sensibilidad que mide la proporción de positivos que son correctamente identificados como tales y la Especificidad que de manera opuesta mide la porción de negativos correctamente clasificados como negativos, ecuaciones 2.14 y 2.15.

$$Sensibilidad = \frac{VP}{VP + FN}, \tag{2.14}$$

$$Especificidad = \frac{VN}{FP + VN}. \tag{2.15}$$

		Clase Actual: P	Clase Actual: N
Predicción	P	Verdaderos Positivos	Falsos Negativos
	N	Falsos Positivos	Verdaderos Negativos

Figura 2.7: Matriz de confusión.

2.8.2 Análisis ROC

Las curvas ROC (Receiver Operative Characteristics) son comúnmente usadas para representar problemas de decisión binaria en Aprendizaje Automático más conocido como Machine Learning. Una curva ROC está definida como una gráfica de Sensibilidad en el eje de las ordenadas contra 1-Especificidad en el eje de las abscisas. Un clasificador discreto produce un par (*sensitividad*, *especificidad*) correspondiente a un punto en el espacio ROC. Los clasificadores mostrados en la Figura 2.8 son discretos. Un clasificador que cae en el punto $B(0,0)$ nunca emite un falso positivos pero tampoco verdaderos positivos. El punto $A(0,1)$ representa un clasificador perfecto, clasifica correctamente tanto los positivos como los negativos sin emitir falsos positivos o negativos, por lo tanto un buen clasificador caerá cerca de esquina superior izquierda. Un clasificador que emite un punto por debajo de la línea punteada, por ejemplo, el punto C, se dice que es peor que una elección al azar. Algunos clasificadores como la Re-

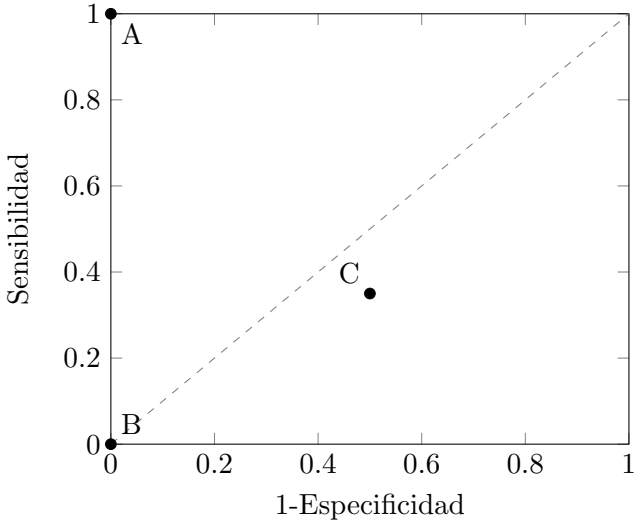


Figura 2.8: Clasificadores discretos en el espacio ROC.

gresión Logística o una Red Neuronal producen la probabilidad de una instancia, es decir, un valor numérico que representa el grado de una instancia de pertenecer a una clase. Esos valores pueden ser estrictamente probabilidades y un valor alto indica una alta probabilidad.

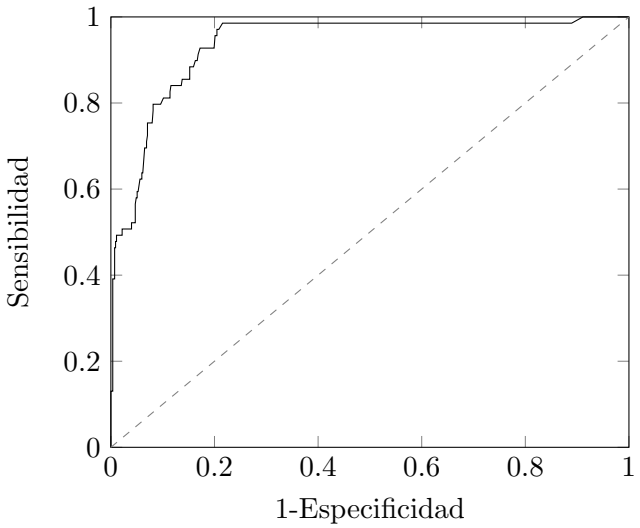


Figura 2.9: Curva en el espacio ROC.

Tal probabilidad puede usarse con un límite o treshold para producir un clasificador binario, si el clasificador produce una salida por encima de cierto límite produce un 1 o un 0 si está por debajo. Cada límite produce un valor en el espacio ROC, variando el límite de 0 a 1 podemos generar una curva ROC como la mostrada en la Figura 2.9.

Un buen clasificador producir una curva que cae en la esquina superior izquierda. El área debajo de la curva ROC es una métrica de desempeño y puede ser calculada usando la regla trapezoidal. Si un clasificador produce una curva con área menor a 0.5 no tendrá un buen desempeño, por lo tanto, un clasificador ideal producirá un área igual a 1.

2.9 Descomposición en Valores Singulares

La descomposición en valores singulares (SVD, por sus siglas en inglés), es una técnica comúnmente usada para el análisis de problemas multiclase o de múltiples salidas, que permite la caracterización de estructuras en los datos. El análisis SVD identifica subespacios que capturan la mayor parte de la varianza en los datos. La proyección de los datos en subespacios de SVD y visualización de los datos en gráficas de dispersión puede revelar estructuras en los datos que pueden ser utilizadas para clasificación. Una descomposición en valores singulares de una matriz X de $N \times p$ es una factorización en el producto de tres matrices:

$$X = UDV^T, \tag{2.16}$$

donde U es una $N \times p$ matriz ortonormal ($U^T U = I_p$) cuyas columnas u_j son llamadas *vectores singulares de la izquierda*; V y es una $p \times p$ matriz ortonormal ($V^T V = I_p$) con columnas v_j llamados *vectores singulares de la derecha*, y D es una $p \times p$ matriz diagonal cuyos elementos diagonales $d_1 \geq d_2 \geq \cdots d_p \geq 0$, conocidos como los valores singulares. Por convención, el orden de los vectores singulares está dado en orden descendente de los valores singulares, con el valor singular más alto en el índice superior izquierdo de la matriz D. Para cualquier matriz X, la secuencia de valores singulares es única y si todos los valores singulares son

distintos, por lo tanto, la secuencia de vectores singulares también es única. Sin embargo, cuando un conjunto de valores singulares es igual, los correspondientes vectores singulares abarcan algunos subespacios. Cualquier conjunto de vectores ortonormales que abarca este subespacio se puede usar como vectores singulares. Los métodos usados para deducir la SVD pueden ser usados como un algoritmo, existen técnicas más sofisticadas, cualquier paquete de cálculo matricial contiene implementaciones numéricas estables de SVD. Por lo tanto, los detalles prácticos de la SVD son muy complicados y no serán discutidos.

Capítulo 3

Experimentación

3.1 Introducción

La experimentación con colonias de abejas se dividió en dos etapas. En la primera etapa se analizaron dos colonias de abejas en condiciones controladas, fue destinada a la prueba del sistema de monitoreo, al ajuste de la ganancia de los micrófonos, a analizar el consumo de energía del sistema, a corregir fallos en los códigos y principalmente a evaluar el desempeño de la metodología propuesta en la identificación de patrones en el sonido de las colonias de abejas. En la segunda etapa, una vez corregidos la mayor parte de los errores en el sistema de monitoreo, y corroborar que efectivamente la metodología propuesta es útil para tal fin, el objetivo fue poner a prueba el sistema de monitoreo en un ambiente real, es decir, en un apiario. Esta vez se analizaron cinco colonias de abejas bajo distintas condiciones biológicas. Se estudió el caso de una colonia con reducida población y el caso de colonias sin reina. En las siguientes secciones se explicará la metodología de análisis, las características de las colonias estudiadas, la configuración del sistema de monitoreo y finalmente se presenta la metodología que fue utilizada para encontrar los parámetros de monitoreo que por un lado reduzcan el espacio de almacenamiento de las muestras y al mismo tiempo garanticen un buen rendimiento del modelo de clasificación.

3.2 Colonias de abejas

Los experimentos se realizaron en colonias de abeja carniola (*Apis mellifera carnica*), una subespecie de abeja domestica de Europa. El apiario, Figura 3.1, de donde provienen las abejas de estudio se localiza en la comunidad de San José de la Era del municipio de Vetagrande, en el estado de Zacatecas. Durante la primera etapa se monitorearon dos núcleos[1]de abejas, cada uno con su respectiva abeja reina. Los núcleos se extrajeron del apiario para posteriormente ser trasladados a la ciudad de Zacatecas. De esta manera fue posible realizar inspecciones frecuentes y como se mencionó anteriormente identificar y corregir errores antes de realizar los experimentos en campo. Posterior a establecer los núcleos en la ciudad, uno de ellos perdió su reina y su población fue decayendo. La colmena que conservó su reina aumentó su población y se encontraba construyendo nuevos panales contiguos a los existentes. Durante esta etapa, se creó un modelo de clasificación con muestras de ambas colonias, los resultados de estos análisis mostraron que la metodología puede identificar diferencias entre estas colonias, sin

[1]Son colonias fundadoras de nuevas colmenas, compuestas de 3 o 4 bastidores o marcos de madera.

Figura 3.1: Apiario en San José de la Era.

embargo, era necesario el análisis de más colonias para reforzar lo antes encontrado. Durante la segunda etapa, los experimentos se realizaron en el apiario. Esta vez, se analizaron cinco colonias de abejas, cuatro colmenas con reina y una con reducida población y sin reina.

3.3 Condiciones en las colonias de abejas

Las condiciones o casos estudiados fueron el reducido número de abejas y el estado de orfandad de una colonia sana. El reducido número de abejas puede considerarse un caso especial de la falta de reina, es decir, la forma que se puede presentar es debido a que la reina haya muerto y que al no haber reina la población haya decaído, o al fenómeno de enjambrazón que de la misma manera que en el caso anterior la colonia haya quedado huérfana. La segunda condición, estado de orfandad, es difícil detectar el momento preciso cuando una colonia sana pierde su reina, por lo tanto, para emular esta condición las abejas reinas fueron removidas de sus colonias. Se experimentó con cinco colonias de abejas identificadas simplemente del 1 al 5 y se realizaron dos experimentos, tabla 3.1. Cuatro de las colonias tenían reina, colmena 1, 2, 4 y 5. Las colonias 1 y 2 en comparación con la colonia 4 y 5 se encontraban en excelente estado de salud, un mayor volumen poblacional y mayor actividad de pecoreo[2]. La colonia restante aunado a la reducida población no poseía reina y se perdió de manera natural. Todas comparaciones se realizaron únicamente por inspección visual. El grupo de colonias se monitoreó desde el 15 de marzo hasta 2 de mayo del 2018. El primer experimento fue analizar las colonias tal y como se encontraban, sin realizar ningún cambio, para observar las diferencias entre las colonias con reina y la colonia de reducida población. Así mismo, observar los cambios en los patrones de las colonias de manera individual a través del tiempo. El segundo experimento fue el remover las reinas de las colonias 2 y 4 para encontrar un patrón distintivo, recordando que la colonia 2 era de las más fuertes y saludables, y la colonia 4 en condiciones de volumen poblacional y de actividad de pecoreo menor a las colonias 1 y 2. Las señales de este experimento se analizaron a partir del segundo día después de la remoción de la reina. Justo después de la remoción de la reina, las colmenas se encontraban bastante

[2]Actividad que realizan las abejas de recoger el néctar de las flores.

excitadas y debido a la intensidad de la señal de audio los micrófonos se saturaron, por lo que estas señales no pudieron ser analizadas. Posterior a 24 horas después de la remoción de la reina, las colonias se encontraban en calma, esta vez las señales se encontraban en valores de amplitud normales.

3.4 Sistema de monitoreo

El área de la apicultura de precisión es una rama relativamente nueva donde el desarrollo se ha dado a paso lento [15, 16, 18, 17]. La inexistencia de bases de datos con información de patrones acústicos de colonias de abejas obliga a los investigadores a obtener sus propias bases de datos, además de tener que desarrollar sus propias herramientas para el monitoreo de colonias de abejas. Sin embargo, en la literatura existen distintas propuestas que se enfocan en metas similares, el desarrollo de sistemas autónomos y el uso de tecnologías inalámbricas para el acceso remoto [27, 24, 48, 25, 13, 49]. Basado en las diversas investigaciones se implementó un sistema de monitoreo enfocado principalmente en la adquisición y almacenamiento de las señales de audio para su posterior análisis. En la Figura 3.2, se muestra un esquema del sistema de monitoreo.

La encargada de ejecutar las tareas es una Raspberry Pi 2 (RPi2). La RPi2 es una computadora de bajo costo y dimensione reducida, puede ser programada en varios lenguajes como C, Python, entre otros. Una ventaja frente a dispositivos embebidos como un microcontrolador, es la facilidad al manejar las muestras recabadas, pudiendo almacenar varios vectores o arreglos de 20000 elementos sin saturar la memoria. Sin embargo, el principal inconveniente es su elevado consumo de energía (1 - 1.75 W) si se compara con un Arduino (125 mW). Se empleó un micrófono amplificado de condensador omnidireccional que incorpora ganancia ajustable a través de un potenciómetro, el módulo encargado de amplificar la señal y la reducción de ruido es el MAX4466. La característica de los micrófonos omnidireccionales es que permiten la captación del sonido independientemente de su origen. Su respuesta en frecuencia va de los 20 - 20k Hz. La ganancia puede ser ajustada de 25x a 125x, una característica importante cuando se monitorea sonido de abejas. La RPi2 carece de convertidor analógico digital (A/D) porque lo que fue necesario de un dispositivo adicional. Para realizar dicha función se utilizó un microcontrolador dsPIC33EP512, cuenta con seis entradas analógicas configurables a resoluciones de 10 ó 12 bits. Las entradas analógicas pueden ser muestreadas de manera simultánea. Para comunicar el dsPIC con la RPi2 se utilizó el bus SPI (Serial Peripherical Interface) que permite la transferencia de información a alta velocidad. Finalmente, para alimentar el sistema se empleó un panel solar de 10 W, con corriente máxima de salida de 0.58 A a 22 Vcc. Conectado a través de un controlador para panel solar a una batería de ácido-plomo de 12 V y 7 Ah.

3.4.1 Instalación de los micrófonos

Para la captura de sonidos durante la primera etapa de experimentos los micrófonos se posicionaron en la parte interior de la tapa de la colmena, Figura 3.3 y cerca del núcleo de abejas, Figura 3.4. En la primera etapa al ser las colonias de abejas de reducida población (solo cuatro bastidores) la colmena no tenía alza melaria[3], como en el caso del segundo experimento

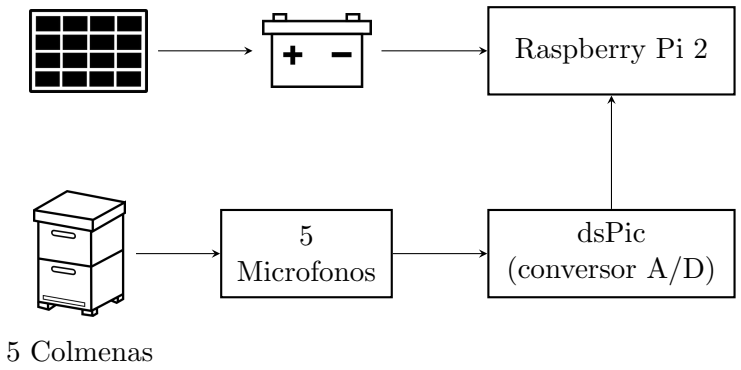


Figura 3.2: Configuración del sistema de monitoreo.

donde las colonias más fuertes tenían dos alzas melarias. Durante la segunda etapa, los sensores se posicionaron en un marco especial de madera, Figura 3.3(b), entre las alzas melarias y la cámara de la reina [4], Figura 3.4. En ambos casos se buscaba que los sensores se ubicaran lo más próximo posible a la cámara de la reina, que es donde se llevan a cabo la mayor parte de las actividades. Pero que al mismo tiempo no fuese invasivo y afectara las actividades de las abejas. Los micrófonos se cubrieron con malla metálica para evitar que fueran cubiertos con cera. No obstante, en ocasiones la malla metálica era cubierta de cera, por lo que era necesario removerla para evitar la obstrucción de la señal de audio

(a) (b)

Figura 3.3: Instalación de los micrófonos en el interior de las tapas de las colmenas (a) para el primer experimento y en el marco de madera (b) para el segundo experimento.

3.5 Metodología de análisis

El esquema de la Figura 3.5 muestra la metodología para el análisis identificación y clasificación de patrones en el sonido de colonias de abejas.

- **Adquisición de sonidos de abejas**. Es la adquisición de las señale de audio para

[3] Caja prevista para el almacenamiento de miel.
[4] Caja donde la reina pone huevecillos y nacen las crías.

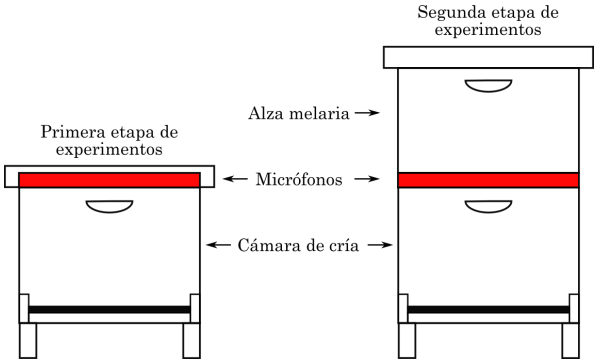


Figura 3.4: Instalación de los micrófonos en la camara de la cria en el primer experimento, y en un marco entre la cámara de la cría y la alza melaria en el segundo experimento.

Tabla 3.1: Experimentos realizados en colonias de abejas, CR = con reina, SR = sin reina.

		1	2	3	4	5
Colmena						
Experimento	Baja población	CR	CR	SR	CR	CR
	Falta de reina	SR	SR(removida)	SR	SR(removida)	CR

la creación de la base de datos. Los parámetros de monitoreo fueron una frecuencia de muestreo de 4 kHz, se realizaron adquisiciones en intervalos de 10 min con muestras u observaciones de 30 s de duración, la elección de estos parámetros será explicada en la siguiente sección.

- **Extracción de características**. Posterior a la adquisición de las muestras el siguiente paso es la extracción de características aplicando una metodología de Coeficientes Cepstrales en el escala de Mel, para posteriormente calcular de cada coeficientes de Mel la media, mediana, varianza, desviación estándar, curtosis y sesgo, finalmente el último paso del preprocesamiento es la normalización Z.

- **Selección de características y regularización**. En un siguiente paso se hizo una selección de las características para eliminar aquellas que sean irrelevantes para el modelo, el mismo paso permite también la regularización del modelo, es decir, evita que el modelo tenga sobreajuste o un ajuste pobre. Para analizar el problema multiclase o de múltiples salidas se utilizó una técnica conocida como uno contra todos, también llamada uno contra el resto, implica entrenar un clasificador binario por cada clase del problema, con las muestras de clase como positivas y el resto como negativas. Cada modelo devuelve la probabilidad de pertenecer a la clase. Las salidas o clases del modelo son: colonias con reina, sin reina (removida) y colonias de reducida población y sin reina.

- **Evaluación del desempeño**. Finalmente, cada modelo de clasificación binario fue evaluado obteniendo el área debajo de la curva ROC.

- **Análisis de conglomerados**. Se realizó un análisis mediante SVD para observar el comportamiento de la base de datos y la separación entre las muestras que representan

distintas condiciones. Si bien este análisis no es un requerimiento para la clasificación fue de especial ayuda para entender el comportamiento de los datos y la forma de realizar los modelos de clasificación.

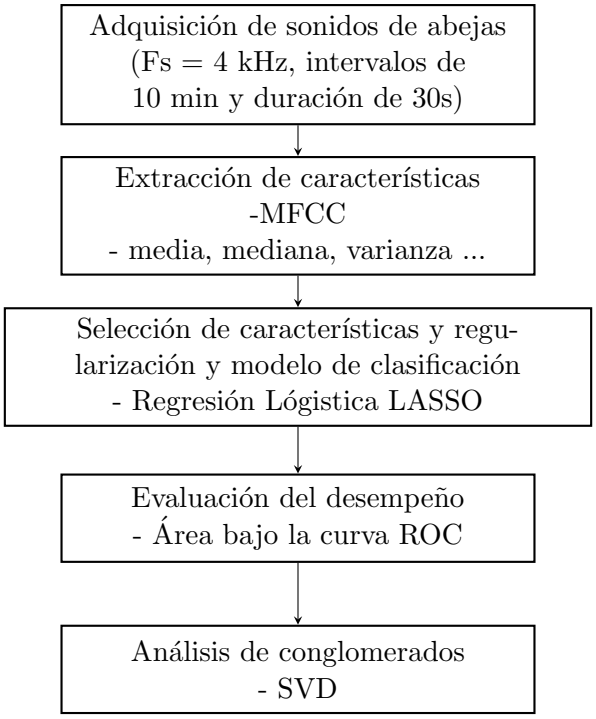


Figura 3.5: Etapas en la identificación, clasificación y análisis de sonidos de abejas.

3.6 Parámetros de monitoreo

Existe un número limitado de estudios del sonido de abejas como un organismo, es decir, al sonido que producen las colonias de abejas en conjunto y no de manera individual o producto de la comunicación entre individuos de la colonia. El sonido producido por colonias de abejas se encuentra en el rango desde aproximadamente 50 Hz y de manera habitual no supera 1 kHz [50, 1, 27]. Para obtener una buena calidad de las de las señales de audio se estableció la frecuencia de muestreo en 4 kHz que es el doble del requerido por el teorema de Nyquist. Sin embargo, los parámetros que han sido poco estudiados es los intervalos a los que se adquieren las muestras y la duración de éstas, y la forma en que impactan en el desempeño del modelo de clasificación. Algunos ejemplos de parámetros de monitoreo que se encuentran en la literatura se muestran en la tabla 3.2.

Se realizó un análisis para determinar los parámetros de monitoreo que, por un lado, permitan un balance entre un modelo que logre un buen rendimiento de clasificación, y que al mismo tiempo, no requiera excesivo espacio de almacenamiento. Lo anterior se verá reflejado en una reducción se de los tiempos de preprocesamiento y procesamiento de los datos, y una disminución en el consumo de energía del sistema de monitoreo. La propuesta de los parámetros de monitoreo fue: realizar tomas de muestras en intervalos de 10 min con duración

Tabla 3.2: Ejemplos de parámetros de muestreo en la literatura.

	Frecuencia de adquisición de muestras (min)	Duración de las muestras (s)
Mezquida, [51]	60	8
Howard, [39]	60	60
Cejrowski, [40]	15	1

de 30 s. El tiempo de grabación se estableció en 30 s en base a [46], parámetro que se utilizó para la clasificación de géneros musicales. Y la frecuencia en que de adquisición (10 min), permitirá la construcción de diferentes conjuntos de datos y evaluar el impacto de los intervalos de adquisición el rendimiento global de los modelos. A partir de las muestras adquiridas se crearon subconjuntos de datos de manera aleatoria segmentando las señales simulando señales de menor duración: 0.25, 0.5, 1, 5, 10 y 20 s. Al mismo tiempo, cada subconjunto se dividió en mas conjuntos variando el número de muestras: 120, 144, 180, 240, 360 y 720, lo que corresponde a adquirir muestras en intervalos de 60, 50, 40, 30, 20 y 10 min. En total se formaron 36 subconjuntos de datos, de cada subconjunto se evaluó el rendimiento obteniendo el área bajo la curva ROC y el error de predicción. Así mismo, estos datos se compararon con el requerimiento en espacio de cada subconjunto de muestras, para la selección del mejor modelo.

Capítulo 4

Resultados y Discusión

4.1 Análisis espectral de la señales

Para analizar la evolución de las señales en tiempo y frecuencia se obtuvieron espectrogramas mediante el paquete *signal* de Octave. Las espectrogramas mostrados en la Figura 4.1 corresponden al primero experimento, las colonias 1, 2, 4 y 5 son saludables, y la colonia 3 una reducida población, en los cuales es evidente la diferencia. En las colonias sanas, las bandas de mayor energía se encuentran entre los 250 y 300 Hz, que de manera habitual se encuentran en colonias sanas [20, 11, 13, 52]. Además, se observan otras bandas alrededor de los 100 Hz. Ambas bandas están ausentes en el espectrograma de la colonia 3, donde no se observan bandas definidas. En estos mismos espectrogramas se puede ver que la señal de la colonia es más débil en comparación con las colonias sanas, lo que podría estar directamente relacionado con la cantidad de abejas.

En el segundo experimento, en el cual se removieron las reinas de las colonias 2 y 4, los espectrogramas muestran que las bandas de mayor energía en los espectrogramas del primer experimento, Figura 4.2, ahora se han reducido, lo cual es particularmente evidente en la colmena 4. Sin embargo, es difícil hacer deducciones de los espectrogramas, porque no todos muestras poseen las mismas características. Este análisis únicamente proporciona información acerca de la naturaleza de las señales, de este análisis se pudo corroborar que las señales obtenidas tienen las características similares a las encontradas en la literatura. Además, sirven de apoyo en el ajuste de la ganancia en los micrófonos.

4.2 Extracción de características

El conjunto de datos consta de 144 grabaciones por cada colonia en un periodo de 24 hrs, cada muestra se grabó a intervalos de 10 minutos con 30 s de duración. La extracción de características se realizó en el software Octave con la función *mfcc*. Después de la extracción de características se calcularon la media, mediana, varianza, desviación, estándar, curtosis y sesgo de cada coeficiente de Mel, lo anterior con la intención de reducir aún más el conjunto de datos. El último paso es la normalización de los datos antes de introducir la base de datos al modelo de clasificación o de realizar el análisis SVD. El conjunto de datos está dispuesto de forma matricial y se compone de un total de 72 características (columnas) y 720 observaciones (filas).

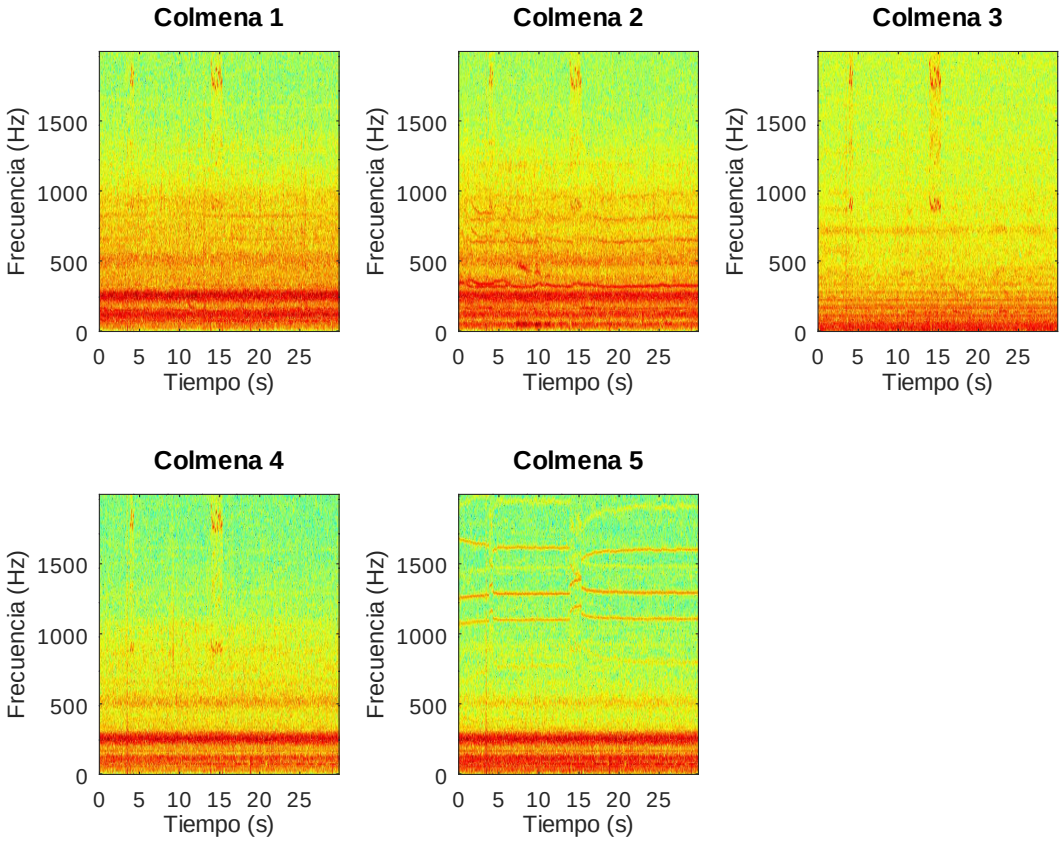


Figura 4.1: Espectrogramas de las colonias mostrando bandas de mayor energía entre los 250 a 300 Hz exceptuando por la colonia 3 que tiene una reducida población y no tiene reina.

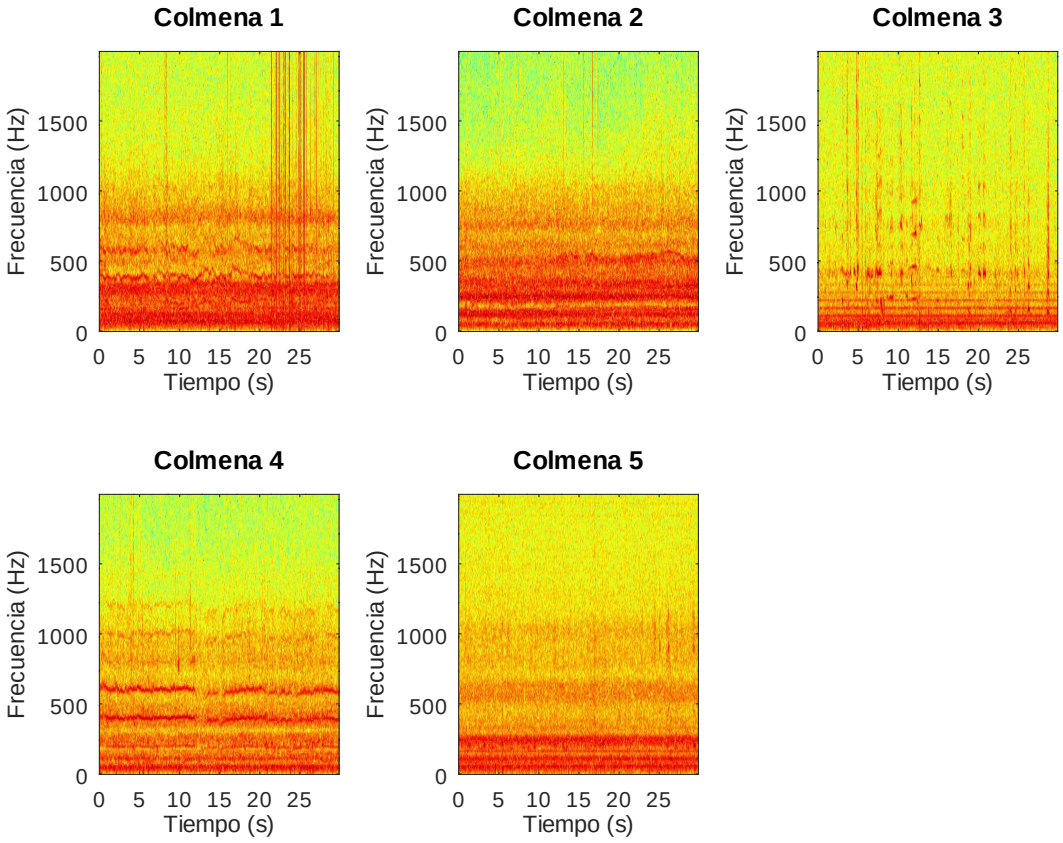


Figura 4.2: Espectrogramas posteriores a la remoción de la reina en las colmenas 2 y 4. En comparación con los espectrogramas anteriores las bandas de mayor energía parecen haberse reducido.

Tabla 4.1: Selección de características para cada clase del modelo multivariado.

Clase	Coeficientes de Mel	
	Media	Mediana
Colonias sanas (con reina)	3, 10, 12	5, 6, 9
Sin reina (removida)	1, 12	2, 3, 5, 6, 8, 10
Reducida población (sin reina)	9, 12	1, 2, 5, 9, 10

4.3 Selección de características, regularización y clasificación del modelo multiclase

La selección de características y regularización se realizó usando los datos del segundo experimento, es decir, posterior a la remoción de las reinas de las colonias. Para eliminar las características innecesarias del modelo se llevó a cabo una selección de mediante un modelo de regresión logística Lasso, usando el paquete *glment* en el software R. La selección se realizó aplicando la técnica de uno contra todos. En el caso bajo análisis se establecen tres clases, éstas son: colonias sanas (con reina), colonias sin reina (removida), y colonia de reducida población. Por lo tanto, se generan tres modelos de clasificación binarios. El conjunto de datos fue dividido de manera aleatoria en un conjunto de entrenamiento (70 %) y un conjunto de prueba (30 %). Los resultados de la selección de características se muestran en la tabla 4.1. En los tres modelos de clasificación las características más descriptivas fueron la media y la mediana, cada condición en la colonia está descrito por diferentes características. Por otro lado, la curtosis la desviación estándar el sesgo y la varianza, son irrelevantes para el modelo.

4.4 Evaluación de los modelos

Para evaluar el rendimiento de cada uno de los modelos se obtuvo el área bajo la curva ROC. Los tres modelos tiene un porcentaje de clasificación correcta por encima del 90 %, fila 1 tabla 4.2.

Tabla 4.2: Evaluación del rendimiento del modelo multiclase.

Clase		1 Con reina	2 Sin reina (removida)	3 Sin reina y reducida población
AUC	Un día	0.994	0.9766	0.997
	Dos días	0.999	0.996	0.999

La prueba anterior se hizo con los datos correspondientes a un día o 24 hrs de monitoreo. Para hacer una prueba que simularía condiciones más parecidas a las reales, se entrenó un modelo con datos de un día y se evalúo con datos de las muestras adquiridas un día después, sería el equivalente a una evaluación 50 % de datos de entrenamiento y 50 % de datos de prueba. Los resultados son incluso un moderadamente más altos, fila 2 tabla 4.2, que en el caso del modelo entrenado con datos de un día. De estos últimos resultados se podría pensar que un modelo entrenado con datos de un solo día puede ser útil para generalizar.

Para corroborar lo antes dicho, se realizó un análisis SVD en las muestras de las colonias de manera individual para observar el comportamiento de las muestras a través de los días, si los patrones se conservan o sufren modificaciones. Lo que resulta interesante del modelo de clasificación y de la evaluación del rendimiento es que fue posible establecer un análisis multiclase e identificar distintos estados o patrones en el sonido de abejas, cada sonido es característico y puede ser fácilmente clasificado por el modelo.

4.5 Análisis de conglomerado

Para analizar de manera visual la forma en que se agrupan las muestras se realizó una descomposición en valores singulares (SVD, por sus siglas en inglés) para crear gráficas de dispersión, este tipo de análisis es equivalente a realizar un análisis por componentes principales (PCA, por sus siglas en inglés). El primer grupo de muestras analizadas son las correspondientes al primer experimento, cuatro colonias sanas y una colonia de reducida población (colonia 3), Figura 4.3. En esta imagen es evidente que se forman tres grupos o conglomerados, el primero por muestras de las colonias 1 y 2, un segundo grupo formado por muestras de las colonias 4 y 5, todas las anteriores sanas y con reina, y un último grupo formado por muestras de la colonia 3. Es notable que las muestras de la colonia con baja población y sin reina, forma un grupo muy aparte del grupo de las muestras de las colonias sanas. Además, la separación entre las agrupaciones en las colonias sanas posiblemente esté relacionado con las características biológicas de la colonia, es decir, estado de salud, volumen poblacional, actividad de pecoreo, etc. Posiblemente exista alguna peculiaridad en el sonido que haga estas condiciones diferenciables al modelo. Como se mencionó antes, de manera visual se pudo percibir que la actividad de pecoreo era mayor en las colonias 1 y 2, con respecto a las colonias 4 y 5. Este mismo agrupamiento de las muestras se observó durante el tiempo de monitoreo de las colonias hasta la remoción de las reinas.

Después de remover las reinas de las colonias 2 y 4, se realizó nuevamente un análisis mediante SVD, Figura 4.4. En este caso las agrupaciones sufrieron cambios, ahora puede considerarse un grupo formado por las muestras de las colonias 1 y 5, que corresponde a las colonias donde no se removió la reina. Por otro lado, las muestras de las colonias 2, 3 y 4, todas colonias sin reina. El mismo patrón se observó únicamente por dos días, debido a que la constante nubosidad durante los días que se llevó a cabo los experimentos, como consecuencia, no fue posible para la celda solar captar la energía suficiente para recargar las baterías y mantener el sistema de monitoreo en funcionamiento.

4.5.1 Análisis de conglomerados por días

El siguiente paso fue analizar las muestras de las colonias de manera individual a través de los días y observar, si las agrupaciones de las muestras sufren alteraciones o si permanecen constantes. Se formaron conjuntos de datos de cuatro días no consecutivos, por cada 24 hrs de monitoreo se obtuvieron 144 muestras. En la Figura 4.5 se muestran los resultados del análisis, el conjunto de muestras en color azul corresponde al primer día de monitoreo, el conjunto en color verde corresponde al cuarto día de monitoreo, el conjunto en color naranja al sexto día y por último el conjunto azul marino al noveno día. Todos los datos corresponden al primer experimento, es decir, antes de remover las reinas. Observando las gráficas 1 y 2 se puede apreciar ligeros cambios en los conglomerados sin que sean significativos, y como se

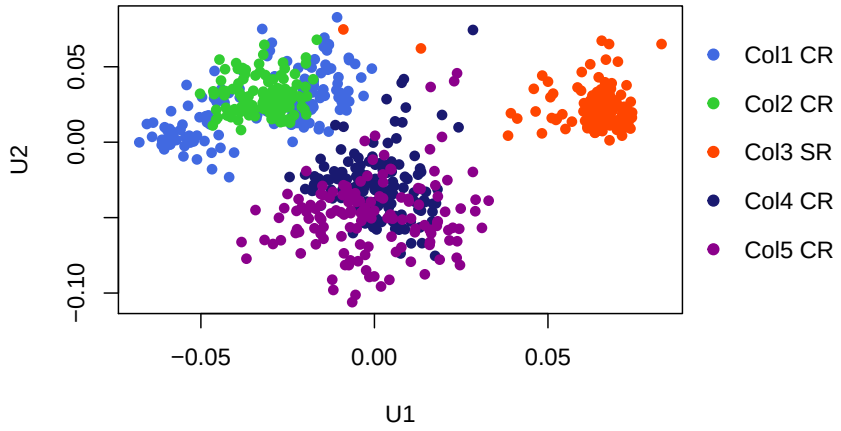


Figura 4.3: Descomposición en valores singulares, CR - con reina y SR - sin reina, toda las colonias con reina excepto por la colonia 3.

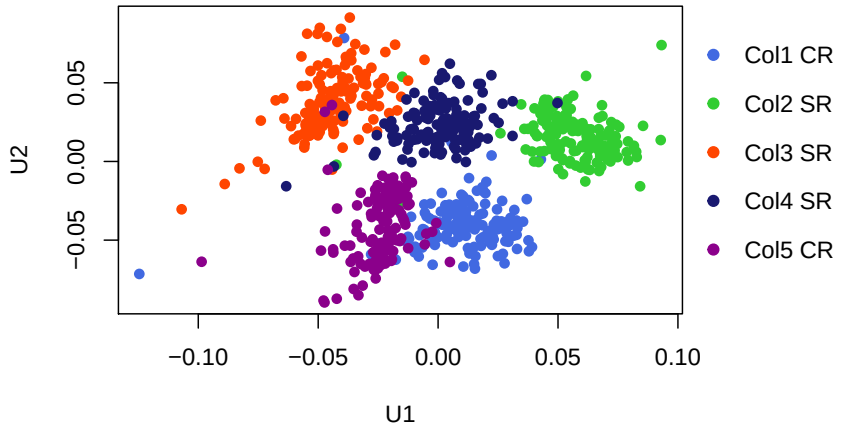


Figura 4.4: Descomposición en valores singulares, colonia 1 y 5 con reina, colonia 2 y 4 reina removida y colonia 3 inicialmente sin reina y baja población.

venía mencionando estos cambios pueden obedecer a mejores condiciones de salud y mayor actividad. En el resto de las colonias la organización de las muestras tiene cambios mínimos.

Se realizó una nueva comparación esta vez se hizo con muestras de las colonias antes y después de la remoción de la reina, la diferencia en tiempo entre las muestras es de 38 días, Figura 4.6. Una vez más se compararon las colonias de manera individual, en las gráficas las muestras en color azul corresponden al estado inicial de las colonias (colonia 1, 2, 4 y 5 con reina y la colonia 3 con reducida población) y las muestras en verde después de remover las reinas de colonias 2 y 4. Estas gráficas indican que en todas las colonias hubo cambios, incluso en las que no se removió la reina. De estos resultados se deduce que los patrones pueden sufrir variaciones a través del tiempo y que no simplemente están relacionados a la ausencia de reina. Si observamos gráfica de la colmena 1 en la Figura 4.6, presentan dos conglomerados que en un modelo de clasificación pueden ser fácilmente identificables.

El alto índice de clasificación correcta obedece a la naturaleza de las muestras con las que fue entrenado el modelo, si bien el conjunto de datos fue divido en 70 % entrenamiento y 30 % para prueba ciega. Las muestran tiene pocas variaciones a través del día, como se hace notar

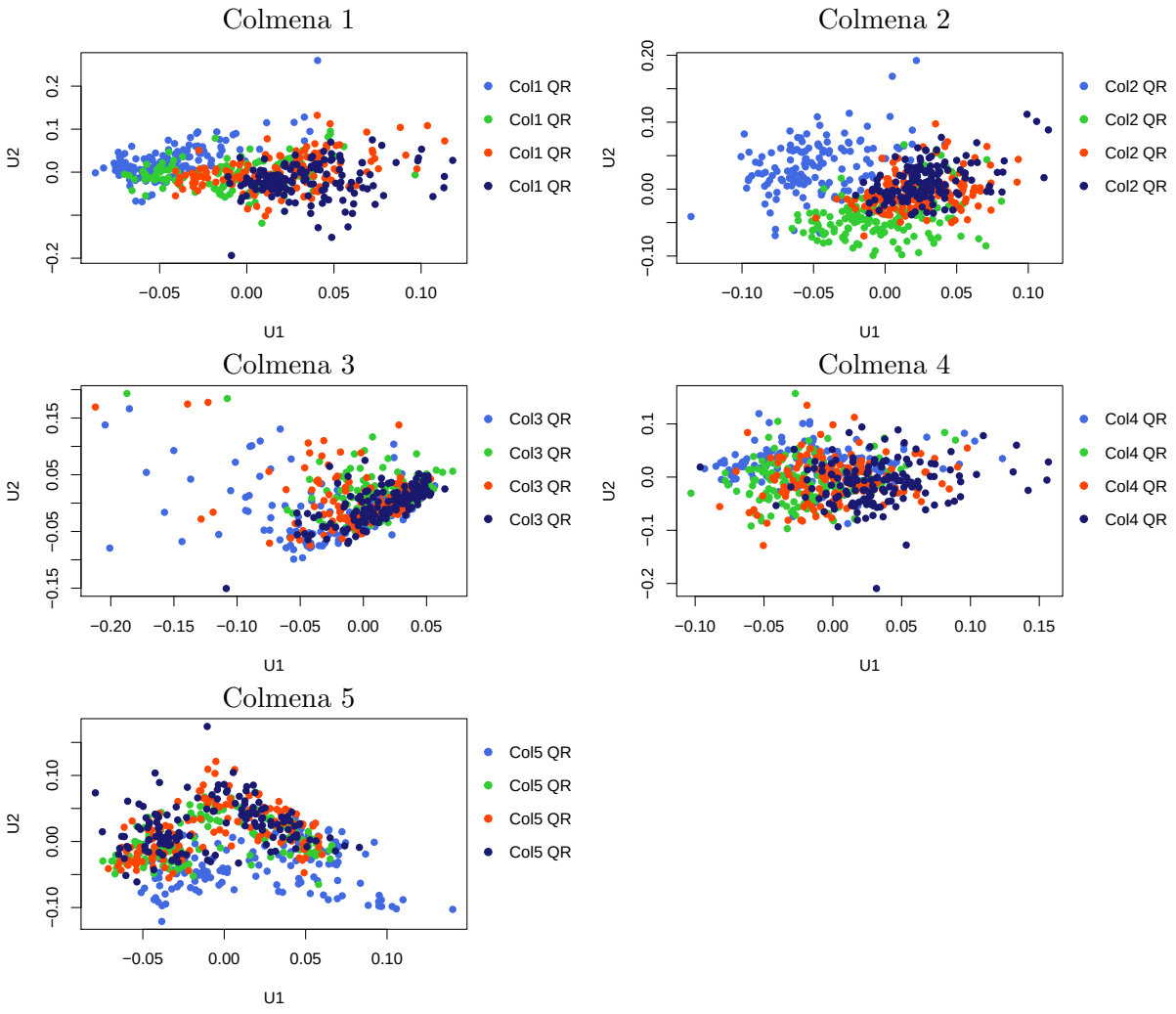


Figura 4.5: Análisis SVD de las colonias consigo mismas a través de distintos días, azul primer día, verde cuarto día, naranja sexto día y azul marino noveno día.

en el estudio de Perez [1] donde espectrogramas de señales de colonias de abejas adquiridos durante el día muestra pocas variaciones, además esto se evidencia en los análisis SVD en la Figura 4.3 y 4.4, donde los conglomerados que forman las muestras se compactan mostrado poca variabilidad en los datos, es decir, que las muestras son muy parecidas. Considerando lo anterior, la frecuencia a la que se adquieren las observaciones podría reducirse. En un inicio se estableció la adquisición de muestras a intervalos de 10 minutos, una reducción en el tamaño e intervalos de adquisición de las muestras conlleva a una reducción del espacio de almacenamiento, tiempos de preprocesamiento y procesamiento y reducción del consumo de energía.

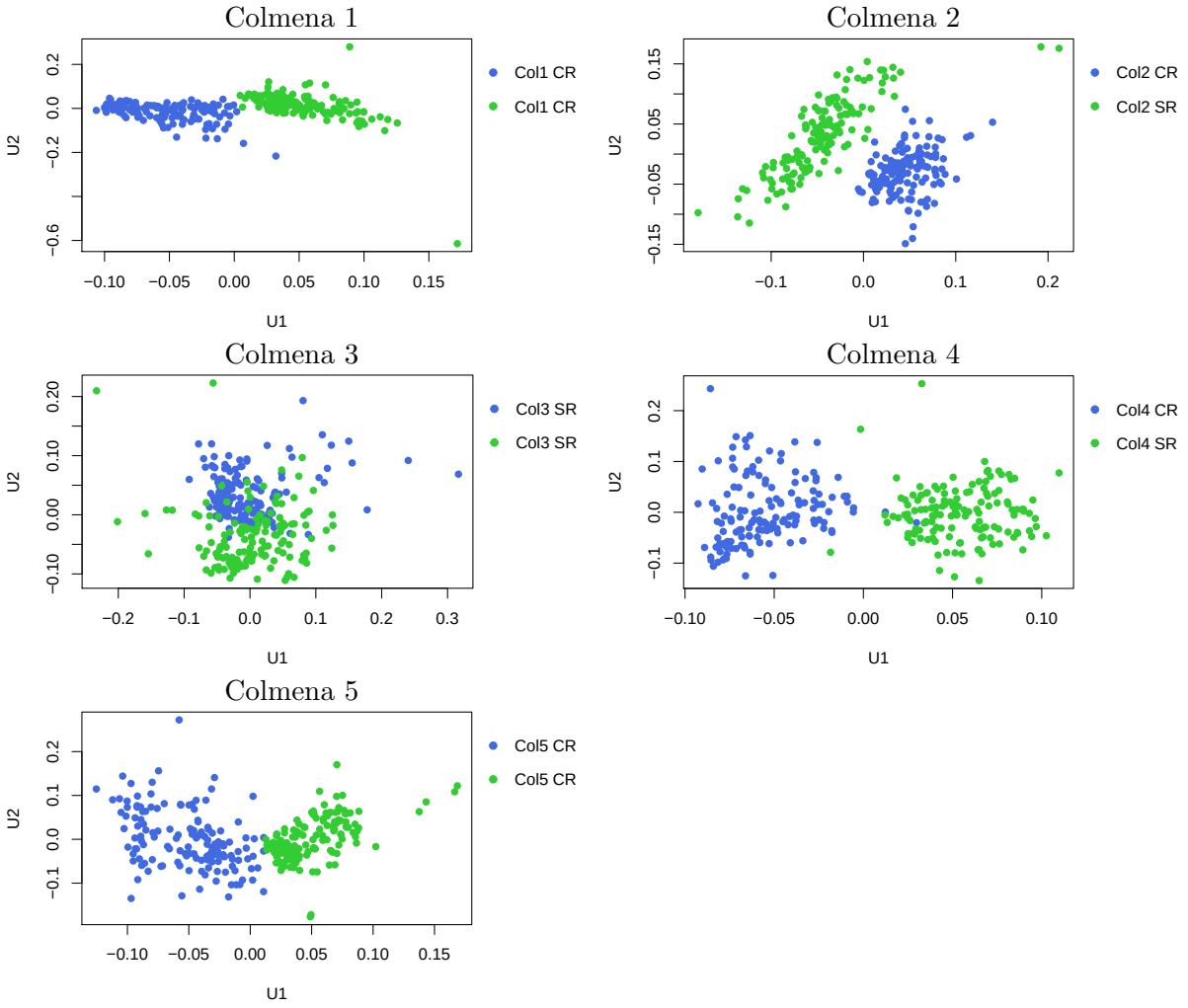
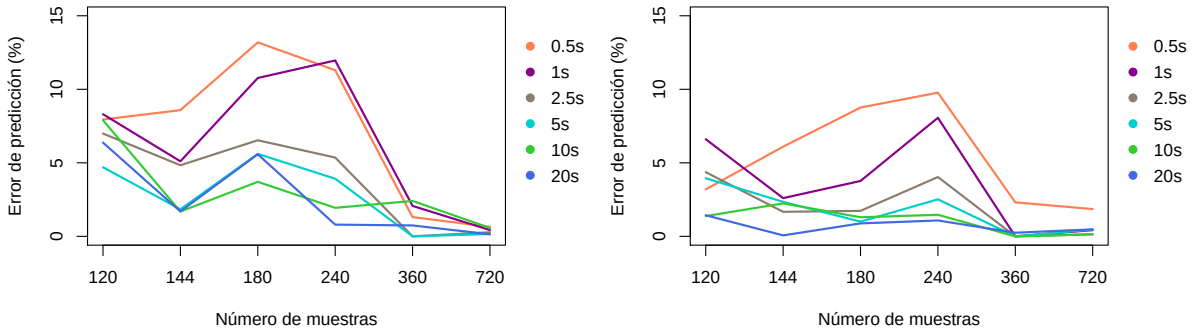


Figura 4.6: Comparativa de las colonias en periodos de un 36 dias de diferencia. En la colmena 1 y 5 las muestras en azul y verde corresponden a la colonia con reina. En la colmena 2 y 4 las muestras en azul corresponden a la colonia con reina y en verde posterior a la remoción de la misma. Finalmente en la colmena 3 ambos grupos de muestras son de la colonia sin reina.

4.6 Evaluación de los parámetros de monitoreo

Para encontrar los parámetros de muestreo que tengan un balance entre una alta tasa de correcta clasificación y que el espacio de almacenamiento en memoria sea mínimo, se evaluó el rendimiento de 36 modelos descritos en la sección 3.6. De los modelos se obtuvo el error de predicción en los datos de prueba y de entrenamiento, y el área bajo la curva ROC. En la primera gráfica, Figura 4.7 del error de predicción contra número de muestras, lo primero que se puede notar es que conforme el tamaño de las muestras aumenta en duración, el error de predicción disminuye, tanto en los datos de prueba, Figura 4.7(a), como en el error en los datos de entrenamiento, Figura 4.7(b). Las muestras mayores con duración mayor a 1s tiene el menor error, sin que la diferencia entre ellas sea significativa. Cuando en cualquiera de los conjuntos de datos sin importar la duración de las muestras el error de predicción se ve

disminuido cuando los conjuntos de datos llegan a las 360 muestras, muy cercano al obtenido con 720 muestras.



(a) Error de predicción en los datos de prueba.

(b) Error de predicción en los datos de entrenamiento.

Figura 4.7: Comparativa del error de predicción contra número de muestras de conjuntos de datos con distintas duraciones de muestras.

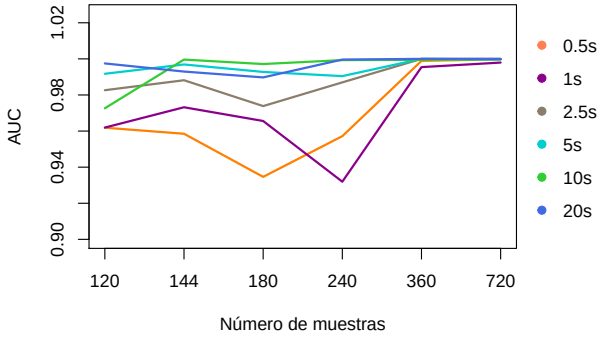


Figura 4.8: Evaluación de los modelos según la longitud de las muestras y el número de muestras.

En las gráficas de la Figura 4.8, de manera lógica las curvas ROC siguen el mismo comportamiento mostrado en las curvas de error de predicción, cuanto mayor es la duración de las muestras más cercano es el valor de AUC a 1. Igual que en el caso anterior, los modelos con muestras de longitud mayor a 5 s tienen el mejor comportamiento sin variaciones drásticas como las que se observan en las muestras de 0.5, 1 y 2.5 s. Cuando la cantidad de muestras en los modelos aumenta el error disminuye, pero incluso el modelo con muestras de 0.5s los modelos entregan un valor de área bajo la curva por encima de 0.9, por lo que se consideran como excelentes modelos.

Considerando las muestras inicialmente adquiridas a intervalos de 10 min y una longitud de 30 s por un total de 5 colonias se requiere de un espacio de almacenamiento de 500 MB. Si esto lo multiplicamos por 30 días de monitoreo, se requiere de 15 GB de almacenamiento, lo que es una cantidad considerable de información. Para la creación de modelos de clasificación que puedan generalizar se requeriría una mayor cantidad de información de colonias, por lo que es necesario una disminución del tamaño y cantidad de muestras. Para analizar los requerimientos de espacio de almacenamiento según la duración y el número de muestras se obtuvo el gráfico de la Figura 4.9, así mismo, en la tabla 4.3 se muestra el espacio requerido

por cada observación dependiendo de su duración. Con los datos de error de predicción y área bajo la curva ROC de los modelos estudiados, es posible hacer una mejor elección del tamaño de muestras que requieran un menor espacio en memoria. Por ejemplo, un modelo entrenado con muestras 5 s, el error de predicción y de prueba está por debajo del 5 %, Figura 4.7(a),además el AUC fue cercano a 0.98. Por otro lado, el requerimiento de espacio de almacenamiento con muestras en intervalos de 60 min es de aproximadamente 12 MB, en un periodo de 30 días se requeriría únicamente de 360 MB. Esto una reducción significativa del espacio de almacenamiento. Si bien como se dijo anteriormente, el modelo con muestras de 2.5 es suficiente, una selección más acertada es 5s, si en un futuro se desea agregar más condiciones al modelo, la identificación podría no resultar tan sencilla y agregar más información al modelo de predicción podría beneficiarle.

Tabla 4.3: Espacio requerido por las muestras según su duración.

Duración de las muestras (s)	30	20	10	5	2.5	1	0.5
Tamaño de las muestras (MB)	0.704	0.390	0.196	0.098	0.049	0.02	0.01

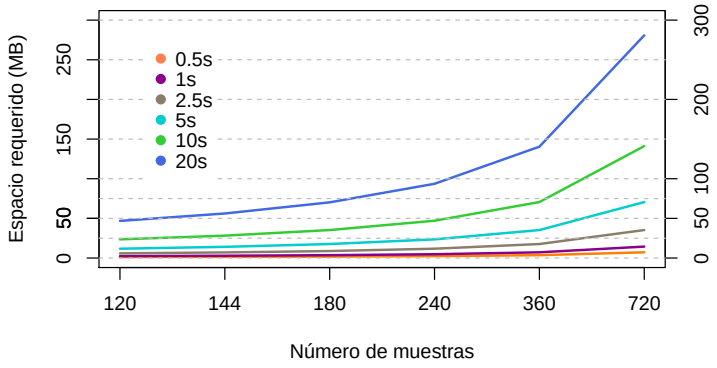


Figura 4.9: Espacio requerido por los conjuntos de datos de acuerdo a la longitud y número de muestras.

Conclusiones y trabajo a futuro

En el trabajo de investigación se desarrolló una metodología para el análisis, identificación y clasificación de patrones en el sonido de abeja, que son característicos de su estado de orfandad y bajo volumen poblacional, dichos patrones fueron comparados con patrones de colonias de abejas sanas, el modelo basado en MFCC y una regresión logística penalizada es sensible a los cambios en el sonido.

La selección de características mediante Regresión Logística Lasso reduce de manera significativa la cantidad de características del modelo, empleando únicamente la media y mediana de los coeficientes de Mel. Además, los modelos obtenidos están regularizados y con un alto grado de clasificación. Con la técnica de uno contra todos fue posible la construcción de un modelo multiclase para la clasificación de cada una de las condiciones que se estudiaron en las colonias. Sin embargo, aún se requiere más investigación, habría que comprobar que efectivamente colonias en condiciones similares ya sea sin reina, sanas, o baja población siguen la misma tendencia que se presentó en la Figura 4.4, es decir, que forman conglomerados fácilmente clasificables y que los conglomerados característico en cada condición puede tener ligeras variaciones. Las MFCC son lo bastante sensibles para detectar esos ligeros cambios en el sonido que pueden estar relacionados con el estado biológico de la colmena (volumen poblacional, estado de salud, actividad de pecoreo, etc.), como se observó en la Figura 4.3 entre las colonias con sanas existía una separación entre los grupos de datos.

La parte más fuerte del análisis se centró en el uso de gráficas de dispersión mediante SVD para la visualización y entendimiento de los conglomerados en las muestras, Figuras 4.3 y 4.4. Se demostró que los datos de distintas colonias de abejas forman agrupaciones fácilmente identificables. Sin embargo, a partir de estas muestras no es posible construir un modelo que tenga la capacidad de generalizar únicamente con datos un solo día, como se observó los patrones sufren cambios, Figura 4.6. Para que un modelo clasifique de manera correcta los estados en una colonia, habría que brindar información del ciclo anual de la abeja y de colonias muy variadas, como ya se observó los patrones de las colonias experimentan cambios través del tiempo. Otro aspecto relevante de los análisis SVD fue la comparación de una colonia consigo misma antes y después de remover la reina, crear un modelo con información de una sola colonia puede conducir a conclusiones erróneos. Como se observó en las gráficas 4.6 de las colmenas 1 y 5, donde no hubo remoción de reina las muestras sufrieron cambios que no estaban relacionados con el estado de orfandad. Un caso similar sucedió en [40], donde se realizó un experimento de remoción y introducción de una nueva reina, los patrones encontrados antes de la remoción de la reina y posteriores a la introducción de la reina eran distintos, y posiblemente se deban a la diferente reina. Del análisis de los parámetros de monitoreo 4.6 es posible hacer una reducción de los intervalos de adquisición de muestras que la inicialmente establecida en 10 min, además la longitud de las muestras (30s) también puede reducirse sin afectar de manera considerable el rendimiento del clasificador. Podríamos considerar distintos esquemas de monitoreo, por ejemplo, adquirir muestras en intervalos de una hora con dura-

ción de 2.5 s, cuyo error de predicción es menor en comparación con las muestras de menor duración, y al mismo tiempo tener valor de área bajo la curva ROC por encima de 0.95. El beneficio de esta reducción es, una disminución en el tiempo de procesamiento de los datos, pudiendo significar que en un dispositivo dedicado al reconocimiento se vea beneficiado en un ahorro de memoria, una disminución en el uso de recursos computaciones y una reducción en el consumo de energía. Lo anterior se cumple para el esquema de tres de las condiciones estudiadas (sana, sin reina y baja población), sin embargo, al introducir más clases o salidas al modelo pudiese ser que la longitud o número de muestras no sea suficiente para brindar experiencia al modelo.

El trabajo a futuro de manera lógica será el incluir información de más colonias al modelo de clasificación en diferente estado anual, y el estudio de más condiciones como es el caso de la varroasis provocado por el ácaro varroa o colonias en condición de preenjambrazón. Mejorar de la autonomía del sistema de monitoreo sustituyendo la Raspberry Pi convencional por una Raspberry pi Zero de menor consumo energético. Y una mejora en los micrófonos de ganancia ajustable a unos de ganancia automática para evitar la saturación de estos.

Referencias

[1] N. Pérez, F. Jesús, C. Pérez, S. Niell, A. Draper, N. Obrusnik, P. Zinemanas, Y. M. Spina, L. C. Letelier, and P. Monzón, "Continuous monitoring of beehives' sound for environmental pollution control," pp. 326–330, 2016.

[2] D. P. Abrol, *Pollination Biology: Biodiversity conservation and agricultural production*, 1st ed. Dordrecht: Springer Netherlands, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780125839808500193 http://link.springer.com/10.1007/978-94-007-1942-2

[3] B. R. Ochoa and R. C. Ortega, "Situación actual y perspectiva de la apicultura en México," *Claridades agropecuarias*, vol. 199, pp. 3–34, mar 2010.

[4] Unep, "Unep Emerging Issues: Global Honey Bee Colony Disorders and Other Threats To Insect," *Agriculture*, p. 16, 2011.

[5] S. G. Potts, J. C. Biesmeijer, C. Kremen, P. Neumann, O. Schweiger, and W. E. Kunin, "Global pollinator declines: Trends, impacts and drivers," *Trends in Ecology and Evolution*, vol. 25, no. 6, pp. 345–353, 2010.

[6] P. Jean-Prost, *Apicultura: conocimiento de la abeja. Manejo de la colmena.* Mundi-Prensa, 2007.

[7] R. F. A. Moritz and E. E. Southwick, *Bees as Superorganisms.* Berlin, Heidelberg: Springer Berlin Heidelberg, 1992.

[8] H. N. Wehmann, D. Gustav, N. H. Kirkerud, and C. G. Galizia, "The sound and the fury - Bees hiss when expecting danger," *PLOS ONE*, vol. 10, no. 3, pp. 1–17, 2015.

[9] C. G. d. G. Sagarpa, "Manual Básico Apícola," Coordinación General de Ganadería, SAGARPA, Tech. Rep., 2014.

[10] E. K. Eskov, *Acoustic Signaling in Social Insects*, Nauka, Moscow, 1979.

[11] E. K. Eskov and V. A. Toboev, "Changes in the structure of sounds generated by bee colonies during sociotomy," *Entomological Review*, vol. 91, no. 3, p. 347, Jun 2011.

[12] A. F. Rybochkin, "The Background Noise as a Source of Information about a Bee Colony," *Pchelovodstvo*, vol. 3, pp. 15–17, 2007.

[13] S. Ferrari, M. Silva, M. Guarino, and D. Berckmans, "Monitoring of swarming sounds in bee hives for early detection of the swarming period," *Computers and Electronics in Agriculture*, vol. 64, no. 1, pp. 72–77, 2008.

[14] R. Sudarsan, C. Thompson, P. G. Kevan, and H. J. Eberl, "Flow currents and ventilation in Langstroth beehives due to brood thermoregulation efforts of honeybees," *Journal of Theoretical Biology*, vol. 295, pp. 168–193, 2012.

[15] A. Zacepins, V. Brusbardis, J. Meitalovs, and E. Stalidzans, "Challenges in the development of Precision Beekeeping," *Biosystems Engineering*, vol. 130, pp. 60–71, 2015.

[16] A. Zacepins and V. Brusbardis, "Precision Beekeeping (Precision Apiculture): Research Needs and Status in Latvia," *Agriculture and Forestry*, vol. 61, no. 1, pp. 135–141, 2015.

[17] A. Zacepins, E. Stalidzans, and J. Meitalovs, "Application of information technologies in precision apiculture," 2012, p. Paper 1023.

[18] A. Zacepins and E. Stalidzans, "Information processing for remote recognition of the state of bee colonies and apiaries in precision beekeeping (apiculture)," *Biosystems and Information technology*, vol. 2, no. 1, pp. 6–10, 2013.

[19] A. Kviesis and A. Zacepins, "System architectures for real-time bee colony temperature monitoring," *Procedia Computer Science*, vol. 43, pp. 86–94, 2015.

[20] F. Edwards-Murphy, B. Srbinovski, M. Magno, E. M. Popovici, and P. M. Whelan, "An automatic, wireless audio recording node for analysis of beehives," *26th Irish Signals and Systems Conference*, pp. 1–6, 2015.

[21] F. Edwards-Murphy, E. Popovici, P. Whelan, and M. Magno, "Development of an heterogeneous wireless sensor network for instrumentation and analysis of beehives," in *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, 2015, pp. 346–351.

[22] M. Giammarini, E. Concettoni, C. C. Zazzarini, N. Orlandini, M. Albanesi, and C. Cristalli, "BeeHive Lab Project - Sensorized Hive for Bee Colonies Life Study," in *Intelligent Solutions in Embedded Systems (WISES)*, 2015, pp. 121–126.

[23] O. A. M. Reyes, Á. A. M. Avila, G. Sebastian Eslava, and G. B. Rozo, "Beekeeping monitoring module," *2012 IEEE 4th Colombian Workshop on Circuits and Systems, CWCAS*, 2012.

[24] W. Chen, C. Wang, J. Jiang, and E. Yang, "Development of a monitoring system for honeybee activities," in *2015 9th International Conference on Sensing Technology (ICST)*, Taiwan, 2015, pp. 745–750.

[25] F. Edwards-Murphy, M. Magno, L. O'Leary, K. Troy, P. Whelan, and E. M. Popovici, "Big brother for bees (3B) - Energy neutral platform for remote monitoring of beehive imagery and sound," in *2015 6th IEEE International Workshop on Advances in Sensors and Interfaces, IWASI*, 2015, pp. 106–111.

[26] G. J. Tu, M. K. Hansen, P. Kryger, and P. Ahrendt, "Automatic behaviour analysis system for honeybees using computer vision," *Computers and Electronics in Agriculture*, vol. 122, pp. 10–18, 2016.

[27] A. Qandour, I. Ahmad, D. Habibi, and M. Leppard, "Remote Beehive Monitoring Using Acoustic Signals," *Acoustics Australia*, vol. 42, no. 3, pp. 204–209, 2014.

[28] A. Zacepins and J. Meitalovs, "Implementation of multi-node temperature measurement system for bee colonies online monitoring," *15th International Carpathian Control Conference, ICCC*, pp. 695–698, 2014.

[29] A. Zacepins, A. Kviesis, E. Stalidzans, M. Liepniece, and J. Meitalovs, "Remote detection of the swarming of honey bee colonies by single-point temperature monitoring," *Biosystems Engineering*, vol. 148, pp. 76–80, 2016.

[30] F. Kronenberg and H. C. Heller, "Colonial thermoregulation in honey bees (Apis mellifera)," *Journal of comparative physiology*, vol. 148, no. 1, pp. 65–76, 1982.

[31] A. Zacepins and T. Karasha, "Application of temperature measurements for bee colony monitoring: A review," *Engineering for Rural Development*, pp. 126–131, 2013.

[32] J. Meitalovs, a. Histjajevs, and E. Stalidzans, "Automatic Microclimate Controlled Beehive Observation System," *8th International Scientific Confernce 'Enginieering for Rural Development'*, pp. 265–271, 2009.

[33] E. K. Es'kov and V. A. Toboev, "Heating of wintering bee bodies related to external air temperature," *Entomological Review*, vol. 89, no. 1, pp. 111–112, 2009.

[34] J. A. Shaw, P. W. Nugent, J. Johnson, J. J. Bromenshenk, C. B. Henderson, and S. Debnam, "Long-wave infrared imaging for non-invasive beehive population assessment," *Opt. Express*, vol. 19, no. 1, pp. 399–408, Jan 2011.

[35] E. Stalidzans and A. Berzonis, "Temperature changes above the upper hive body reveal the annual development periods of honey bee colonies," *Computers and Electronics in Agriculture*, vol. 90, pp. 1–6, 2013.

[36] D. Atauri Mezquida and J. Llorente Martinez, "Short communication. Platform for beehives monitoring based on sound analysis. A perpetual warehouse for swarm's daily activity," *Spanish Journal of Agricultural Research*, vol. 7, no. 4, pp. 824–828, 2009.

[37] T. J. Brundage, "Acoustic Sensor for Beehive Monitoring," mar 2010.

[38] J. J. Bromenshenk, C. Henderson, R. Seccomb, S. Rice, and R. Etter, "Honey bee acoustic recording and analysis system for monitoring hive health," 2009. [Online]. Available: http://www.google.com/patents

[39] D. Howard, O. Duran, G. Hunter, and K. Stebel, "Signal Processing the acoustics of honeybees (APIS MELLIFERA) to identify the "queenless"state in Hives," *Proceedings of the Institute of Acoustics*, vol. 35, pp. 290–297, 2013.

[40] T. Cejrowski, J. Szymański, H. Mora, and D. Gil, "Detection of the bee queen presence using sound analysis," in *Intelligent Information and Database Systems*, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, Eds. Cham: Springer International Publishing, 2018, pp. 297–306.

[41] J. Campbell, L. Mummert, and R. Sukthankar, "Video monitoring of honey bee colonies at the hive entrance," *Visual observation & analysis of animal & insect behavior, ICPR*, vol. 8, no. 1, pp. 1–4, 2008.

[42] M. Bencsik, J. Bencsik, M. Baxter, A. Lucian, J. Romieu, and M. Millet, "Identification of the honey bee swarming process by analysing the time course of hive vibrations," *Computers and Electronics in Agriculture*, vol. 76, no. 1, pp. 44–50, 2011.

[43] H. Human, R. Brodschneider, V. Dietemann, G. Dively, J. D. Ellis, E. Forsgren, I. Fries, F. Hatjina, F.-L. Hu, R. Jaffé, A. B. Jensen, A. Köhler, J. P. Magyar, A. Özkýrým, C. W. W. Pirk, R. Rose, U. Strauss, G. Tanner, D. R. Tarpy, J. J. M. van der Steen, A. Vaudo, F. Vejsnæs, J. Wilde, G. R. Williams, and H.-Q. Zheng, "Miscellaneous standard methods for Apis mellifera research," *Journal of Apicultural Research*, vol. 52, no. 4, pp. 1–53, 2013.

[44] A. Zacepins and E. Stalidzans, "Architecture of automatized control system for honey bee indoor wintering process monitoring and control," *13th International Carpathian Control Conference, ICCC*, pp. 772–775, 2012.

[45] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-frequency cepstral coefficients on the power spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2001, pp. 1764–1769.

[46] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *Transctions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[47] L. Ma, B. Milner, and D. Smith, "Acoustic Environment Classification," *Transactions on Speech and Language Processing*, vol. 3, no. 2, pp. 1–22, 2006.

[48] E. K. Eskov and V. A. Toboev, "Analysis of statistically homogeneous fragments of acoustic noises generated by insect colonies," *Biophysics*, vol. 55, no. 1, pp. 92–103, 2010.

[49] S. Gil-Lebrero, F. Quiles-Latorre, M. Ortiz-López, V. Sánchez-Ruiz, V. Gámiz-López, and J. Luna-Rodríguez, "Honey Bee Colonies Remote Monitoring System," *Sensors*, vol. 17, no. 1, p. 55, 2016.

[50] D. G. Dietlein, "A method for remote monitoring of activity of honeybee colonies by sound analysis," *Journal of Apicultural Research*, vol. 24, no. 3, pp. 176–183, 1985.

[51] D. Atauri Mezquida, "El sonido producido por las colonias de Apis mellifera como indicador de su estado fenológico y sanitario ." Tesis Doctoral, Universidad Rey Juan Carlos, 2010.

[52] J. Tlačbaba, M. Černý, P. Dostál, and A. Přidal, "The acoustic emission in the nest of the honey bee depending on the extreme weather conditions," *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, vol. 62, no. 1, pp. 245–254, 2014.

Bibliografía

[1] K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," *Analytica Chimica Acta*, vol. 703, no. 2, pp. 152–162, 2011.

[2] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[3] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[5] D. J. Hand and R. J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[6] T. Fawcett, "An introduction to ROC analysis environments support," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[7] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*. New York, New York, USA: ACM Press, 2006, pp. 233–240.

[8] Universidad de Stanford, "Coursera," 2017. [Online]. Available: https://es.coursera.org/learn/machine-learning

[9] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, oct 2013.

[10] T. Fischetti, *Data Analysis with R*. Birmingham, UK.: Packt Publishing Ltd., 2015.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, segunda ed., ser. Springer Series in Statistics. New York, NY: Springer New York, 2009, vol. 1.

[12] E. Mayor, *Learning Predictive Analytics with R*. Birmingham, UK.: Packt Publishing Ltd., 2015.

[13] J. Brownlee, *Machine Learning Mastery with R*, Melbourne, Australia, 2017.

[14] K. Ramasubramanian and A. Singh, *Machine Learning Using R*. Berkeley, CA: Apress, 2017.

[15] D. P. Berrar, W. Dubitzky, and M. Granzow, *A Practical Approach to Microarray Data Analysis*, 1st ed. Springer US, 2003.

Anexo 1

Publicaciones

- Antonio Robles-Guerrero, Tonatiuh Saucedo-Anaya, Efrén González-Ramírez and Carlos E. Galván-Tejada, "Frequency Analysis of Honey Bee Buzz for Automatic Recognition of Health Status; a preliminary study". International Conference on Computer Networks Applications (ICCNA), *Advances in Computer Network Applications, Proc. Research in Computer Science*, vol. 142, 2017.
- Antonio Robles-Guerrero, Tonatiuh Saucedo-Anaya, Efrén González-Ramírez and José Ismael De la Rosa Vargas, "Analysis of a multiclass classification problem by Lasso Logistic Regression and Singular Value Decomposition to identify sound patterns in queenless bee colonies". *Computers and Electronics in Agriculture*, vol. 159, pp. 69-74, 2019.

ISSN 1870-4069

Frequency Analysis of Honey Bee Buzz for Automatic Recognition of Health Status: A Preliminary Study

Antonio Robles-Guerrero[1], Tonatiuh Saucedo-Anaya[2], Efrén González-Ramírez[1], Carlos E. Galván-Tejada[1]

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, Mexico

[2] Universidad Autónoma de Zacatecas, Unidad Académica de Física, Zacatecas, Mexico

aroblesp@uaz.edu.mx, gonzalez_efren@hotmail.com, ericgalvan@uaz.edu.mx, tsaucedo@fisica.uaz.edu.mx

Abstract. The study of honey bee health has received special attention in the last years. Researchers has been monitoring physical variables to determinate the status of the colony. This is a first approach in the development of a real time monitoring system to provide useful information to beekeepers that will help them to prevent colony losses. This study presents an analysis of the sound from two colonies of bees in the Mel frequency domain. The first is a healthy colony with queen and the second one is a hive with no queen and with a reduced population. Sound samples were acquired for each colony and characterized using Mel Frequency Cepstral Coefficients (MFCC). To summarizes the information, statistical descriptors was obtained for each Mel coefficient. An exploratory analysis of samples revealed two different hive characteristics; the presence and lack of a queen bee. For honey bee buzz recognition, a Logistic Regression Model was used. The preliminary results show that it is possible to classify both characteristics obtaining high classification rates using a reduced set of features.

Keywords: honey bee, remote sensing, beehive monitoring, queenless state.

1 Introduction

Pollinators are essential for diet diversity, biodiversity, and the maintenance of natural resources. The honey bee is the most important pollinator. Approximately 73% world cultivated crops depend on some variety of bees [1]. The colony health is influenced by external factors, such as, increase of pathologies,

Computers and Electronics in Agriculture 159 (2019) 69–74

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Analysis of a multiclass classification problem by Lasso Logistic Regression and Singular Value Decomposition to identify sound patterns in queenless bee colonies

Antonio Robles-Guerrero[a,*], Tonatiuh Saucedo-Anaya[b], Efrén González-Ramírez[a], José Ismael De la Rosa-Vargas[a]

[a] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, 98000 Zacatecas, Zac., Mexico
[b] Unidad Académica de Física, Universidad Autónoma de Zacatecas, Calzada Solidaridad esq. Paseo, La Bufa s/n, 98060 Zacatecas, Zac., Mexico

ARTICLE INFO

Keywords:
Queenless state
Beehive monitoring
Bee sound
Sound analysis

ABSTRACT

This study presents an analysis of a multiclass classification problem to identify queenless states by monitoring bee sound in two possible cases; a strong and healthy colony that lost its queen and a reduced population queenless colony. The sound patterns were compared with patterns of healthy queenright colonies. Five colonies of Carniola honey bee were monitored by using a system based on a Raspberry Pi 2 and omnidirectional microphones placed inside the hives. Feature extraction was carried out by Mel Frequency Cepstral Coefficients (MFCCs) method. A multiclass model with three outcome variables was constructed. For feature selection and regularization, a Lasso logistic Regression model was used along with one vs all strategy. To provide visual evidence and examine the results, data was analyzed by scatter plots of Singular Value Decomposition (SVD). The results show that is possible to detect the queenless state in both cases. Queenless or healthy colonies can generate slightly different patterns and the data clusters of the same condition tend to be close. The proposed methodology can be applied for the analysis of more conditions in bee colonies.

1. Introduction

The queen bee is the most important bee in a colony, she plays a vital role inside the hive, she is the only one who lays eggs; around 2000 eggs per day. The queen bee maintains the order of the colony primarily by the utilization of pheromones which transmit information within colony members. It is common to find queenless colonies, the queen bee could accidentally die due to illness or bad beekeeping practices. When a colony loses its queen, all bees tend to change their behavior from a state of organized activity to one disorganized (Butler, 1954). Loss of the queen bee during the period when there are no males can result in the death of the colony. Apart from the traditional techniques (beehive inspection) there are no alternative methods for queenless state identification.

Researchers have been analyzing different parameters by using sensors to determinate the colony state. The most studied parameters are temperature, humidity, vibrations, weight and sound (Zacepins et al., 2015; Meikle and Holst, 2015; Stalidzans and Berzonis, 2013; Bencsik et al., 2011). The application of technology and data analysis results in the development of Precision Beekeeping (PB). The PB has been defined as an apiary management strategy based on the monitoring of individual bee colonies to minimize resource consumption and maximize the productivity of bees (Henry et al., 2019; Zacepins et al., 2015, 2012; Kviesis et al., 2015). To determine the colony state, i.g., changes in bee sound while preparing to swarm were studied in Ferrari et al. (2008) and Eskov and Toboev (2011); the results claim that sound can be used to predict this event, monitoring the changes in the sound frequency prior to swarming process. Varroa is probably the most dangerous mite of honey bees (Moritz and Southwick, 1992). Infestation can change the sound produced by a healthy colony, in Qandour et al. (2014), a comparison of a healthy colony with an infested one was analyzed, the results show that there are differences in sound produced by infested colonies. In Bromenshenk et al. (2015) a review of studies about bee as biosensor to detect the presence of toxic chemicals and bee pest is presented, preliminary results show that colonies tend to be noisier when exposed to organic insecticides. The queenless state detection by sound analysis was previously studied in Howard et al. (2013), the results show that the signal spectrum emitted by queenrigth colonies is different from queenless colonies, however, the classification method used was not successful. Another example is

* Corresponding author.
E-mail addresses: aroblesp@uaz.edu.mx (A. Robles-Guerrero), tsaucedo@fisica.uaz.edu.mx (T. Saucedo-Anaya), ismaelrv@ieee.org (J.I. De la Rosa-Vargas).

https://doi.org/10.1016/j.compag.2019.02.024
Received 11 September 2018; Received in revised form 16 January 2019; Accepted 20 February 2019