Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Geolocalización de pájaros mediante radiotelemetría y machine learning

Tesis de Licenciatura en Ciencias de la Computación

Axel Ariel Lew

Director: Agustín Gravano

Buenos Aires, 2020

GEOLOCALIZACIÓN DE PÁJAROS MEDIANTE RADIOTELEMETRÍA Y MACHINE LEARNING

El campo de Inteligencia Artificial, y en particular el Aprendizaje Automático, es cada día más utilizado y tiene cada vez mayor adopción tanto en diferentes industrias como en investigaciones de diversas áreas. Se pueden encontrar usos en campos como medicina, psiquiatría o finanzas, donde su uso trae grandes beneficios. En esta tesis nos proponemos utilizar herramientas del área de aprendizaje automático para realizar un trabajo colaborativo con investigadores del área de biología, quienes estudian el comportamiento del *Tordo Pico Corto*, una especie de pájaro que tiene un comportamiento parasitario. A partir de radiotransmisores colocados en ejemplares de esta especie, y de antenas colocadas en el área de estudio, utilizaremos el aprendizaje automático para poder estimar la posición de estos y así realizar diferentes estudios y análisis sobre el comportamiento de los mismos. La hipótesis central que vamos a analizar sugiere que los pájaros de esta especie tiene un comportamiento social monógamo. A través de diferentes experimentos, analizando el tiempo que cada pareja de pájaros pasan juntos, y la distancia a la que se encuentran entre sí, encontramos evidencia que sugiere que esta hipótesis es cierta.

Palabras claves: Comportamiento Animal, Parasitismo, Radio-telemetría, Aprendizaje Automático

AGRADECIMIENTOS

A mi familia, por estar ahí siempre, porque esto no hubiese sido posible sin ellos.

A mis amigos y compañeros de cursada, por tantos buenos momentos, por su apoyo y por hacer de esta una experiencia única.

A mi director Agustín Gravano y a Romina Scardamaglia por su ayuda y apoyo durante esta tesis, y a todos los docentes del DC por todo lo aprendido en este camino.

Índice general

1. INTRODUCCIÓN A LA PROBLEMÁTICA

Dentro de las áreas de investigación de las ciencias biológicas, se encuentra el estudio del comportamiento animal. Esta área estudia por qué los animales se comportan como lo hacen, y la manera en que interactúan con el entorno, el ecosistema y otros seres vivos. El estudio del comportamiento animal es importante, no solamente por el lado científico, de querer conocer y entender el universo y los elementos (en este caso, animales) que lo componen, sino que su estudio también puede tener impacto en la conservación de la flora y fauna. Conocer cuál puede ser el impacto de introducir nuevas especies o la reducción en cantidad de alguna especie en un ecosistema, nos puede ayudar a la preservación del mismo.

En esta tesis en particular, nos interesa el estudio de los pájaros, y más precisamente la especie Tordo Pico Corto (*Molothrus rufoaxillaris*). El *Tordo Pico Corto* es una especie de ave que se caracteriza por realizar parasitismo de nidos, que consiste en delegar la responsabilidad de la cría de sus descendientes en otras especies (llamados hospedadores). En el caso del ave de estudio, esta deja los huevos en los nidos de otra especie la cual, si no puede reconocer que el huevo no es suyo, proveerá el cuidado parental necesario, empollará y criará al pajarito (hasta que eventualmente pueda reconocer que es un pájaro de otra especie). El *Tordo Pico Corto* parasita casi exclusivamente a la especie Tordo Músico (*Agelaioides badius*) [1].

En el estudio de esta especie y su comportamiento se encuentran la Dra. Romina Scardamaglia y el Dr. Juan Carlos Reboreda, del Departamento de Ecología, Genética y Evolución de la Universidad de Buenos Aires. En diversos papers, como [1] y [2], se estudian y describen ciertas características de esta y otra especie. En estos se utiliza radiotelemetría junto con el seguimiento de los investigadores para las observaciones de las actividades de los pájaros. La zona geográfica de estudio de estos pájaros es en la reserva *El Destino*, cerca de la ciudad de Magdalena, en la Provincia de Buenos Aires. En este área conviven el *Tordo Pico Corto*, el *Tordo Músico* y otras especies de pájaros y animales.

El *Tordo Pico Corto* suele agruparse en grandes grupos para pasar la noche en algún sitio, denominado dormidero. Los beneficios de estos dormideros pueden ser disminuir los riesgos de depredadores y mayor probabilidad de encontrar pareja. También se cree que estos dormideros pueden servir como centros de información, donde individuos que no tengan información sobre donde obtener comida, o sitios para parasitar, puedan obtenerla [2]. El *Tordo Pico Corto* muestra alta fidelidad con su respectivo dormidero. Su área de acción, es decir, el área donde se mueve para realizar las actividades de parasitismo, suele estar cercana a su dormidero.

Para poder estudiar a esta especie de pájaros, el grupo de investigación de biología capturó ejemplares utilizando trampas de tipo *walk-in* cebadas con mijo. Cada pájaro fue marcado con anillos de colores para poder distinguirlos, y con un pequeño transmisor de frecuencias de radio.

Los pájaros que vamos a analizar en este trabajo entraron en las trampas mencionadas de a pares (o de a tríos en algún caso). Estos pares de pájaros atrapados juntos son las

potenciales parejas que observaron los investigadores, las cuales se esperan que muestren altos niveles de asociación durante el día.

Complementando los radiotransmisores puestos en los pájaros, se colocaron en la reserva cuatro antenas para poder recibir las señales de los mismos.

Vamos a utilizar la información recolectada por estos radiotransmisores para estudiar el comportamiento de estos pájaros.

Fig. 1.1: Imagen del área de estudio junto con las cuatro antenas, generada utilizando Google Earth.

La característica parasitaria que presenta esta especie, algo poco común, plantea preguntas interesantes sobre el comportamiento de estos individuos. Este comportamiento requiere habilidades especiales para llevar a cabo el parasitismo, como por ejemplo, una buena memoria espacial para recordar dónde se encuentran los nidos de los hospedadores.

Sobre estas características particulares de esta especie, entre las preguntas que nos interesa contestar, se encuentra su comportamiento monógamo social. Hay evidencia, en base a observaciones y estudios previos, que sugiere que esta especie tiene un comportamiento monógamo social. Esto no es tan frecuente en pájaros que realizan parasitismo. La existencia de la monogamia social nos plantea otras preguntas, sobre el rol de macho en la pareja, y si colabora en las actividades de parasitismo.

Estas son algunas de las preguntas que nos interesa perseguir en este trabajo.

1.1. Hipótesis

Nos interesa analizar el comportamiento del Tordo Pico Corto. Contar con información sobre la ubicación de los pájaros durante el día nos permitiría analizar el comportamiento individual y grupal de los mismos. Estos análisis nos pueden ayudar a contrastar las observaciones hechas por los biólogos y contribuir a la confirmación o refutación de las hipótesis planteadas.

Con el objetivo de encontrar la posición en la que se encuentra un pájaro en un momento dado, vamos a utilizar técnicas de *machine learning*. De esta forma, vamos a poder utilizar la información de los radiotransmisores y las antenas para poder estimar, con cierto margen de error, la posición de los pájaros. Nuestro objetivo es, usando las herramientas que disponemos, que el error de nuestras predicciones sea el mínimo posible, y que nos permita luego realizar análisis de forma de llegar a resultados correctos. Es decir, buscamos tener un error que no tenga impacto negativo en los análisis posteriores.

La hipótesis principal que vamos a atacar en este trabajo, es si el Tordo Pico Corto tiene un comportamiento monógamo social. Para esto, una vez que tengamos por donde se movieron todos los pájaros durante un periodo de tiempo, analizaremos si podemos detectar parejas de pájaros, en las cuales coincidan sus patrones de movimiento. En otras palabras, vamos a decir que dos pájaros son pareja si pasan gran parte de su tiempo juntos.

La segunda pregunta que nos interesa responder es sobre el rol del macho en esta relación monógama social, en particular, si los machos acompañan a su pareja a realizar las actividades de parasitismo. Queremos saber si los machos salen del dormidero junto con su pareja, cerca de las 4am al parasitar, o si salen en su horario habitual, cerca de las 6am.

Finalmente la tercera hipótesis de nuestro interés, es entender la actividad de las parejas cerca de nidos activos, es decir, nidos a los cuales es posible parasitar. Buscamos entender si las parejas pasan más tiempo juntos cerca de estos nidos activos, para la puesta de los huevos, o si pasan más tiempo juntos, por ejemplo, investigando a qué nidos pueden parasitar. Esto nos ayudaría a entender mejor la participación del macho en las actividad de parasitismo.

2. DATOS

Tenemos el objetivo de poder determinar la ubicación de los pájaros en un momento dado, a partir de las señales emitidas por los radiotransmisores colocados en los mismos. Las antenas reciben las señales de estos transmisores junto con una intensidad o potencia, que determina la claridad con que llega la señal. Una potencia muy grande implica que la señal proviene de un lugar cercano a la antena. En cambio, una potencia muy chica ocurre cuando la señal llega muy débil a la antena, ya sea por la distancia entre ambos o por otros factores que interfieran con la señal.

Conocer la ubicación de los pájaros nos va a permitir realizar análisis respecto a qué tanto tiempo o qué tan cerca suelen estar unos pájaros con otros, entre otras cosas, para validar o refutar nuestras hipótesis. Podríamos también validar dónde es que duermen los pájaros, por dónde se mueven en el día y sacar nuevas ideas o hipótesis de cómo se comportan estos pájaros.

Al encarar esta problemática desde el lado de *machine learning* necesitamos datos (en particular, datos etiquetados), para poder entrenar una máquina de aprendizaje, es decir, un modelo, que sepa estimar la posición de un pájaro, a partir de las emisiones de su radiotransmisor. Los datos etiquetados, son datos que han sido "clasificados" de cierta manera, que nos permite utilizarlos para alimentar a nuestra máquina de aprendizaje para resolver la tarea deseada. Para nosotros, los datos van a corresponder a las señales recibidas por las antenas, y la "clasificación" de los mismos, va a ser una posición en el mapa.

2.1. Datos de calibración

Nos interesa tener, para cualquier punto *(x,y)* en el mapa, cuál es la potencia de la señal que recibirían las cuatro antenas si hubiera un pájaro en esa posición. De esta manera, al querer predecir la posición de un pájaro a partir de las recepciones de una emisión, de la forma $\langle s_1, s_2, s_3, s_4 \rangle$, podemos tomar alguno de los puntos en el mapa que tengan similares potencias.

Ese es el rol que cumplen los datos de calibración que tenemos. Son recepciones de las antenas de emisiones generadas manualmente en el campo, donde sabemos la posición desde donde fue hecha cada una. Nos van a permitir entrenar una máquina de aprendizaje que sepa usar las relaciones entre las intensidades de las señales recibidas y las posiciones para predecir, lo mejor posible, las posiciones de los pájaros.

Estos datos de calibración fueron obtenidos por la Dra. Scardamaglia, utilizando transmisores iguales a los colocados en los pájaros. Para emular una altura (fija) a la que puede estar un pájaro, se colocó el transmisor en la punta de un palo de cuatro metros. Parando en 598 puntos en el campo, con una distancia de 100 metros entre sí de manera uniforme, se utilizó un GPS profesional para obtener la posición exacta en la que estaba parado al momento de la emisión. Tenemos entonces una grilla de 23×26 puntos donde se obtuvieron datos. En realidad, no se realizaron emisiones en todos los puntos en la grilla, ya que en

algunos era evidente que no iba a llegar señal a las antenas, al estar muy lejos o en áreas con alta densidad de árboles.

Por cada punto, se obtuvieron dos minutos de emisiones, con una emisión cada 5 segundos (al igual que los pájaros), rotando cada 30 segundos el ángulo de emisión en 90 grados. Se generaron entonces 24 emisiones por punto. Notar que una emisión de la señal no necesariamente va a tener una recepción en alguna o todas las antenas; por ejemplo, si la emisión ocurre muy lejos de la posición de la antena.

Los datos de calibración se componen entonces por:

- Una lista de tuplas conteniendo la posición en el mapa, fecha y hora de inicio y de fin de cada calibración, junto con un identificador de transmisor. Por ejemplo:

  `<(463268.431054228,6111790.741639590), 27/01/2018,17:50:05,17:52:05,12>`
  `<(463353.767051577,6111842.545118690), 27/01/2018,17:45:05,17:47:05,12>`

- Las señales recibidas por cada antena, que contiene la fecha y hora, el identificador del transmisor y la potencia de la señal. Por ejemplo:

  `<02/06/2018,12:04:04,18,146>` `<25/02/2018,6:26:40,17,12>`

La posición en el mapa se encuentra en coordenadas UTM. Las potencias de las señales recibidas pueden tener un valor entre 1 y 255.

La cuarta antena, la antena *D2*, fue puesta en funcionamiento después de las primeras tres antenas. Entonces la recolección de los datos se realizó en dos etapas. En una primera etapa, se obtuvieron para todos los puntos, las señales recibidas por tres antenas. Luego, en una segunda etapa, se obtuvieron las señales de las cuatro antenas para todos los puntos. Para algunos puntos entonces podemos tener hasta 48 emisiones.

2.2. Análisis de los datos de calibración

Nos interesa comprender la particularidades de los datos con el fin de entender si es posible usarlos para llegar a un modelo de aprendizaje automático que pueda predecir la posición de los pájaros.

En un mundo ideal, a cada punto le correspondería una misma intensidad de señal en todo momento. Si yo estoy quieto parado en una posición emitiendo señales, esperaría que la intensidad de la señal recibida por las antenas no varíe. Dado un período de emisión de un punto de calibración, las antenas recibirían la misma potencia por cada ángulo de emisión, y tendrían algo de la forma `[40,...,40,52,...,52,57,...,57,49,...,49]`. En realidad, hay muchas cosas que pueden influir en la potencia que terminan recibiendo las antenas. Por ejemplo, la estructura de la vegetación que pueda interferir con la señal o la misma volatilidad propia de la señal de radio o de los equipos radiotransmisores.

Parte del análisis consiste en medir el ruido que agrega el medio a las recepciones, para validar cómo podría impactar esto en nuestros predictores. También nos interesa explorar los datos en busca de *insights* que nos permitan tener mejores modelos o entender por qué están o no andando de la manera esperada.

Tenemos cuatro antenas, ubicadas en lugar distintos del campo (ver figura 1.1), por lo tanto pueden tener comportamientos distintos. Vale la pena entonces, segregar los análisis

por antena. Transformando un poco los datos que tenemos, llegamos a obtener para cada período de dos minutos de emisión desde un punto, cuatro listas de la forma $[p_1, p_2, ..., p_{24}]$ donde cada lista representa las recepciones de una antena, y p_i es la potencia recibida por ella, o 0 en caso que la antena no haya recibido ninguna señal.

Anteriormente habíamos dicho que los transmisores emitían cada cinco segundos, y que emitíamos señales durante dos minutos para los datos de calibración, y por lo tanto tendríamos 24 emisiones totales. En realidad, los transmisores emiten cada 5 segundos aproximadamente, y no comenzamos a contar el tiempo exactamente desde la primer emisión. Por lo tanto, para algunos puntos podemos tener 23 o 25 emisiones.

Observemos cómo son las potencias de cada punto de calibración, que recibe cada antena. Al estar un punto cerca de una antena, lógicamente este va a tener una mayor potencia. En la figura 2.1 podemos ver el área de influencia de cada antena, es decir, hasta dónde llega a recibir las señales de los transmisores. Las X marcan la ubicación de las antenas. Para cada punto se toma el promedio de las intensidades de todas las emisiones recibidas por esa antena. Los puntos que se muestran con el color más claro, son los puntos que no tuvieron recepción de ninguna de las señales emitidas.

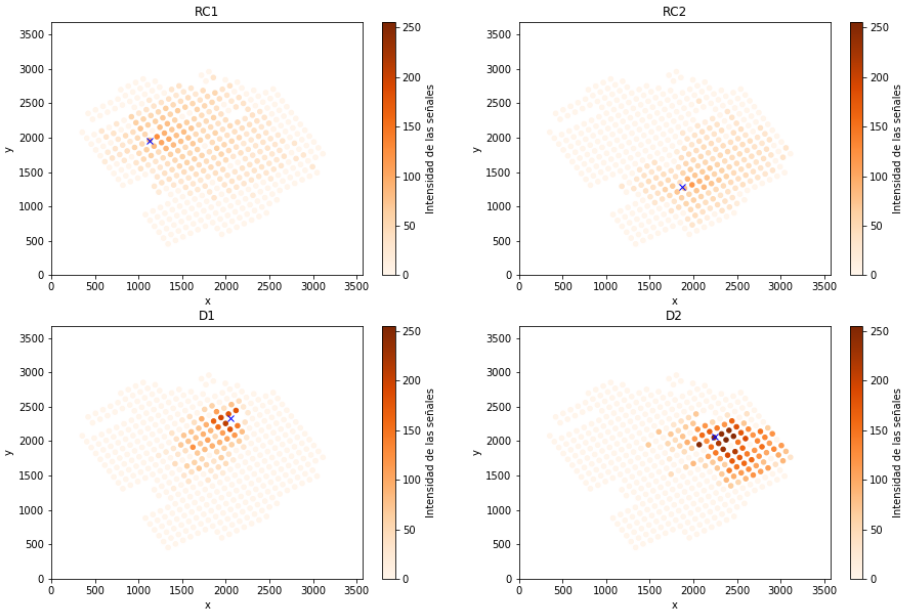

Fig. 2.1: Mapa de intensidades de las potencias recibidas por cada antena

Podemos ver que las antenas *RC1* y *RC2* son las que tienen mayor alcance; reciben señales de puntos más lejanos en el mapa. Esta diferencia se puede deber a las características propias del terreno. Áreas con mayor densidad de árboles, o con bajo relieve, son más propensos a sufrir interferencia en las señales y que no lleguen para puntos lejanos. Por otro lado, la antena *D2*, y luego la antena *D1*, son las que tienen mayor potencia para los puntos cercanos a ellas, llegando a tener recepciones con potencia cercana a 255 (máxima

potencia posible).

Esto nos indica que vamos a ser buenos para predecir puntos cercanos a las antenas y no tan buenos para puntos lejanos. Al mismo tiempo, en los puntos que podamos recibir las señales en las cuatro antenas, pensaría que vamos a ser mejores que en los puntos que usan menos antenas. Si la única antena con recepción de una señal es la antena *RC1*, con potencia P, van a existir varios puntos distintos que tengan una potencia igual o similar a P. Veamos entonces en la figura 2.2 cuántas antenas están al alcance de cada punto, donde una antena está al alcance de un punto si recibió la señal de alguna de las emisiones del punto.

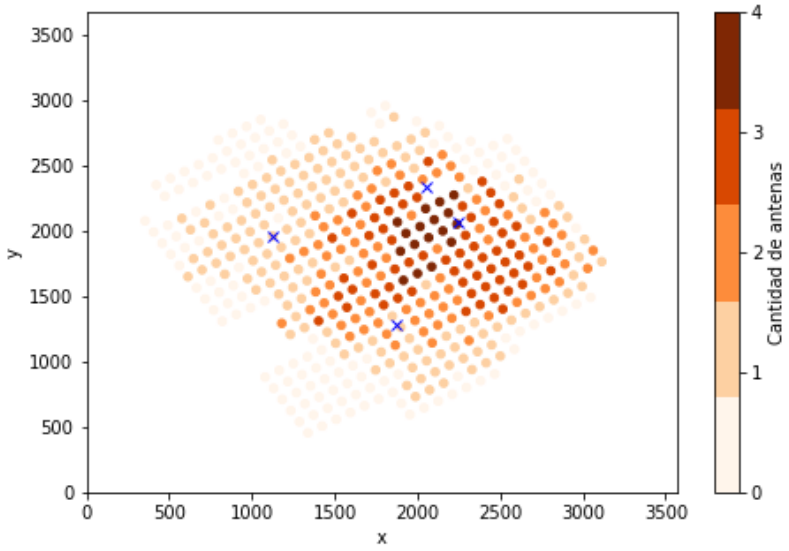

Fig. 2.2: Mapa de cantidad de antenas alcanzadas por cada punto

En una región importante del mapa, tenemos 2, 3 o 4 antenas para cada punto. Creemos que en esa región vamos a poder ser buenos prediciendo las posiciones. En cambio, en la región noroeste, al utilizar solamente la antena *RC1*, vamos a tener dificultades para predecir posiciones con cierta precisión. Pero lo que sí vamos a poder determinar, en ese caso, es que la señal provino de la región noroeste. Viendo esto, podríamos decidir más adelante, enfocarnos en solo una sub-región óptima del mapa.

En la figura 2.3 podemos ver cuántos puntos están al alcance de 0 a 4 antenas. Si bien la mayoría de los puntos están al alcance de cero o de una antena, la cantidad que tiene más de una antena al alcance no es despreciable, y es donde mejor vamos a poder predecir las posiciones.

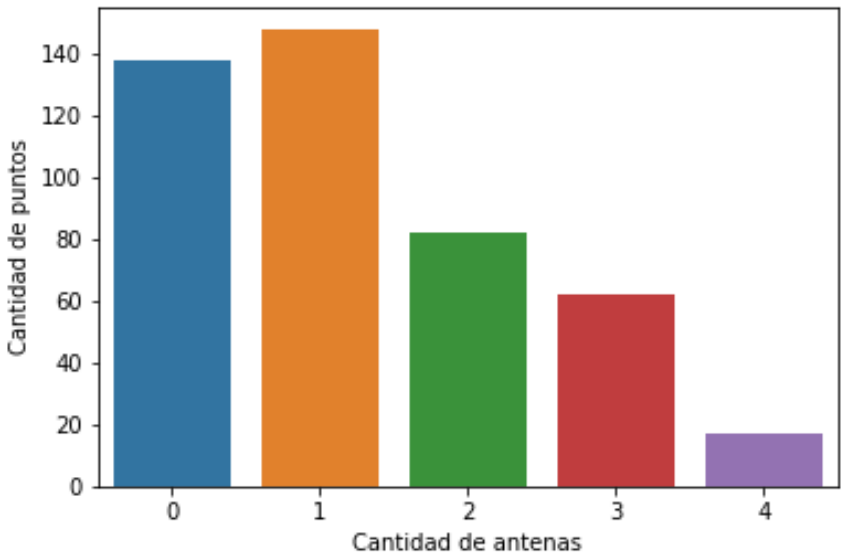

Fig. 2.3: Cantidad de puntos que usan n antenas

En un período de emisión de dos minutos desde un punto, durante el cual se realizan 24 emisiones, sabemos que no necesariamente van a llegar todas las señales a las antenas. La señal puede ser débil y sumada al ruido del medio puede no ser captada por la antena. Esperamos que, mientras mayor intensidad tengan las potencias recibidas para un punto, más probable es que estas señales lleguen a las antenas. Esto se puede ver en la figura 2.4 donde la cantidad de recepciones recibidas por punto aumenta al aumentar el promedio de la potencia de todas sus emisiones.

En la figura 2.5 se ve reflejada la relación distancia-potencia, donde, tal como se esperaría, al aumentar la distancia de un punto hacia la antena, disminuye el promedio de la potencia de todas las emisiones del punto.

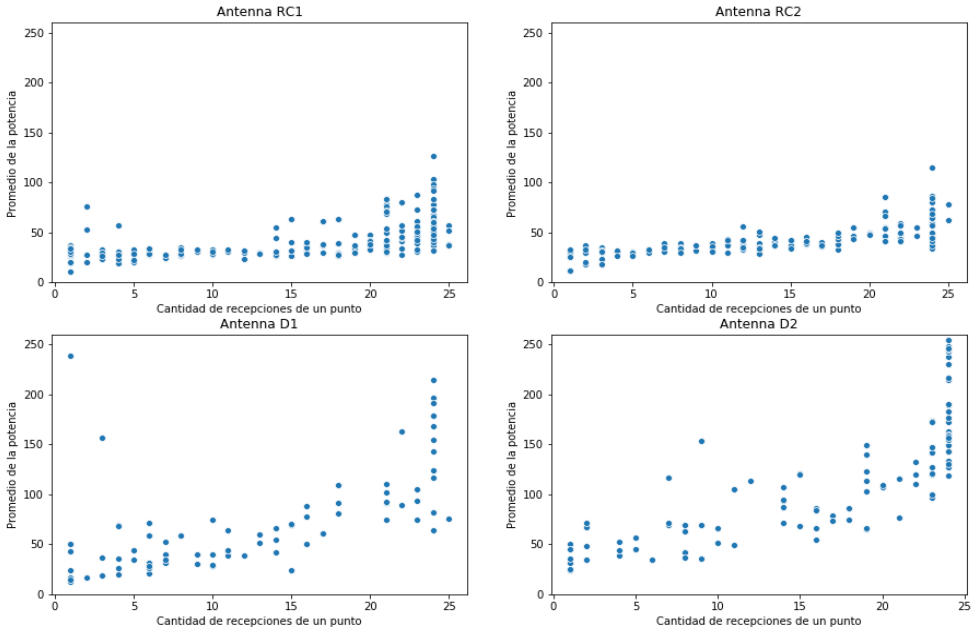

Fig. 2.4: Relación entre la cantidad de recepciones y la potencia recibida

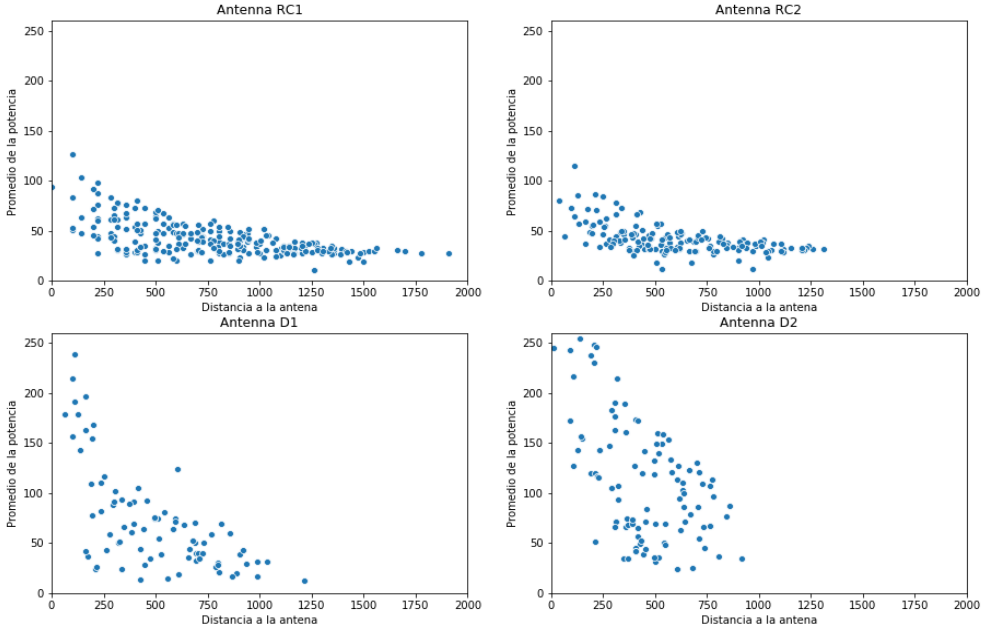

Fig. 2.5: Relación entre la distancia a la antena (en metros) y la potencia recibida

Para ver con más claridad las diferencias entre las antenas, la figura 2.6 muestra, para cada antena, la distribución de las potencias de las señales recibidas. Al igual que en los mapas en la figura 2.1, se puede observar que las antenas *RC1* y *RC2* reciben señales de menor potencia que las otras. Además, tienen una menor dispersión, ya que las potencias de las señales recibidas suelen estar cercanas entre sí. En cambio en las otras antenas la dispersión es mayor, porque hay muchos puntos cercanos a las antenas con potencias muy grandes, y los puntos más alejados con potencias leves.

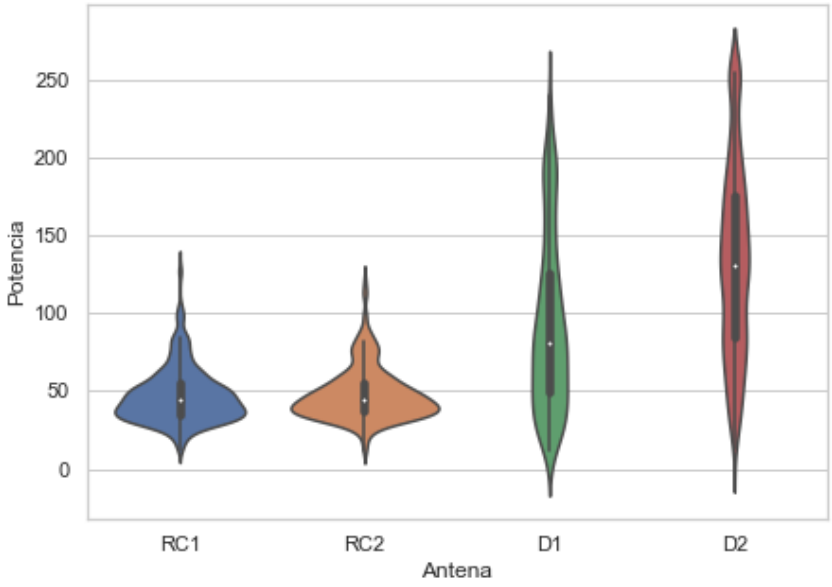

Fig. 2.6: Distribución de las potencias recibidas por cada antena

También nos interesa saber, para cada período de emisión de dos minutos, cuánto puede variar la potencia de las señales recibidas por ángulo, para cada antena. Como dijimos antes, en un mundo sin ruido, tendríamos una misma potencia por cada ángulo de emisión. Mientras más varíen las potencias que recibe una antena desde un mismo punto y ángulo, más difícil va a ser llegar a un buen modelo de aprendizaje automático, porque se dificulta más el mapeo entre potencia recibida y distancia a la antena. En la figura 2.7 vemos para cada antena, la distribución de los coeficientes de variación de las potencias recibidas de las emisiones de cada punto de calibración. El coeficiente de variación, definido como proporción de la desviación estándar sobre la media, nos permite describir la dispersión de las potencias independiente de la magnitud de las mismas.

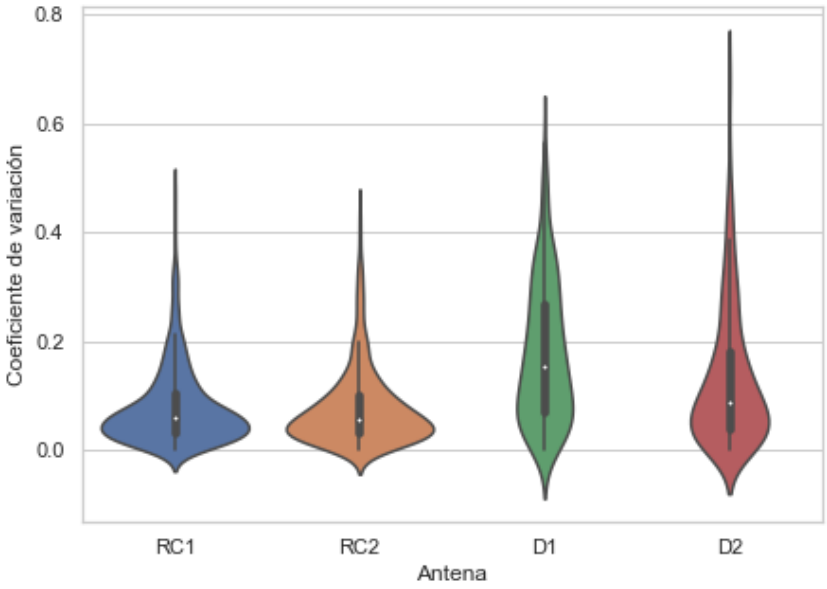

Fig. 2.7: Distribución de los coeficientes de variación de las potencias de cada punto recibidas por cada antena

2.3. Datos de los pájaros

En la sección anterior, hemos hablado de los datos de calibración que tenemos, y que utilizaremos para entrenar un modelo de aprendizaje automático. Realizamos análisis y validaciones de los mismos en busca de entender y encontrar particularidades en los datos que nos permitan crear modelos más efectivos.

Además, tenemos los datos recolectados de los pájaros. Al igual que los datos de calibración, son las emisiones hechas por los radiotransmisores, en este caso, colocados en los pájaros. Concretamente, lo que tenemos son las señales recibidas por las cuatro antenas, donde cada señal informa su intensidad y el identificador del dispositivo (o sea, del pájaro) que realizó la emisión.

De esta forma, nuestros datos consisten en un conjunto de recepciones de las antenas de la forma <02/06/2018,12:04:04,18,146> conformada por la fecha, la hora, el identificador del transmisor y la potencia de la señal.

Otra forma de ver los datos, que nos va a ser más útil para nuestro análisis y posterior uso, es pensarlos de la forma $< Fecha, Hora, Id, P_1, P_2, P_3, P_4 >$ donde P_i es la potencia de la señal recibida por la antena i en ese instante, con valor 0 si aquella antena no recibió ninguna señal.

A diferencia de los datos de calibración, en este caso no sabemos en qué lugar se encontraba el pájaro al momento de la emisión. Eso es precisamente lo que trataremos de estimar más adelante. Otra distinción que vale la pena mencionar, es que al ser emisiones obtenidas de chips colocados en los pájaros, estas emisiones pudieron haber sucedido mientras

el pájaro estaba en movimiento, o a una gran altura. Con lo cual, tenemos una dimensionalidad extra, ya que las potencias recibidas no dependen solamente de las posiciones x e y (coordenadas), sino que también dependen de z, la altura.

Recordemos que la metodología para la selección de pájaros a los cuales ponerles el transmisor consistió en poner trampas en el campo, y tomar los pájaros que entraran de a pares en las mismas, en distintos momentos. Por esto, para algunos pájaros tenemos más días con emisiones que otros. Tenemos la garantía de que desde el 10/01/2018 hasta el 25/01/2018 todos los pájaros tenían el transmisor puesto y funcionando. Vamos a excluir de nuestros análisis a dos pájaros, el macho 15 y la hembra 16, ya que cuentan con una cantidad insignificante de emisiones con recepciones de alguna antena (249 y 43 emisiones respectivamente durante estos 15 días). Estos valores resultan insuficientes para cualquier tipo de análisis a realizar.

En la figura 2.8 podemos ver la distribución de las recepciones de emisiones de los pájaros a lo largo de los días. Tomamos los 16 días para los cuales tenemos emisiones de todos los pájaros, y solamente el horario diurno, de 8 a 19hs, donde tenemos la confianza de que los pájaros no se encontraban en el dormidero. En la figura se muestra qué proporción de las emisiones totales tuvieron recepción de alguna antena para cada día.

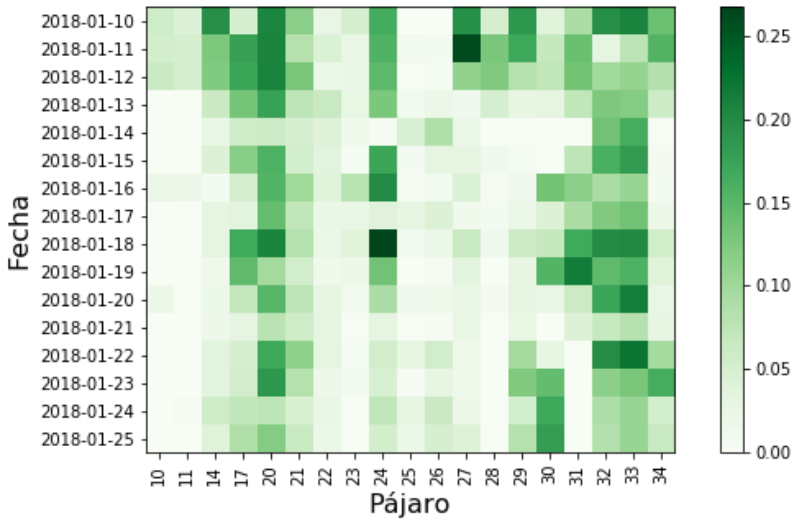

Fig. 2.8: Proporciones de los días donde tenemos emisiones sobre cada pájaro

Se pueden apreciar las diferencias entre los distintos pájaros y las frecuencias de las recepciones de sus emisiones. Cuál es el área de acción de cada pájaro (es decir, el área del campo por donde se suelen mover) puede impactar en la cantidad de recepciones que haya tenido. En los boxplots de la figura 2.9 se pueden apreciar estas diferencias de manera más precisa, mostrando mediana, percentiles y outliers.

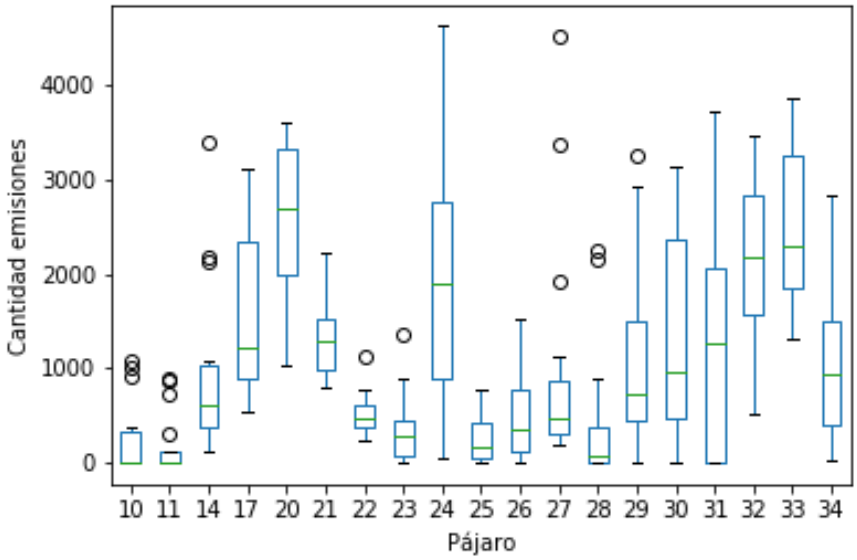

Fig. 2.9: Boxplots que muestran la distribución de la cantidad de emisiones de cada pájaro por día

2.4. Análisis de los datos de los pájaros

Al igual que con los datos de calibración, nos interesa entender las particularidades que puedan tener las emisiones de los pájaros. Estas emisiones de los pájaros son las que utilizaremos más adelante para predecir la posición en la que se encontraban los pájaros, utilizando un modelo de aprendizaje automático, entrenado con los datos de calibración. Entonces nos interesa que las emisiones de los pájaros sean lo más parecidas posibles a las emisiones hechas en la etapa de calibración. Si son diferentes, entonces las métricas de precisión (por ejemplo, qué tan lejos cae la predicción hecha respecto a la posición real) que saquemos del modelo no serán representativas (para nuestro caso de uso).

En las tablas 2.1 y 2.2 podemos ver la cantidad de recepciones que tuvo cada antena y la cantidad de emisiones de pájaros que tuvieron recepción de una, dos, tres o cuatro antenas respectivamente. Podemos apreciar que existen diferencias con respecto al mismo análisis en los datos de calibración. En las emisiones de los pájaros, hay una proporción mucho mayor, casi exclusiva, de emisiones con recepción de solo una antena. En cambio, en los datos de calibración, hay más emisiones que alcanzan más de una antena. Esto podría afectar a nuestros modelos, ya que aprenderían a predecir posiciones utilizando emisiones con recepción de más de una antena, pero al querer usar el modelo, utilizaríamos mayormente emisiones con recepciones de solo una antena.

Antena	Cantidad de recepciones
RC1	144142
RC2	158994
D1	57062
D2	29993

Tab. 2.1: Cantidad de recepciones de señales de cada antena

Cantidad de antenas	Cantidad de recepciones
1	296303
2	38436
3	5204
4	351

Tab. 2.2: Cantidad de emisiones que utilizan cada cantidad de antenas

3. TÉCNICAS DE ML PARA GEOLOCALIZACION DE PÁJAROS

El aprendizaje automático, o *machine learning* es un área de las ciencias de la computación que estudia algoritmos que mejoran su performance gracias a la experiencia. Un programa de computadora aprende de la experiencia E respecto a una clase de tareas T y medida de performance P si su performance en tareas de T, medidas por P, mejoran con la experiencia E [3]. Dentro del área de *machine learning* tenemos el aprendizaje supervisado. Es la tarea de aprender una función que mapea una entrada a una salida basado en ejemplos de pares entrada-salida (datos etiquetados). Es el método mas común en problemas de clasificación, regresión y ranking.

Un feature es una variable de entrada que representa una observación de un hecho o fenómeno. Un problema de *machine learning* puede usar uno o más features $x_1, x_2, ..., x_n$. Por ejemplo, en un sistema de detección de spam, los features pueden ser palabras en el texto del mail, ubicación del remitente, hora del envío del mail, etc. Un label o etiqueta, es el valor que nos interesa predecir. Por ejemplo si un mail es spam o no, o el precio de un inmueble.

Un modelo define una relación, es decir, una función de features a labels. Entrenar un modelo significa mostrarle ejemplos y que gradualmente "aprenda" la relación entre feature y label. Inferir significa aplicar el modelo entrenado a nuevos ejemplos no conocidos, usar el modelo para realizar predicciones.

Comúnmente, al trabajar con modelos de aprendizaje automático supervisado, se cuenta con un conjunto de datos que se utilizan para entrenar y evaluar el o los modelos. Este conjunto de datos es dividido en, al menos, dos partes, una que se usará para entrenar el modelo y otra para evaluar el mismo. Esto nos permite entender qué tan bien cumple el modelo con la tarea buscada y comparar su performance contra otros modelos. Existen diferentes estrategias sobre cómo separar el conjunto inicial en uno de entrenamiento y uno de validación, con el objetivo de que las evaluaciones sean lo más justas posibles, y evitar llegar a conclusiones erróneas sobre la performance de los modelos.

3.1. Técnicas de ML para geolocalizacion de pájaros

Tenemos el objetivo de encontrar un modelo que nos sepa predecir, dadas las recepciones de las antenas de las señales emitidas por los radiotransmisores, la posición en que se encuentra un pájaro en un momento dado.

Comenzamos la búsqueda de este modelo partiendo de un baseline simple que iremos iterando y complejizando a medida que veamos que sea necesario. Vamos a crear un modelo que para predecir tome como entrada una tupla $\langle P_1, P_2, P_3, P_4 \rangle$ donde P_i es la potencia de la señal recibida por la antena i, con valor 0 si no recibió ninguna señal. El objetivo es que el modelo prediga una posición (x, y) en el mapa, tal que en esa posición un transmisor tenga recepciones de las antenas con valores iguales (o muy similares) a los pasados como entrada.

3.1.1. Algoritmos de regresión

K nearest neighbor:

K vecinos más cercanos (*KNN*) es un algoritmo de aprendizaje automático que consiste en utilizar, para estimar el valor o label (en nuestro caso una posición en el mapa) de una nueva instancia, las instancias similares a ella. La similitud entre instancias se define como la inversa de la distancia entre los features de las mismas, para alguna definición de la función distancia. En *KNN*, se toman las k instancias más similares, y a partir de los valores de estos, se calcula la predicción final. Una función de distancia típica es la distancia euclidiana, definida como

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{3.1}$$

para dos puntos $q = (q_1, ..., q_n)$ y $p = (p_1, ..., p_n)$. Luego de obtener los k puntos a menor distancia, se realiza un promedio o promedio ponderado de los labels de los mismos para determinar la posición de la instancia consultada.

Redes neuronales artificiales:

Las redes neuronales artificiales son otro algoritmo de aprendizaje automático supervisado, que aprende a partir de ejemplos de datos etiquetados. Basadas o inspiradas en el cerebro humano como lo indica el nombre, buscan aprender una función f que mapea features a labels, a partir de los ejemplos, de forma que esa función minimice cierta métrica o función de pérdida determinada.

El cerebro humano esta compuesto por miles de millones de neuronas. Las neuronas pueden activarse cuando reciben señales de otras neuronas conectadas a ella, dependiendo de las intensidades de estas señales y la potencia de las conexiones. Las redes neuronales artificiales imitan este comportamiento de las neuronas biológicas, utilizando conceptos matemáticos.

En las redes neuronales artificiales, una neurona es una función matemática, que recibe una o más entradas, provenientes de otras neuronas, y devuelve una salida, el resultado del computo de sus entradas. Las conexiones en las redes neuronales artificiales tienen un peso que indican cuánto influye la activación de una neurona en la otra.

Una red neuronal artificial se compone de un conjunto de "neuronas" que están conectadas entre si, permitiendo el envío de señales entre ellas. A partir de ejemplos, durante el entrenamiento se van fortaleciendo las conexiones entre las neuronas que ayudan a la tarea buscada (minimizar cierta función de costo), y reduciendo las que perjudican a la misma.

En nuestro caso, la función de costo que queremos minimizar es el error entre la predicción y la posición real de los pájaros. A diferencia de otros modelos, las redes neuronales nos van a permitir correlacionar las coordenadas x e y, ya que los pesos de las conexiones que se computan, se utilizan para predecir en simultáneo el eje x y el eje y.

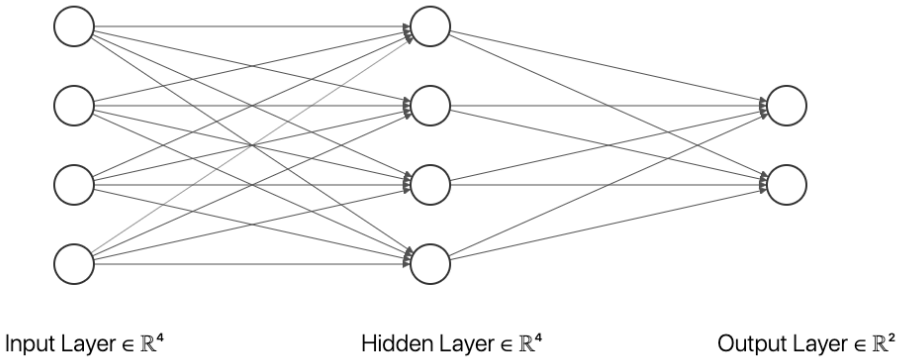

Fig. 3.1: Ejemplo de red neuronal utilizada

Decision Tree:

Decision Tree, o árbol de decisión, es un algoritmo de aprendizaje automático que se caracteriza por predecir a partir de reglas simples, inferidas a partir de los datos de entrenamiento. Cada nodo interno del árbol corresponde a una validación sobre algún feature y las hojas del árbol corresponden a un label o valor, si se trata de un problema de clasificación o regresión respectivamente. Durante el entrenamiento de los árboles de decisión, se decide qué condiciones y features se utilizarán en cada nodo para particionar los datos.

Una de las ventajas más importantes de este algoritmo, es que los modelos creados son simples de entender e interpretar.

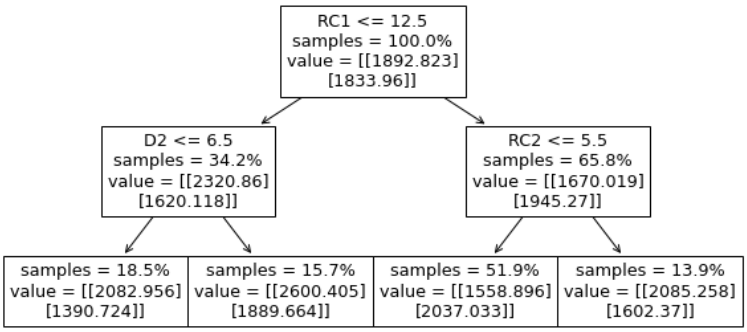

Fig. 3.2: Ejemplo de árbol de decisión

Gradient Boosting:

Gradient boosting (*GB*) es un algoritmo de aprendizaje automático basado en el ensamble de modelos. A partir de la combinación de muchos modelos débiles, se busca obtener un modelo robusto. Los ensambles son construidos a partir de modelos de árboles de decisión. En cada iteración se va agregando de a un árbol, de manera que este ayude a corregir los errores en las predicciones de los modelos anteriores. Los modelos son entrenados utilizan-

do una función de costo diferenciable a elección, y utilizando un algoritmo de gradiente descendente, minimizando el error.

3.1.2. Algoritmos de clasificación

En las siguientes secciones, vamos a utilizar los modelos obtenidos en esta parte, para realizar distintos análisis sobre las emisiones de los pájaros que tenemos y tratar de validar o refutar nuestras hipótesis.

Idealmente, el modelo de regresión nos permitiría trazar la trayectoria de los pájaros en cada instante. Los primeros análisis sobre estos modelos nos dan evidencia de que esto no es posible, tanto por la (no tan buena) calidad de los datos de calibración, como por la calidad de las emisiones de los pájaros.

Para algunos de estos análisis no es necesario conocer la posición exacta de los pájaros, ya que quizás nos alcance con saber si dos pájaros están o no cerca.

En ese caso, nos puede bastar con dividir el campo en regiones, y crear un modelo que nos sepa predecir en qué región se encuentra un pájaro en un momento dado. Este modelo va a ser un modelo de clasificación, el cual creemos que puede tener mejores resultados, ya que podría generalizar mejor.

Además, mientras más grandes sean las subregiones, y por lo tanto haya menor cantidad, más probable es que predigamos bien a qué región pertenece una emisión. Entonces tenemos una forma de aumentar la precisión de nuestro modelo, aunque perdiendo información que nos aporta el mismo.

Si partimos el campo en dos regiones de igual tamaño y nuestro modelo consiste en simplemente devolver una de las regiones al azar, cada una con probabilidad 0,5, entonces el *accuracy* (es decir, la proporción de predicciones correctas sobre el total de predicciones) sería 0,5. En cambio, si dividimos el campo en tres regiones iguales, entonces el *accuracy* sería 0,333.

Al dividir el campo en más regiones, estamos disminuyendo la precisión. Pero, por otro lado, ganamos información. Si el campo tiene tamaño 3000x3000 metros, al dividir en dos regiones (de 1500x3000 metros cada una), si dos pájaros están en la misma región, sabemos que están a una distancia de a lo sumo $\sqrt{1500^2 + 3000^2} = 3354{,}1$ si ambos pájaros se encuentran en extremos opuestos. En cambio, si está dividido en tres regiones (de 1000x3000 metros cada una), la distancia máxima posible entre los pájaros en una misma región es de $\sqrt{1000^2 + 3000^2} = 3162{,}3$ metros.

3.1.3. Algoritmos de clasificación + regresión

Generalmente, los algoritmos de clasificación no solo devuelven la clase predicha para una instancia, sino que también son capaces de devolver la probabilidad de pertenencia a esa clase. Y, adicionalmente, pueden devolver la probabilidad de pertenencia de todas las clases.

Volviendo a nuestro caso de uso. Supongamos que tenemos dos áreas A y B limítrofes, y el clasificador predice que el pájaro se encuentra en A con probabilidad 0,55, en B con probabilidad 0,45 y el resto de las regiones con probabilidad nula. Intuitivamente

creería que el pájaro se encuentra dentro de A, pero muy cerca del área B. Siguiendo este razonamiento, proponemos un modelo de regresión que utilice las probabilidades de cada área que predice el modelo de clasificación.

Una ventaja de este tipo de modelos, es que podemos aproximar o acotar el error. Si el clasificador tiene un *accuracy* de $x\,\%$ y la distancia máxima posible entre dos puntos de una misma región es y, entonces vamos a tener un error menor a y en el $x\,\%$ de los puntos. Sabiendo del trade-off mencionado previamente entre *accuracy* y distancia máxima entre cada par de puntos, podemos tomar decisiones sobre cuál es el error deseado para cierto porcentaje de los puntos.

3.1.4. Series temporales

Como mencionamos anteriormente, vamos a utilizar modelos de machine learning para predecir las posiciones en las que se encontraban los pájaros, a partir de los registros de las antenas. Este enfoque tiene una limitación importante: no saca provecho de la temporalidad de los datos de los pájaros.

Supongamos que conocemos con cierta confianza la ubicación de un pájaro en un momento dado. En el siguiente instante, es decir, en la siguiente emisión hecha cinco segundos después, sabemos que el pájaro se va a encontrar cerca de esa posición. La ubicación de un pájaro en un momento está relacionada con la posición en la que se encontraba previamente y en la que se encontrará posteriormente.

Este enfoque de utilizar la posición anterior en la que creíamos que estaba el pájaro, o utilizar las emisiones previas (y/o posteriores) podría ser de mucha ayuda para tener mejor precisión. Especialmente, nos serviría para detectar saltos anómalos en el recorrido de los pájaros (por ejemplo, si un pájaro se mueve 1000 metros en 5 segundos).

Lamentablemente, tenemos ciertas limitaciones para aplicar este enfoque. Los datos de calibración que poseemos para entrenar los modelos no corresponden a vuelos de pájaros, no son generados a partir del movimiento de un radiotransmisor. No podemos evaluar modelos que usan la temporalidad de los datos con los datos de calibración que tenemos. Utilizar modelos que no han sido evaluados podría ser contraproducente, no tenemos forma de saber si los cambios generan mejoras en la performance o no. Decidimos por esto no utilizar este enfoque, y usar los modelos más simples que no usen la temporalidad de los datos.

3.2. Evaluación

Una parte importante y fundamental del proceso de trabajo de cualquier proyecto relacionado al aprendizaje automático, es cuantificar los resultados de las soluciones. Es importante establecer un criterio para evaluar los distintos modelos y soluciones posibles, de forma de poder notar cuándo uno nos va a servir más que otro para nuestro caso de uso. Las métricas que obtenemos al evaluar un modelo deberían corresponder con el uso que luego le vamos a dar al mismo.

Lo que buscamos en este trabajo es que nuestra predicción se encuentre los más cerca posible de la posición real del radiotransmisor. Un modelo va a ser mejor que otro si sus

predicciones son más cercanas a la posición real. Para cada punto en nuestro conjunto de validación vamos a tener un error d que representa la distancia en metros entre nuestra predicción y la posición real del radiotransmisor. Vamos a medir el error medio absoluto (MAE), definido como:

$$MAE = \frac{\sum_{i=1}^{n} d_i}{n} \tag{3.2}$$

donde n es la cantidad de puntos en el conjunto de validación. Esta métrica nos da una buena intuición de qué tan lejos caen nuestras estimaciones de la posición real.

Si evaluamos muchos modelos o variaciones de los modelos y nos quedamos con el que tenga mejor performance, podría pasar que el modelo que elijamos sea muy bueno solamente para los datos en los que estamos evaluando. En este caso, estaríamos sobre-estimando la performance de nuestro modelo.

Si utilizáramos solo dos conjuntos, uno de entrenamiento y otro de validación, podríamos caer en la situación de que el conjunto de validación, al ser chico, por más que sea obtenido de manera aleatoria, tenga algún sesgo. Entonces las métricas que obtengamos de la evaluación podrían ser mejores o peores de lo real o esperado.

K-Fold cross-validation minimiza el impacto que esto puede tener, ya que se realizan múltiples evaluaciones y después se ponderan los mismos. Esta metodología es utilizada con frecuencia en el área de *machine learning* ya que permite obtener estimaciones menos sesgadas, es decir, más precisas sobre la performance de los modelos.

Utilizamos *K-Fold cross-validation* como metodología de evaluación de los modelos. Vamos a particionar nuestros datos en k conjuntos. Cada dato es una emisión desde un punto en el mapa, junto con la información de qué intensidades recibió cada antena. Omitimos las emisiones que no tuvieron recepción de ninguna antena ya que no aportan información (de nuestro interés). Cada conjunto es utilizado una sola vez como conjunto de validación, mientras los otros k-1 se utilizan como datos de entrenamiento.

Antes de comenzar las evaluaciones de los modelos, lo primero que hacemos es separar una parte de los datos para utilizarlos como *hold-out* y el resto para entrenar y evaluar los modelos. Una vez que tengamos nuestro modelo ganador, mediremos su performance final contra los datos de *hold-out*, esperando que arroje resultados similares que al utilizar el otro conjunto de datos.

3.3. Resultados

Para cada algoritmo de interés, realizamos las evaluaciones de la manera mencionada previamente y obtuvimos los resultados. Para cada método o algoritmo realizamos un *grid search* para obtener la mejor combinación de hiperparámetros, que nos de la mejor performance posible. En la tabla 3.1 se pueden ver los resultados.

	MAE training	MAE test
KNN	307,28	321,96
DecisionTree	228,35	336,34
GB	295,58	316,49
NN	328,61	331,44
KNNc+GB	306,08	330,71

Tab. 3.1: Métricas obtenidas de algoritmos

En la figura 3.3 se pueden ver algunos ejemplos de las predicciones del modelo *KNN*. En azul se encuentran las posiciones reales de las emisiones y en rojo las predicciones realizadas. Los puntos en verde son predicciones exactas, es decir, emisiones donde la posición real y la predicción fueron iguales. Las lineas marcan la correspondencia entre la posición real y la predicción

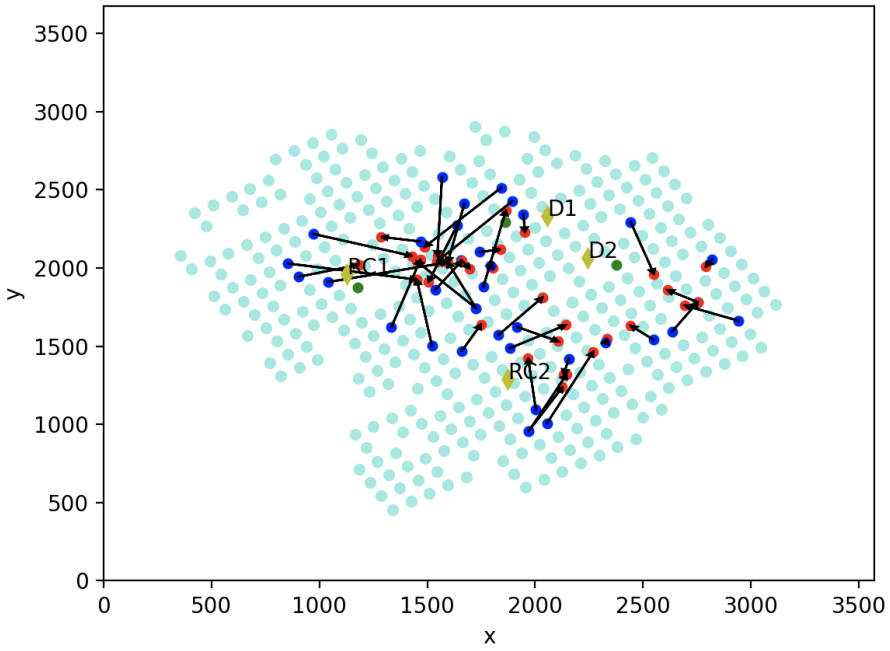

Fig. 3.3: Ejemplo de algunas predicciones con el modelo KNN

Los mejores modelos encontrados fueron *KNN*, *GradientBoosting* y Redes Neuronales. Decidimos que nuestro modelo campeón, el que usaremos para predecir las posiciones de los pájaros, es **KNN**. Por un lado, logra resultados sobre los datos de test comparables a los otros modelos. Además, al ser un modelo más simple que los otros, nos permite tener una mayor interpretabilidad de las predicciones, lo cual podría servirnos para tomar algunas decisiones que mejoren la utilidad del modelo. Con *KNN* podemos saber qué k emisiones del conjunto de datos de entrenamiento se utilizaron para llegar a una predicción. Si para una predicción se utilizan emisiones de múltiples puntos lejanos entre sí, podríamos decidir que no es confiable y omitir tal predicción.

3.4. Cota inferior del error

Sabemos que la señal que emiten los radiotransmisores y las recepciones de las antenas son propensas al ruido y a tener mucha variabilidad. Para las emisiones de un mismo punto de calibración es normal tener recepciones con distintas intensidades. Y por esto mismo es probable que no podamos determinar el origen de una emisión solamente viendo las intensidades recibidas, ya que van a haber varios puntos que hayan tenido alguna emisión con esas intensidades.

Si no existiera esta superposición de las intensidades, entonces un modelo que haga *overfitting* sobre los datos de entrenamiento al evaluarlo sobre los datos de entrenamiento, tendría precisión perfecta, es decir, un error de cero metros. Pero mientras mayor superposición exista, mayor va a ser este error.

Una aproximación a la cota inferior de MAE que pueda tener cualquier modelo que surja de estos datos de calibración se puede calcular viendo qué tan grande es este solapamiento de las intensidades. Por ejemplo, si tenemos dos puntos a y b a una distancia de 100 metros, y ambos tienen alguna emisión de la forma $S_a = S_b = \langle 0, 0, 50, 70 \rangle$, entonces si cuando evaluamos la emisión S_a tenemos un error de cero metros (es decir, predecimos esta emisión en la posición del punto a), al evaluar S_b tendremos un error de 100 metros. Por ejemplo, en la figura 3.2 se puede ver varios puntos de calibración que tuvieron alguna vez una emisión de la forma $\langle 0, 50, 0, 0 \rangle$.

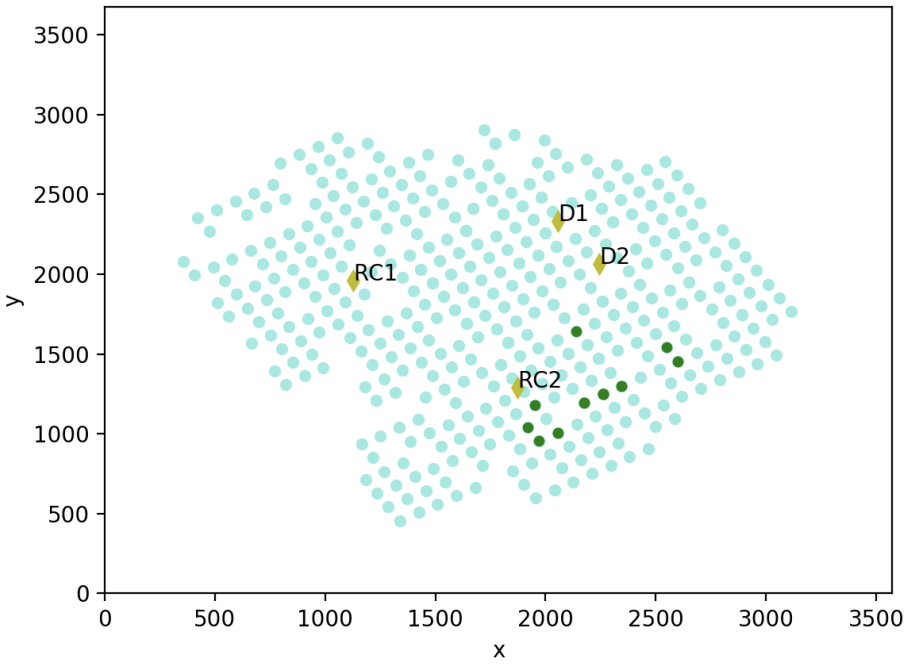

Fig. 3.4: Puntos de calibración que tuvieron emisiones con la misma intensidad

Para aproximar una cota inferior del error, seguimos el siguiente procedimiento:

1. Agrupar las emisiones de todos los puntos de calibración que tuvieron las mismas

intensidades.

2. Encontrar cuál sería la posición óptima a utilizar, es decir, el punto en el mapa tal que la distancia a todos los puntos de calibración es mínima. Esto se conoce como la *Mediana Geométrica*. Encontrar una solución exacta es un problema muy complejo, por lo tanto usamos un algoritmo aproximado. Utilizamos un algoritmo iterativo, que converge monótonamente en la mediana geométrica, descrito en [4].

3. Tomar el promedio de las distancias entre los puntos óptimos y los puntos de calibración.

De esta forma encontramos que el error mínimo que puede tener cualquier modelo es, aproximadamente, **226,89 metros**. Conociendo esto, un error de 321 metros en nuestros modelos parece razonable, siempre y cuando el modelo no esté realizando *overfitting*, sino que esté generalizando bien.

3.5. Limitaciones con los datos de calibración

Comenzamos la tesis hablando del objetivo principal de la misma, que es poder utilizar los datos de emisiones que tenemos sobre los pájaros para validar hipótesis y demás sobre esta especie de pájaros. Pero cuando generamos el modelo predictor a partir de las emisiones de los puntos de calibración, y al evaluar con los mismos, estamos trabajando con datos que no pertenecen a los pájaros.

Si bien estos datos fueron obtenidos usando los mismos radiotransmisores que los utilizados en los pájaros, existen muchas diferencias entre la emisiones de ambos. Los pájaros pueden estar a mayor o menor altura, podrían estar en el piso o en árboles donde la señal podría ser más propensa a interferencia. Los pájaros también se mueven, las emisiones podrían ser en momentos donde el pájaro está en movimiento, o en regiones en las que no tenemos puntos de calibración.

Más importante aún, los pájaros podrían localizarse en cierta región en la mayoría de su tiempo. Nuestros puntos de calibración están distribuidos uniformemente, y los evaluamos de esa manera. Es decir, los evaluamos como si los pájaros también fueran a moverse uniformemente por el campo, lo cual podría no ser completamente cierto.

Cuando utilicemos el modelo con las emisiones de los pájaros, no vamos a tener la performance encontrada y explicada en la sección anterior. Vamos a estar, en cierto sentido, a "ciegas", ya que estaremos utilizando un modelo que no sabemos cómo se desempeña al utilizar estas emisiones de los pájaros.

Para mitigar un poco la incertidumbre que genera utilizar un modelo que no esta correctamente validado, realizamos otras validaciones. Tenemos, gracias a la observación del equipo de biólogos que trabajó en el campo, algunos momentos donde sabemos la ubicación de algún pájaro, con un margen de error de algunos metros. Podemos evaluar entonces qué predice nuestro modelo en los instantes cercanos a esos momentos, y medir el error.

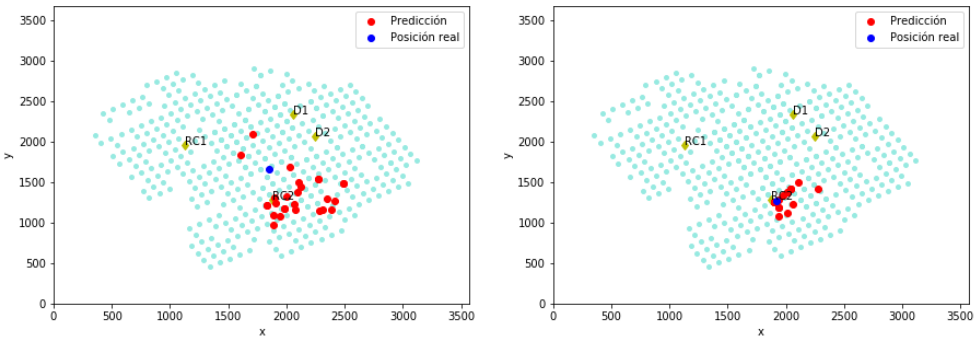

Fig. 3.5: Dos ejemplo de las predicciones en momentos cercanos a la observación de un pájaro

De los 48 puntos de validación que tenemos, encontramos emisiones en un rango de un minuto del momento de la observación para 16 puntos. En estos 16 puntos obtuvimos un error medio absoluto de **487,66 metros**. Si bien este error puede no ser representativo sobre el error que tendríamos al evaluar sobre el vuelo de los pájaros, resulta un valor aceptable y no tan alejado, 166 metros (un 50 % mas), de lo encontrado al evaluar los modelos con los datos de calibración.

4. VALIDACIÓN DE HIPÓTESIS

En esta sección procedemos a responder las preguntas planteadas en la introducción. Proponemos distintos experimentos que nos permitan encontrar evidencia que nos ayuden a responder las preguntas.

4.1. Hipótesis monogamia social

La hipótesis principal que buscamos estudiar en este trabajo es, si la especie de pájaros *Tordo Pico Corto* tiene un comportamiento monógamo social. A partir de distintos puntos de análisis vamos a ver si existe evidencia que sugiera si la hipótesis es verdadera o falsa.

Utilizaremos las señales que tenemos de los pájaros en un período de 16 días, desde el 2018-01-10 hasta el 2018-01-25 inclusive. Si bien tenemos más días con datos de pájaros, solamente en este rango de fechas tenemos garantía de que todos los pájaros estaban en el campo con los radiotransmisores colocados. Sin hacer esto, es probable que las parejas que tienen por ejemplo 30 días de emisiones en común tengan más co-ocurrencias, es decir, estuvieron más veces juntos, que las parejas que tengan solo 15 días de emisiones en común.

Tomamos para analizar solamente el horario diurno, en el cual los pájaros suelen moverse por el campo. Durante la noche, estos pájaros frecuentemente duermen juntos en el dormidero, por lo que no tiene sentido medir co-ocurrencia allí, ya que cualquier par de pájaros pueden estar juntos o cercanos entre sí.

Previo a comenzar nuestros análisis, el equipo de biólogos tenía como hipótesis la existencia de diez potenciales parejas de pájaros, basándose en el hecho de que los pájaros fueron atrapados de a pares en las trampas, sumado a observaciones propias en la reserva. Para agregar un paso más de validación de nuestro estudio, **decidimos no conocer cuáles eran estas parejas candidatas hasta terminar los análisis propios**, de manera de evitar sesgar nuestro trabajo y luego poder verificar que nuestros resultados se asemejen a las observaciones de los biólogos.

4.1.1. Co-ocurrencia de las parejas

Que exista monogamia social implica que se formen parejas de pájaros. Estas parejas deberían pasar mucho tiempo juntas. En este eje de investigación vamos a analizar cuánto tiempo pasan juntos los pájaros, y si hay parejas que pasan mucho más tiempo juntas que con cualquier otro pájaro.

Para analizar la cercanía en la que se encuentran los pájaros, utilizaremos el modelo de aprendizaje automático descripto en la sección anterior, que nos dará una estimación de las posiciones de los pájaros. Como vimos en las secciones anteriores, donde hablamos de la performance de los modelos, sabemos que no tenemos la precisión necesaria como para determinar cuándo dos pájaros están uno al lado, o a pocos metros, del otro. El error estimado que vamos a tener en las predicciones va a ser mayor a 300 metros, en promedio.

Que el modelo tenga un error semejante no es un impedimento para poder utilizarlo para nuestros análisis. Conociendo el error esperado, no buscaremos determinar si dos pájaros se encuentran uno al lado del otro. Buscaremos en cambio determinar, de manera más general, si se encuentran en una misma región. De esta forma, el modelo nos va a ser útil para determinar en qué región se encuentra cada pájaro, sin la necesidad de tener una muy alta precisión.

Definimos la co-ocurrencia de dos pájaros como el hecho de estar en la misma región en el mismo tiempo. Como las emisiones de las señales ocurren cada 5 segundos, entonces "en el mismo tiempo"quiere decir "con menos de 5 segundos de diferencia". ¿Qué es una región? Si el modelo que predice las posiciones de los pájaros a partir de las señales recibidas fuera exacto, entonces podríamos decir que dos pájaros están juntos si están a menos de 10 o 20 metros de distancia. Como sabemos que nuestro modelo no tiene tan buena performance, debemos ser menos restrictivos y utilizar regiones más grandes. Definimos distintas posibles regiones para utilizar en los análisis:

- Una única región comprendida por el área dentro de las cuatro antenas: recordemos que las antenas fueron puestas estratégicamente conociendo cuál es la región del campo donde se cree que mayormente suelen realizar las actividades de parasitismo.
- Regiones con un radio de X metros centrados en las antenas: cerca de las antenas es donde mejor podemos predecir las posiciones, dado que las intensidades de las señales son más fuertes cerca de las antenas
- Regiones regulares de $N \times M$: nos permite tener una granularidad mayor que las regiones anteriores.

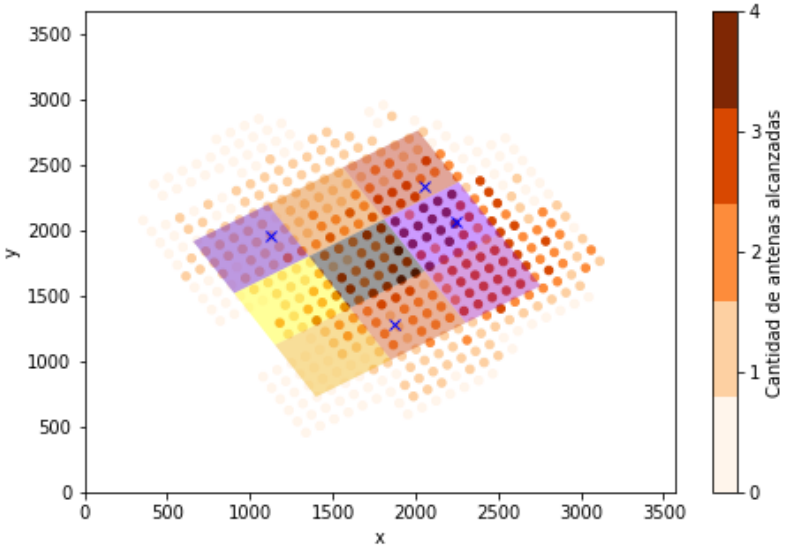

Fig. 4.1: Regiones regulares de $N \times M$

Decidimos utilizar regiones regulares de 300x300 metros como se ilustra en la figura

4.1. La elección de estas regiones regulares nos permite que cada antena caiga en una sola región.

Vamos a contar entonces, para cada pareja potencial de pájaros, cuantas veces co-ocurren. Utilizaremos como dijimos previamente las señales que tenemos de los pájaros sobre 16 días. Las co-ocurrencias observadas se muestran en las figuras 4.2 y 4.3 mediante una escala cromática que asigna colores claros a los valores más elevados. En la primer figura se muestra para cada par de pájaros, es decir, en cada posición de la matriz, cuantas veces co-ocurrieron. En la segunda figura se muestra qué proporción de las ocurrencias de las hembras fueron con cada macho. Por ejemplo, la hembra 21 el 60,6 % de las veces que estuvo con otro pájaro, fue con el macho 20.

Fig. 4.2: Co-ocurrencias entre cada pareja posible de pájaros

Fig. 4.3: Distribución del tiempo de co-ocurrencia de las hembras

A simple vista, podemos destacar a las parejas 20-21, 33-32, 34-29 y 30-31 que tienen valores de co-ocurrencia muy grandes en comparación con otros pájaros. Adicionalmente en las parejas 23-24 y 25-26, si bien no tienen valores muy altos de co-ocurrencia, podemos ver que las hembras tienen una proporción alta de sus co-ocurrencias con el macho respectivo.

Vemos que hay pares de pájaros que se destacan sobre los demás, que tiene valores altos de co-ocurrencia, es decir, estimamos que en muchos momentos se encuentran en puntos del campo cercanos. Lo que no sabemos con certeza es si esto se debe a que son parejas, y por eso pasan mucho tiempo juntas. Quizás los pájaros 33 y 32 son los que más co-ocurrencias tienen, solamente por ser los pájaros con más emisiones. Entonces podría ser posible que algunas de las diferencias vistas en la matriz sean una consecuencia azarosa de la distribución y la frecuencia de las emisiones.

4.1.2. Test estadístico

Proponemos entonces un test estadístico para determinar la significancia de la cantidad de co-ocurrencias de las posibles parejas. El objetivo es ver si ese valor es suficiente para determinar que la pareja es la ganadora no como consecuencia azarosa de la distribución y cantidad de emisiones de los pájaros, sino por la existencia de algún tipo de vínculo real entre los mismos.

Corrimos un test estadístico *Chi-Square* sobre la matriz de co-ocurrencia encontrada. Obtuvimos un **p-valor de 9.999e-05**. Esto significa que la tabla de co-ocurrencias se desvía significativamente de los valores esperados si las filas y columnas fueran independientes. En la figura 4.4 se pueden ver los residuos del test para cada pareja de pájaros. Cada residuo indica cuánto se desvía el valor observado de la celda respecto del valor esperado. Podemos ver siete parejas que se destacan sobre el resto, 10-11, 20-21, 25-26, 30-31, 29-34, 32-33 y 23-24. Una vez concluida nuestra experimentación, **el equipo de biólogos nos confirmó que estas siete parejas encontradas coinciden con sus observaciones en el campo.**

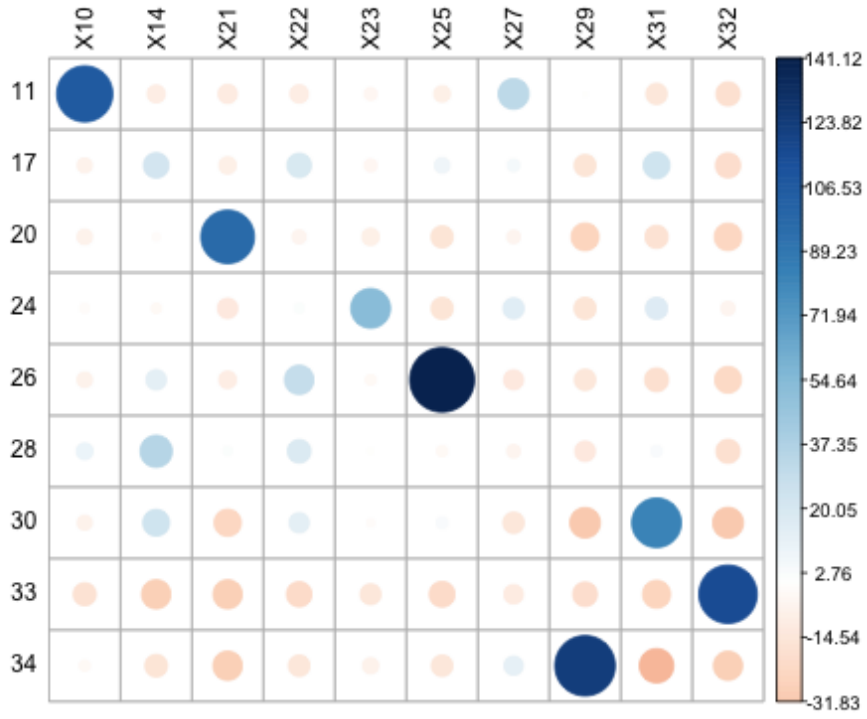

Fig. 4.4: Residuos resultado del test Chi-Square

4.1.3. Análisis de distancias promedio

En el punto anterior estuvimos analizando cuánto tiempo pasan juntos cada pareja de pájaros para poder determinar si hay ciertas parejas que pasan más tiempo que con los demás. Nos interesa también saber qué tan cerca se encuentran los pájaros entre sí. Si existieran parejas de pájaros, entonces estos pares deberían estar muy cerca (a metros) durante una buena cantidad de tiempo.

Como nuestro modelo no tiene la suficiente precisión, no podemos detectar cuándo y cuánto pasa esto. Pero sí podríamos generalizar un poco más, y comparar a qué distancia suele estar cada posible pareja de pájaros, esperando que las parejas reales estén a una distancia menor, en promedio, que las otras.

Vamos a tener casos en los que para un momento dado, para un pájaro tengamos una predicción y para otro pájaro no (es decir, si tuvimos señales no recibidas por ninguna antena). En ese caso, ¿a qué distancia se encuentran ambos? El pájaro del cual no obtuvimos señales puede ser que se encuentre muy lejos del campo de análisis, y por eso no lleguen las señales. O también podría ser que esté dentro de la región, pero que no hayan llegado las señales por alguna razón.

Vamos entonces a medir la distancia solamente para los momentos en que ambos pájaros se encuentren dentro la región comprendida por las cuatro antenas. En el boxplot de la figura 4.5, podremos apreciar las diferencias en las distancias de cada pareja posible de pájaros.

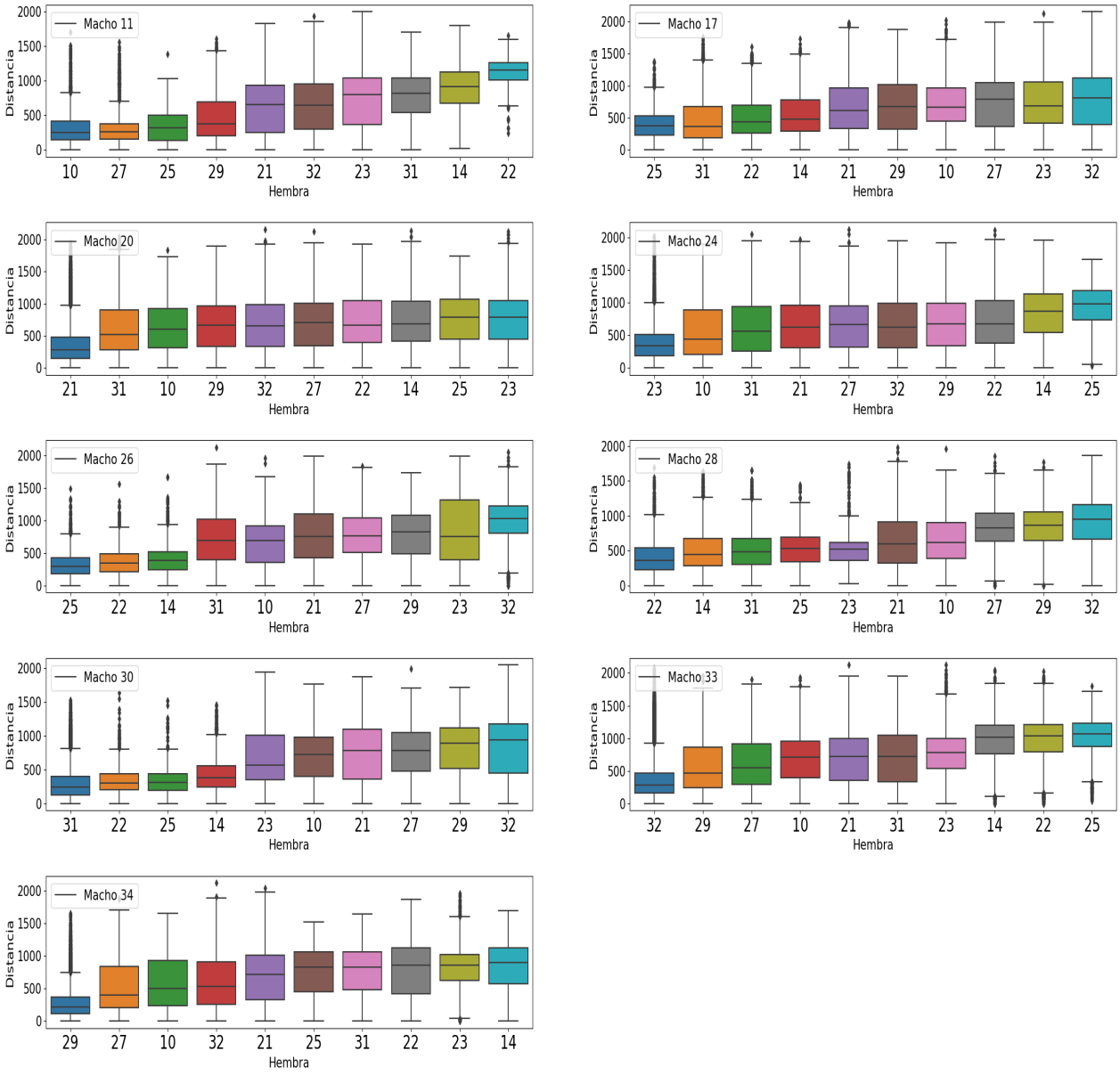

Fig. 4.5: Cuartiles de las distancias entre las ocurrencias de cada par de pájaros

Podemos ver cierta consistencia entre los resultados obtenidos en el análisis de co-ocurrencia y la distancia promedio en que se encuentran las posibles parejas, donde vemos que las parejas que más se destacaron del análisis anterior son, en líneas generales, las parejas que tienen menos distancia entre si en promedio.

4.2. Análisis de mapas de calor

Otro posible punto de análisis, complementario a los otros dos, es el análisis del área de acción de los pájaros, es decir las áreas por donde se mueve mayormente cada uno.

En los análisis anteriores comparábamos la ubicación de los pájaros en un mismo instante, midiendo distancia o si estaban en una misma región. Proponemos ahora analizar, sin tener en cuenta el momento exacto de las emisiones, las áreas del campo por donde se movieron los distintos pájaros durante el día. Creemos que si existen parejas de pájaros con comportamiento monógamo, sus áreas de acción serán similares. Es decir, si la hembra pasa mucho tiempo cerca de la antena RC1, esperamos que el macho correspondiente también lo haga.

En la figura 4.6 se puede ver los sectores del campo por donde se estuvo moviendo un pájaro durante un día. El pájaro estuvo la mayor parte del día cerca de las antenas *RC2* y *RC1*

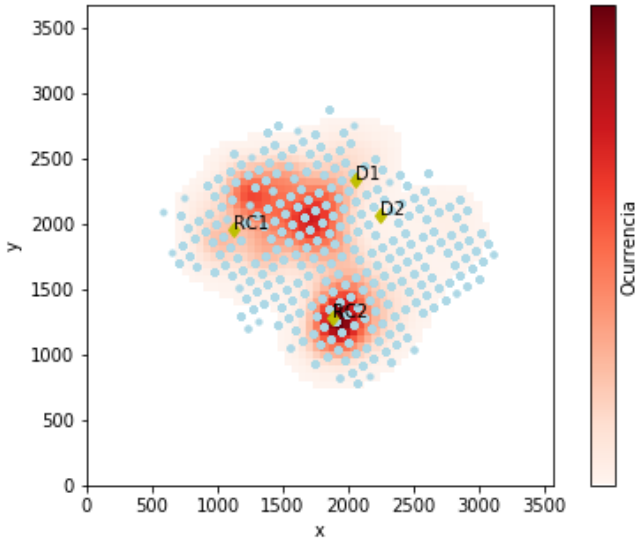

Fig. 4.6: Ejemplo mapa de calor para las predicciones de un pájaro durante un día

Para obtener los mapas de calor, dividimos el campo en 64x64 regiones regulares y contamos la cantidad de veces que estuvo el pájaro en cada región obteniendo una matriz de 64x64 con estos valores. Luego, le aplicamos a esta matriz una convolución con un filtro gaussiano con $\sigma = 2$. Este suavizado nos permite contemplar que si una predicción cayó en la región *(i,j)*, en realidad pudo haber caído en la región continua a la misma, pero con menor probabilidad. Finalmente dividimos cada posición de esta matriz por el total de emisiones del pájaro, para obtener en cada posición una estimación del porcentaje de las veces que el pájaro estuvo en esa región sobre las emisiones totales.

A continuación vamos a medir qué tan similares son los mapas de calor de cada posible pareja de pájaros. Vamos a analizar las similitudes de los mapas por día, y luego

tomaremos el promedio de las similitudes obtenidas. Para asegurar que los mapas de calor sean representativos, para cada pájaro utilizaremos solo los días que tuvieron más de 100 emisiones con recepciones de antenas.

Definimos la disimilitud o discrepancia entre dos mapas *M1, M2* de calor como:

$$\frac{\sum_{i=1}^{64}\sum_{j=1}^{64}|M1_{i,j} - M2_{i,j}|}{2} \tag{4.1}$$

con $M_{i,j}$ el valor calculado para la región (i, j). Notar que la discrepancia es un valor entre 0 y 1, obteniendo 0 si ambos mapas de calor son iguales y 1 los mapas son totalmente distintos. Una discrepancia baja significa una alta similitud en los mapas de calor.

La figura 4.7 muestra los resultados obtenidos fueron los siguientes, siendo cada posición de la matriz el promedio de la discrepancia de cada día entre dos pájaros. Los casilleros en blanco significan que no hubo días en el que ambos pájaros hayan tenido al menos 100 emisiones con recepciones.

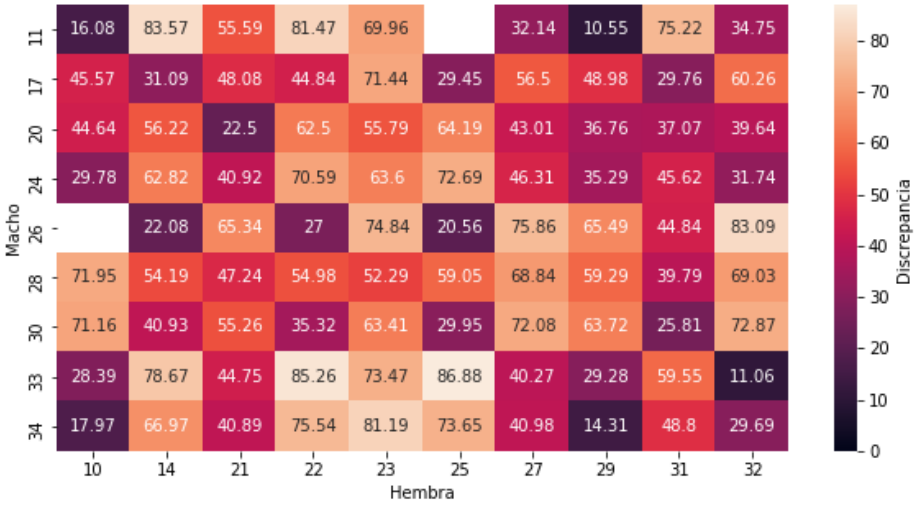

Fig. 4.7: Resultados discrepancia de mapas de calor

Podemos ver que la mayoría de los pájaros tienen a más de una posible pareja con similitud alta (discrepancia menor a 20 %). Por ejemplo entre las hembras 10 y 29, y los machos 11 y 34, todos tienen alta similitud. Esto nos da indicios de que hay pájaros que pueden tener áreas de acciones similares aunque no fueran pareja. Sin embargo, como punto positivo, vemos que varias de las parejas de pájaros destacadas encontradas durante el análisis de co-ocurrencia tienen alta similitud de los mapas de calor, es decir, del área de acción de los mismos. Ejemplos de esto son las parejas 10-11, 25-26, 29-34 y 32-33.

4.3. Hipótesis horario de salida del dormidero

Otra de las hipótesis que nos interesa validar, o refutar trata sobre los horarios de salida del dormidero de los machos y las hembras. Las hembras, cuando se encuentran en etapa de puesta salen del dormidero más temprano de lo usual, cerca de las 4am, hacia el nido que van a parasitar. Caso contrario, el horario normal de salida del dormidero es cerca de las 6am. Una de las funciones que se creen que los machos pueden cumplir en caso de que exista un comportamiento monógamo social, es la de acompañar a las hembras a parasitar. En ese caso, el horario de partida del dormidero de los machos debería coincidir con el de su pareja. Nos proponemos entonces encontrar evidencia de que existe una correlación entre el horario de partida del dormidero de los machos con los horarios de sus parejas.

4.3.1. Uso del predictor para encontrar la hora de partida

Para encontrar a qué hora partió un pájaro del dormidero, podemos utilizar el modelo de *machine learning* que hemos usado y analizado previamente. Sabemos que nuestro predictor tiene un error significativo, por lo tanto debemos validar si tiene sentido y es correcto su uso.

Hay sectores del dormidero en los cuales no hubo ningún punto de calibración para utilizar en el entrenamiento del modelo, así que es posible que el modelo no sepa predecir correctamente esa región. Y la calibración fue hecha a una altura que no sobrepasaba la altura de los árboles. Además el dormidero es una zona densa de árboles, por lo que es probable que haya más ruido y distorsión de las señales allí. Por esto, es muy importante que verifiquemos si el uso del predictor es correcto para este caso de uso.

Los pájaros suelen tener una fidelidad muy alta con el dormidero, es decir, la mayoría de las noches duermen allí. Tenemos entonces una forma de validar el predictor. Podemos estimar las posiciones de todos los pájaros en el horario donde se supone que se encuentran en el dormidero. Si nuestras predicciones indican que se encuentran ahí, entonces tenemos fundamentos para creer que funciona correctamente. Caso contrario, es posible que utilizar el predictor no sea lo deseado.

En la figura 4.8 podemos ver la proporción de las emisiones (con recepción de alguna antena) que caen, según nuestro modelo, en el dormidero. El color muestra la proporción mencionada, y el tamaño del punto muestra la cantidad total de señales recibidas por alguna antena. Estas emisiones ocurrieron entre las 0 y las 4hs, donde deberían haber estado durmiendo en el dormidero. Idealmente, la proporción para cada pájaro en cada día sería cercana a 1, es decir, casi todas las predicciones del modelo deberían caer dentro del dormidero. Pero vemos valores muy bajos, donde la proporción no alcanza al 10 %. En estos casos, en la mayoría, nuestro modelo predice la posición de los pájaros de manera errónea.

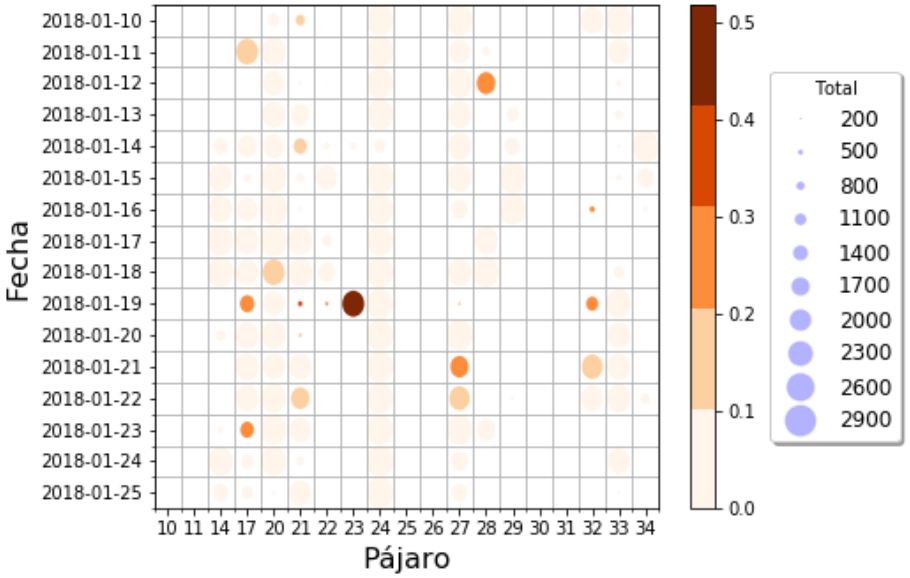

Fig. 4.8: Cantidad de emisiones con recepción de alguna antena y proporción de las mismas que caen dentro del dormidero.

Hay dos grandes motivos posibles de la mala predicción de nuestro modelo. Por un lado, puede ser que nuestro modelo sea malo para predecir en esa región del mapa, como mencionamos antes, por no tener puntos de calibración, o por tener una altura mayor de la que fueron obtenidos los puntos de calibración. La otra causa posible, es que señales de los transmisores sean mucho más ruidosas, por ser un área densa de arboles. Entonces quizás pequeños movimientos dentro del dormidero pueden implicar un cambio grande en las señales recibidas por las antenas. Por estas razones, no podemos utilizar el predictor para realizar este análisis.

4.3.2. Uso de las intensidades de las señales para encontrar la hora de partida

Una alternativa que tenemos, en lugar de utilizar el predictor, consiste en usar las señales recibidas por las antenas. En vez de tratar de predecir en qué momentos un pájaro se encuentra en el dormidero, y en qué momento no, lo que podemos hacer es mirar las intensidades recibidas por las antenas, y en el momento en que varíe mucho, suponer que el pájaro se movió.

Por ejemplo, si una antena recibía de un pájaro siempre intensidades cercanas a 20, y en un momento, empezó a recibir intensidades cercanas a 60, podemos pensar que éste se movió. No sabremos en qué dirección se movió, o si se movió mucho o no. Pero podríamos suponer que si se movió por primera vez a las 6am, entonces su horario de partida fue cercano a esa hora.

Para que esto tenga sentido debemos asumir que no existirán grandes variaciones en las intensidades recibidas por las antenas sin que los pájaros se muevan. Validaremos esto viendo que las intensidades de las señales recibidas por las antenas sean constantes desde

las 0am hasta las 4am, ya que el pájaro debería estar durmiendo, es decir, no debería moverse.

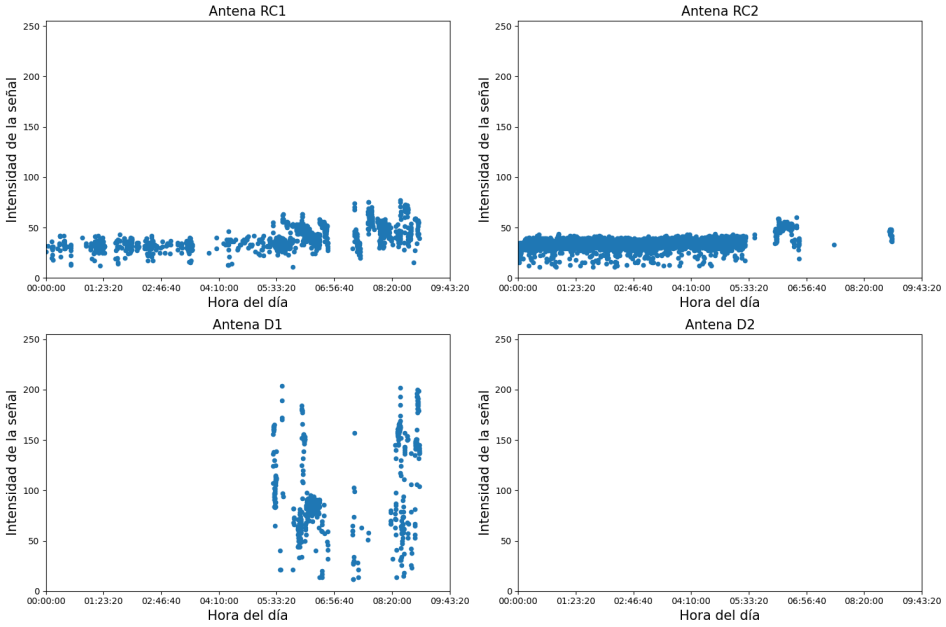

Fig. 4.9: Ejemplo de la distribución de las intensidades de las señales recibidas por las cuatro antenas a lo largo de un día, para el pájaro macho número 20, elegido a modo ilustrativo.

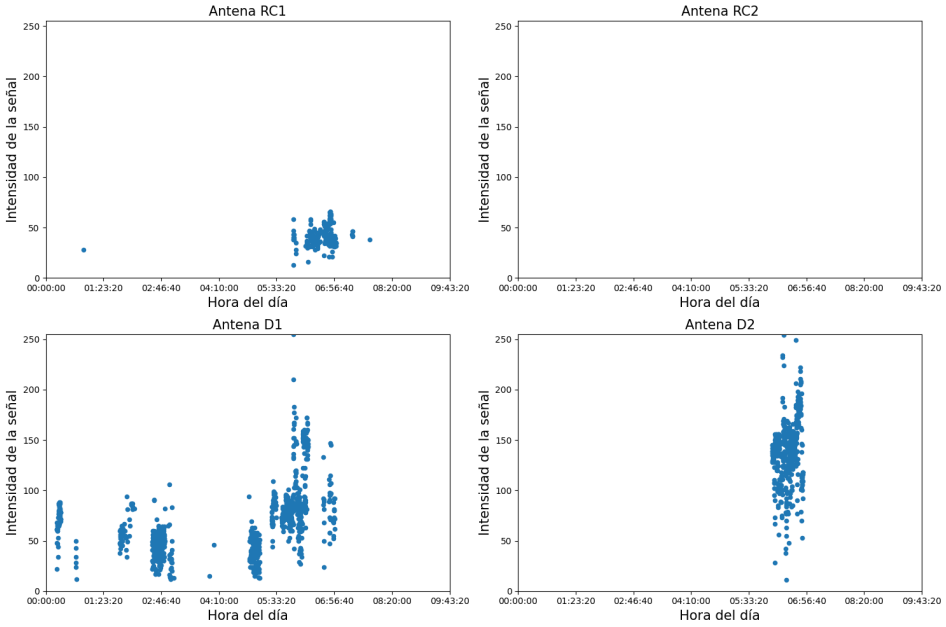

Fig. 4.10: Ejemplo de la distribución de las intensidades de las señales recibidas por las cuatro antenas a lo largo de un día, para el pájaro hembra número 27, elegido a modo ilustrativo.

Encontramos, como vemos en las figuras 4.9 y 4.10, que hay algunos casos donde se distingue fácilmente cuál es el momento donde hay una variación significativa en las señales, y otros casos donde no es tan sencillo establecer el momento. También se distinguen casos donde la primer variación ocurrió antes de las 4am, donde suponíamos que el pájaro estaba durmiendo.

Cada uno de estos gráficos muestran la variación de las intensidades recibidas por cada antena para un pájaro. Nos interesa identificar los momentos en que la señal varíe de forma significativa. Para esto, utilizaremos un algoritmo de *change point detection* cuya tarea es encontrar cambios en el modelo subyacente de una señal o serie temporal [5]. Ejemplo de esto se puede ver en la figura 4.11. De esta manera, podemos establecer el momento de salida del dormidero de cada pájaro como la detección más temprana de cambio en las señales de todas las antenas.

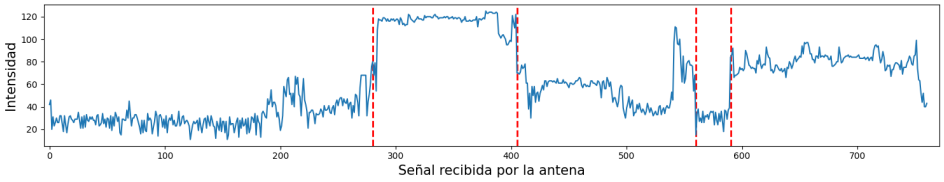

Fig. 4.11: Ejemplo de la detección de puntos de cambio sobre las señales recibidas por la antena RC1 de emisiones del pájaro 20.

Al utilizar estos algoritmos de *change point detection* es posible que haya más de un punto de cambio detectado, por ejemplo, si el pájaro se movió más de una vez. Si buscamos un solo punto de cambio, los algoritmos devolverán el punto que encuentren con el cambio de la señal más significativo. Entonces si ocurrió un cambio en la señal a las 4am y otro más pronunciado a las 7am, diremos que el punto de cambio, y por lo tanto la salida del dormidero fue a las 7am, cuando en realidad hubo un movimiento en un horario previo. Al no conocer el número de puntos de cambio disponemos como alternativa definir un valor de penalidad [5]. El concepto de penalidad es análogo al de término de regularización utilizado frecuentemente en el área, cuyo objetivo es balancear la complejidad de la función o modelo subyacente. El valor de penalidad entonces nos permite controlar la importancia del término de regularización. Con un valor muy chico de penalidad, se detectarán muchos puntos de cambio, incluso aquellos que son causados por ruido en la señal. En cambio, con un valor muy grande de penalidad solo se detectarán los cambios más significantes, o ninguno. Este enfoque tiene la desventaja que requiere encontrar un valor óptimo de penalidad, y que este valor sea apropiado para todas las señales a analizar, caso contrario podría no encontrar puntos de cambio para algunas señales y/o determinar que el ruido en la señal es un punto de cambio para otras. Finalmente, ante la dificultad de hallar un valor justo de penalidad, decidimos utilizar solo un punto de cambio por señal, el punto donde se encuentre el cambio más grande en la intensidad de la señal.

Luego de realizar algunas pruebas, encontramos los mejores resultados en la detección de los puntos de cambio utilizando el algoritmo llamado *Window sliding*. Este es un algoritmo aproximado que computa la discrepancia d entre dos ventanas adyacentes sobre una señal y, utilizando una función de costo c para calcular la discrepancia de la forma $d(y_{u..v}, y_{v..w}) = c(y_{u..w}) - c(y_{u..v}) - c(y_{v..w})$ siendo $y_{u..v}$ la señal de y en el intervalo de tiempo entre los instantes u y v. La función de costo c mide el *goodness-of-fit* de la señal

a un modelo específico, es decir, la diferencia entre los valores observados y los valores esperados por el modelo en cuestión. Cuando dos ventanas cubren segmentos distintos, la discrepancia tiene un valor grande, caso contrario, tiene un valor chico. Entonces para cada índice i, el algoritmo *Window sliding* mide la discrepancia entre la ventana pasada y futura inmediata. Finalmente, realiza una búsqueda de los picos en los valores de las discrepancias, para definir a los mismos como los puntos de cambio.

Como mencionamos previamente, vamos a establecer el momento de salida del dormidero como la detección más temprana de cambio en las señales de las cuatro antenas. Este fue el criterio que mejores resultados, a simple vista, nos otorgaba. Otros criterios considerados para determinar el momento de salida fueron:

- Utilizar los algoritmos previamente mencionados de detección de punto de cambio, pero en vez de utilizar la señal de una antena a la vez para la detección, utilizar una señal n-dimensional compuesta por las recepciones de las cuatro antenas.
- Aplicar técnicas de reducción de dimensionalidad para transformar las señales de las cuatro antenas en una sola señal unidimensional. A través de algoritmos como *principal component analysis*, *Isomap* [6] o *TSNE* [7] buscamos transformar las señales recibidas por las cuatro antenas, es decir, una señal 4-dimensional, en una señal unidimensional. Utilizando tal transformación, pequeños cambios en las señales de las cuatro antenas deberían producir pequeños cambios en la nueva señal.

La ventaja que nos otorga el criterio elegido es que tenemos un mayor control para decir qué hacer en los casos donde alguna emisión de la señal no llega a alguna antena. Caso contrario, los algoritmos podrían detectar a estas emisiones como punto de cambio, cuando en realidad es un comportamiento esperado de la señal.

Entonces, aplicando el procedimiento mencionado, y suponiendo que el primer movimiento del pájaro durante el día corresponde a la salida del dormidero, se obtuvieron los horarios de salida de todos los pájaros desde el 2018-01-10 hasta el 2018-01-25 inclusive. Los histogramas en las figuras 4.12 y 4.13 muestran cuáles fueron los horarios de salida para cada pájaro.

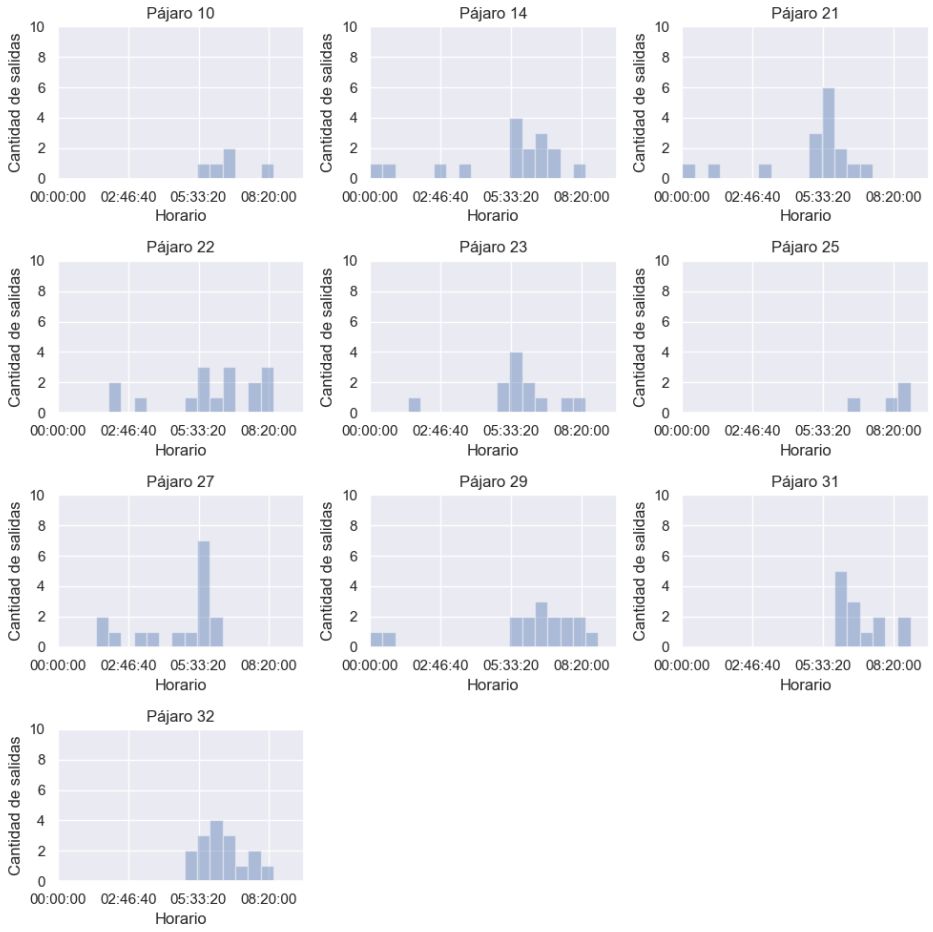

Fig. 4.12: Horario estimado de salida del dormidero, para pájaros hembras

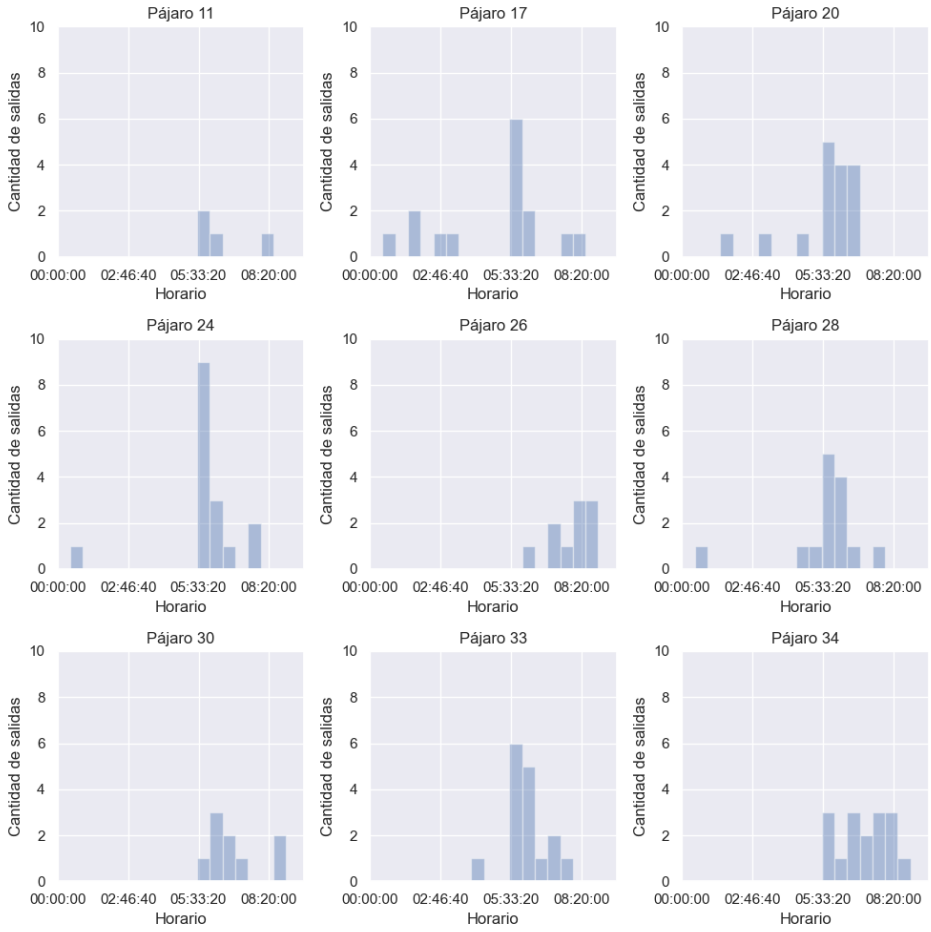

Fig. 4.13: Horario estimado de salida del dormidero, para pájaros machos

Podemos ver que hay varios picos alrededor de las 6am. Sin embargo, observamos muchos casos donde los pájaros "salen", de acuerdo con nuestro método de estimación, antes de las 4am. Esto no debería pasar, ya que antes de ese horario los pájaros deberían estar durmiendo, o sea, no debería haber cambios en las intensidades de las señales. Esto nos da evidencia de que esta idea y método (en especial por los datos de las señales que tenemos) no son suficientemente precisas como para permitirnos un análisis de los horarios de partida de los dormideros.

4.4. Hipótesis nidos activos

En esta sección buscamos estudiar la actividad del Tordo Pico Corto, tanto macho como hembra, en relación con los nidos activos. Los nidos activos son aquellos nidos del hospedador, los cuales son posibles de utilizar para parasitar.

Habiendo hecho pruebas sobre nuestro modelo de aprendizaje automático, tenemos evidencia para creer que no es posible el estudio de los nidos activos, ya que no tenemos precisión suficiente para detectar cuándo un pájaro se encuentra cerca de un nido activo.

Por lo tanto, dejamos este análisis inconcluso.

5. CONCLUSIONES Y TRABAJO FUTURO

En esta tesis realizamos un trabajo colaborativo con un grupo de investigación de biología donde, a partir del uso de herramientas del área de *machine learning* pudimos participar y ayudar en el proceso de validación de hipótesis aportando evidencia que ayude a decidir y justificar los resultados.

El foco principal del trabajo estaba en la hipótesis y el análisis del comportamiento monógamo social del Tordo Pico Corto. Encontramos evidencia que sugiere que la hipótesis de que estos pájaros tienen un comportamiento monógamo social es verdadera. Dentro de los individuos analizados encontramos, a través de distintas estrategias, parejas que se destacaban sobre otras. Se destacan por el tiempo que pasan juntos, por la cercanía que se encontraban, o una combinación de ambas.

En nuestro análisis, encontramos parejas que sobresalían por sobre otras por tener valores mucho mayores de co-ocurrencia, es decir, de veces que estos pares de pájaros se encontraban juntos. El test estadístico nos ayudó a determinar que estos altos valores se deben a que existe un vínculo real entre las parejas halladas. Las parejas más destacables son las parejas 20-21, 32-33, 29-34, 30-31, 26-25, 11-10 y 24-23. Estas parejas también se destacan, en mayor o menor medida, en los análisis posteriores de mapas de calor y distancia promedio. Este comportamiento hallado y las parejas encontradas coinciden con siete de las diez potenciales parejas hipotetizadas en el campo por el grupo de investigación de biología. Para las restantes tres parejas no se observaron asociaciones en los datos.

Con respecto al resto de las hipótesis de interés (salida del dormidero y nidos activos), no logramos suficiente precisión en nuestra metodología que nos permitiera encontrar evidencia significativa para aportar a la validación de las mismas, quedando así para trabajo futuro.

De lo visto durante el transcurso de este trabajo, surgen algunas ideas para mejorar nuestros modelos de aprendizaje automático con el fin de tener estimaciones más certeras de la ubicación de los pájaros dentro del área en estudio. Para esto es fundamental contar con una mayor cantidad de datos de calibración y una mejor calidad de los mismos, para el entrenamiento de estos modelos. Por ejemplo, con el uso de drones con GPS, colocándoles los radiotransmisores, sería posible obtener muchos datos con bajo esfuerzo humano (en comparación con el método actual, que involucra a una persona moviéndose y registrando las señales por la reserva). Además estos drones podrían emular de mejor manera el vuelo de los pájaros, con lo cual resultarían de mejor calidad que los datos actuales y nos permitirían entrenar modelos que utilicen la temporalidad de las emisiones.

El otro punto de mejora posible es sobre la calidad de las emisiones de los pájaros. El radiotransmisor utilizado puede no ser óptimo para la tarea que estamos realizando, conociendo que las señales de radio son muy sensibles a ruido, y quizás otras tecnologías sean mejores para este caso de uso, por ejemplo GPS. Por supuesto que para el cambio de tecnología existen otras limitaciones, como el costo del mismo y el presupuesto disponible, o el impacto ecológico (los transmisores deben ser lo suficientemente livianos para minimizar el posible impacto en la vida de los individuos). Otra forma de mejorar la calidad de

los datos de los pájaros sería agregar más antenas. De esta forma, contaríamos con más información para hacer las predicciones. Estas dos últimas opciones implicarían tener que capturar y recopilar nuevamente datos sobre los pájaros, y además, volver a recopilar los datos de calibración para el entrenamiento de los modelos.

Bibliografía

[1] Romina Scardamaglia and Juan Reboreda. Ranging behavior of female and male shiny cowbirds and screaming cowbirds while searching for host nests. *The Auk*, 131:610–618, 10 2014.

[2] Romina Scardamaglia, Alex Kacelnik, and Juan Reboreda. Roosting behaviour is related to reproductive strategy in brood parasitic cowbirds. *Ibis*, 02 2018.

[3] T.M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[4] Yehuda Vardi and Cun-Hui Zhang. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, 97:1423–6, 03 2000.

[5] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[6] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[7] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.