Buen trabajo Naty.. por su naturaleza los datos son un poco conflictivos, podemos hacer algo con ellos o podrías ir a una base de datos más certera y más “científica”

3.9

*sobre 4.5

Análisis de datos migratorios sobre Cardellina canadensis

Introducción

En Colombia se encuentran cerca de 1898 especies de aves, lo que sitúa al país como uno de los primeros en el mundo en cuanto a la diversidad de este grupo. Dentro de esta amplia riqueza se encuentra un elevado número de especies que presentan comportamientos migratorios recurrentes y cíclicos (aproximadamente 275), que equivalen a 14,5% de las especies de aves presentes en el país. La mayoría de estos desplazamientos se inician en respuesta a un estímulo ambiental y cesan en cuanto dicho estímulo deja de presentarse. El movimiento de un lugar a otro permite el ajuste permanente a las condiciones cambiantes del entorno, mientras estas no sean superiores a los límites de tolerancia de la especie en cuestión. El objetivo de este trabajo es evaluar si una de estas especies migratorias “*Cardellina canadensis*” o reinita del Canadá (Figura 1-a) , ha experimentado cambios en los patrones de migración hacía Colombia en los últimos 20 años, utilizando los datos de la plataforma Ebird.

La especie *Cardellina canadensis* se moviliza en la época migratoria (Octubre a Marzo) a Colombia desde el hemisferio Norte en Cánada (Figura 1-b). Esta catalogada como una especie en peligro según la ley de especies amenazadas del Canadá y entre los mayores problemas para su conservación se encuentran las áreas para invernar en américa del sur, sin embargo la información para determinar su permanencia en el país es poca y dispersa, por ello herramientas de gestión de datos como la plataforma Ebird y los diferentes análisis que puedan surgir de esta información, son esenciales para el manejo y conservación de la especie.

Figura 1. (a) foto de reinita del Canadá y (b) ruta migratoria de la Reinita del Canadá hacía Colombia.

Plataforma Ebird

Ebird es una plataforma de datos desarrollada por el laboratorio de ornitología de la universidad de Cornell. Su principal objetivo es recoger la información derivada de listas de aves para almacenarla y comparirla de forma gratuita, con el fin de habilitar nuevos enfoques para la ciencia, la conservación y la educación.

Ebird es el proyecto de ciencia ciudadana relacionado con biodiversidad más grande del mundo, con más de 100 millones de registros de aves contribuidos a diario que consiste en un marco científico simple: los pajareros indican cuándo, dónde y cómo fueron a pajarear, y luego llenan una lista de chequeo de todas las aves observadas o escuchadas durante la jornada.

La calidad de los datos tiene una importancia crítica. Al ingresar sus registros, los observadores reciben una lista de las especies que son probables para esa fecha y región. Estos filtros para las listas de chequeo son desarrollados por los mejores expertos en distribuciones de aves en el mundo. Cuando se reportan aves inusuales, o se reportan conteos muy altos, los expertos regionales revisan estos registros.

En la figura 2 se presenta la interfaz que presenta la plataforma para Cardellina Canadensis en Colombia en el 2014. Cada mes se divide en cuatro períodos (cada uno de aproximadamente 7 días). En cada período, se calcula el porcentaje de listas que reportan una especie, pero se consideran sólo aquellas listas que reportan todas las especies. Las barras verdes más amplias muestran los períodos en los que es menos probable que no se detecte una especie en particular, mientras que las barras verdes estrechas muestran que dicha especie está presente, pero es detectada con menor frecuencia, y las barras grises representan los periodos sin listas.

Los datos obtenidos de la plataforma para la distribución de Cardellina Canadensis en Colombia cuenta con información semanal sobre: (i)frecuencia, referente al número de listas en las cuales aparece registrada la especie, dividida entre el número total de listas de dicho periodo, (ii) abundancia, es el número promedio de aves reportadas en todas las listas de verificación dentro de un rango de fechas y región especificadas; (iii) aves por hora es la cantidad promedio de aves observadas por hora de observación de aves dentro de un rango de fechas y región especificadas (iv) conteo máximo, es el conteo más alto de una especie presentada en una sola lista de verificación dentro de un rango de fechas y región especificada; (v)totales, es la suma de todas las observaciones de una especie de todas las listas de verificación presentadas dentro de un rango de fechas y región especificadas(vi) conteo promedio difiere de "Abundancia" en que solo incorpora listas de verificación que informaron la especie (sin ceros).

Figura 2. .Opciones de obtención de datos en al plataforma Ebird

Análisis de los datos

1. *Lectura y gráfica de la serie*

A continuación en la figura 3 se muestra la serie de datos que representan el avistamiento de *Cardellina Canadensis* en Colombia, desde Enero de 1998 hasta septiembre de 2018 con una resolución de datos semanal. En la figura 3-a se muestra la serie de la frecuencia de avistamiento de la especie en las listas ingresadas en una semana, dividido entre el total de las listas para esa semana. En la figura 3-b se muestra el conteo máximo de individuos en cada lista. Para el análisis de datos se escogió la serie con la frecuencia de listas, debido a que es menos sesgado que el número máximo de individuos, porque este último no toma en cuenta el número total de listas ingresadas a la plataforma.

Otro argumento válido para no tomar como objeto de análisis la serie de conteo máximo es que al inicio de la serie (1998) (Figura 3 b) no existía esta plataforma, el acceso a internet era muy limitado y no se reconocía la importancia de realizar investigaciones a partir de estos datos, por ello el conteo de especies era muy baja. Pero esta práctica y los datos han ido aumentando con el tiempo, debido al reconocimiento de la plataforma y a que cada día se suma más gente a esta actividad y reconoce su importancia para la toma de decisiones.

En la figura 3-a se observa que las frecuencias son muy altas al inicio de la serie debido a que era menor el número total de listas ingresado en 1998 que en el 2018. Las mayores frecuencias se encuentran en los últimos años, porque el total de listas ha aumentado.

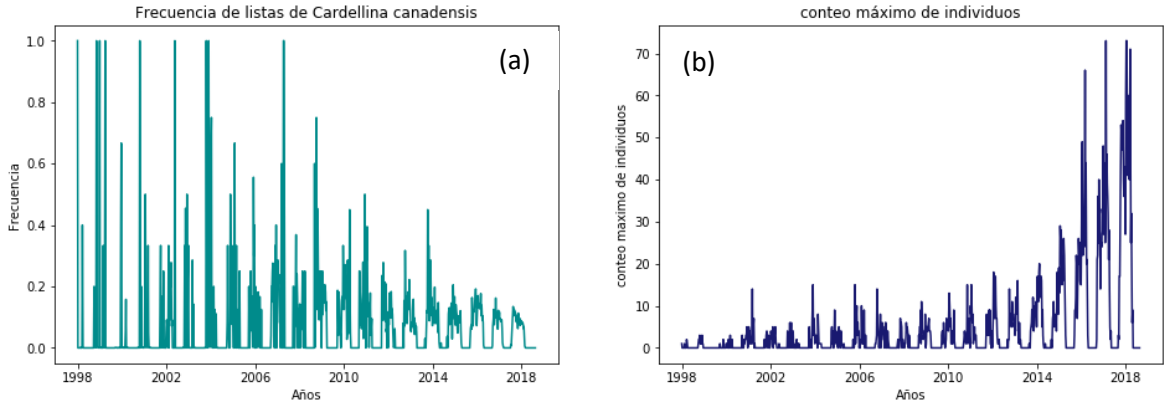
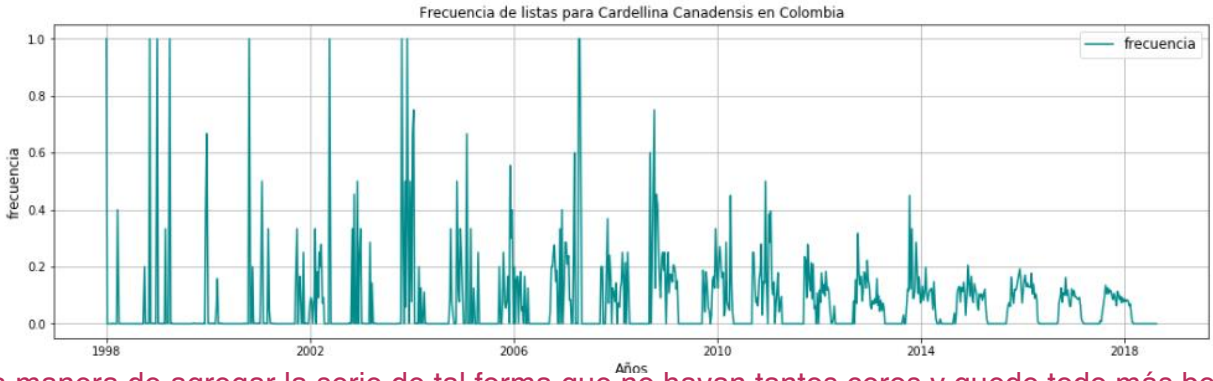


Figura 3. Series de datos de avistameinto de Cardellinas canadensis en Colombia



Busquemos la manera de agregar la serie de tal forma que no hayan tantos ceros y quede todo más bonito :)

Figura 4. Serie de frecuencia de listas de Cardellina cancdensis.en Colombia.

2. *Percentiles y función de distribución de probabilidad (FDP)*

Los percentiles más frecuentes de distribución de la serie original se muestran en la tabla 1. Estos percentiles describen o indican el panorama general de la distribución de los datos, en donde casi la mitad de los datos (hasta P _49) tienen un valor de 0.0, a partir del P_50 empiezan a obtener valores muy bajos; esto indica que, al ser el valor de la mediana(P_50) tan bajo, es muy poco probable observar *Cardellina canadensis* en cualquier semana del año, según esta mediana.

Los datos de frecuencia de avistamiento tienen una clara acumulación en la primera parte de la gráfica de la figura 5, en donde la mayoría de los percentiles (hasta P_90) tienen un valor inferior a 0.2, es decir que el porcentaje de frecuencia de avistamiento para la especie en general es muy baja.

percentil (%)	frecuencia de avistamientos

25	0
50	0.0015
75	0.1065
90	0.2

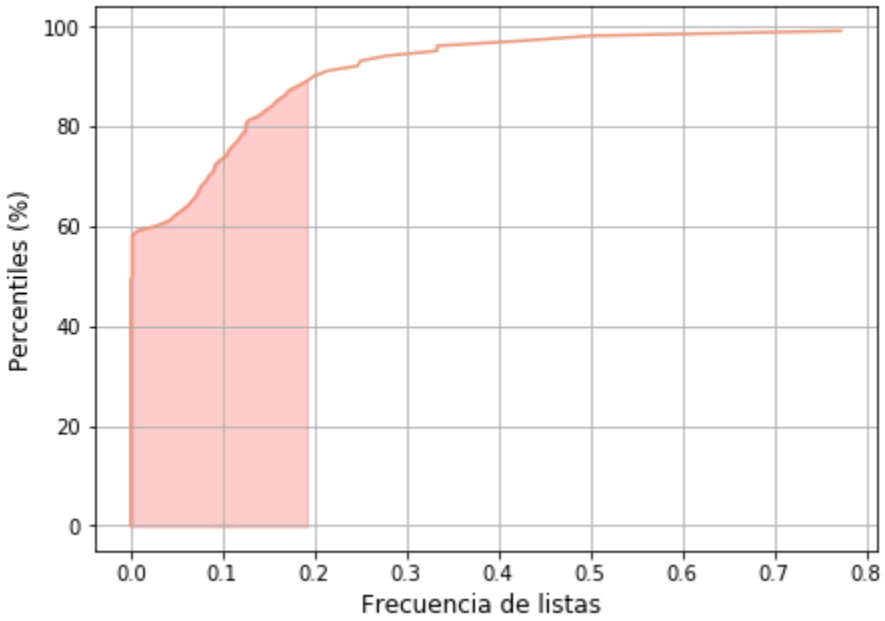


Figura 5. Distribución de los principales percentiles en la serie.

Función de distribución de probabilidad

La función de distribución de probabilidad asigna a un rango de datos la probabilidad de que estos ocurran. En la figura 6 se muestra la función de distribución de probabilidad que representa la última columna de la tabla 2, que además de presentar la probabilidad para la mitad de cada rango de datos (0.5), muestra la frecuencia en la cual se presenta ese rango dentro del conjunto total de datos.

Tabla 1. Frecuencia y probabilidad para los rangos de distribución de los datos.

rango de datos	frecuencia	probabilidad
0	728	0.73387097
0.1	156	0.15725806
0.2	56	0.05645161
0.3	17	0.00907258
0.4	9	0.01108871
0.5	11	0.00302419
0.6	3	0.00201613
0.7	2	0

0.8	0	0.01008065
0.9	10	

En la figura 6 se observa que el primer rango de los datos, es decir de 0-0.1 de frecuencia en el avistamiento de la especie, se repite con mayor frecuencia (728 datos) entre este rango de valores, esto le confiere la probabilidad más alta (0.733).

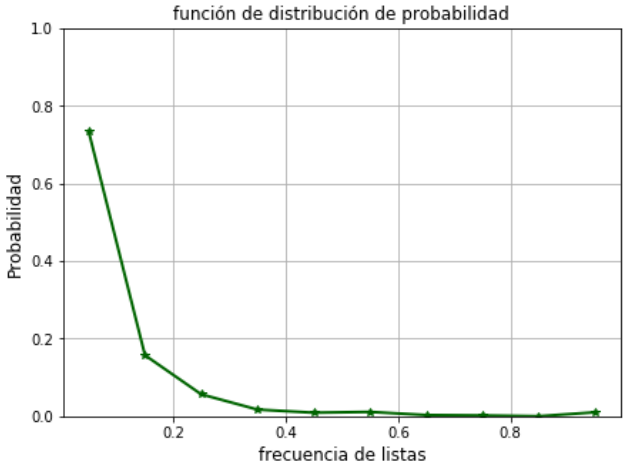
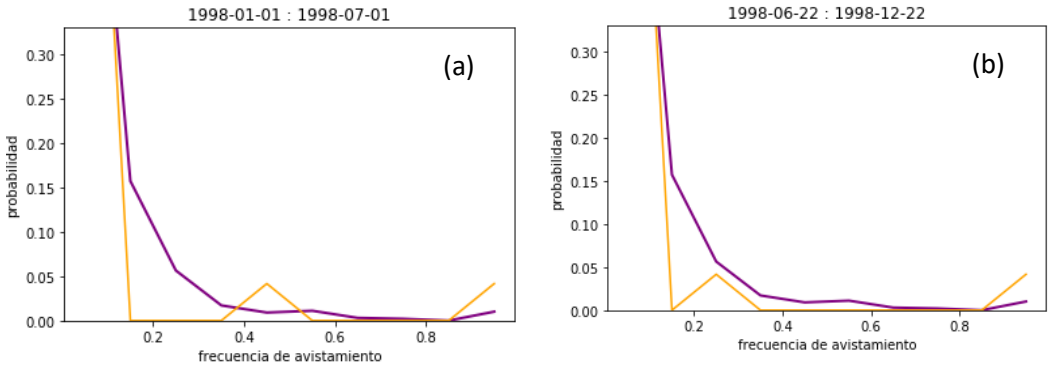


Figura 6. Función de distribución de probabilidad.

En la figura 7 se muestran diferentes panoramas, que comparan en cuatro diferentes épocas en el año como cambia la FDP de la ventana de tiempo que se quiere analizar(color naranja) respecto a la FDP original (morada). En este caso la época migratoria va de Octubre a Marzo, en estos 6 meses hay 24 datos de frecuencia por la resolución semanal de los datos. En la figura 7-a, se observa que al ser una época migratoria, hay una mayor frecuencia de observar la especie (0.4-0.6), mientras que para un mes donde no hay migraciones como Junio (figura 7-b), la frecuencia de avistamiento es más baja (0.2-0.4). Los mismos casos más extremos ocurren en las figuras 7-c y 7-d, en la primera una alta frecuencia de avistamiento en



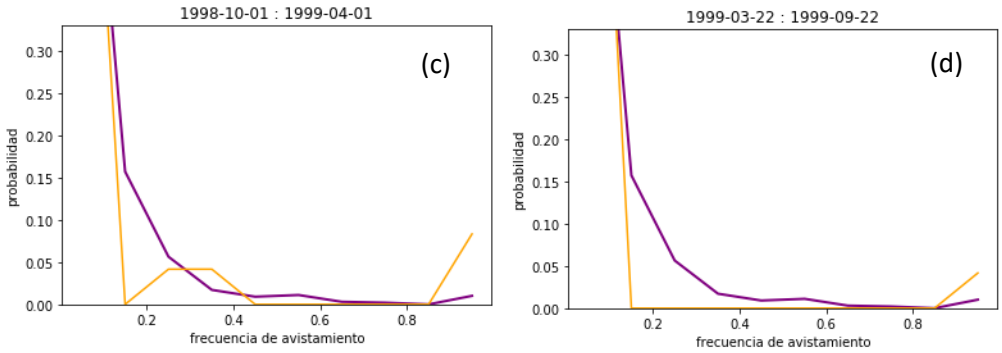


Figura 7. Histogramas para diferentes periodos en la ventana móvil.

3. Estimación de indices de localización, dispersion y simetría

índices de localización

Los índices de localización como la media y la mediana son utilizados para observar la tendencia central de los datos. La media no es una medida robusta (supone una distribución normal de los datos) ni resistente (toma en cuenta los "outliers"), por ello la mediana es una medida más confiable acerca de la tendencia central. En este caso, según la figura 8, la media y la mediana de la serie son muy distintas al inicio de la serie, debido a la falta de observaciones, de tal manera que cualquier dato afectaba más a la media y la mediana se mantuvo por mucho tiempo en 0, la cual si representa la tendencia central de los datos por mucho tiempo. A medida que se fueron alimentando las bases de datos que contenían la especie, tanto la media como la mediana se fueron estabilizando, hasta mostrar un comportamiento muy similar al final de la serie.

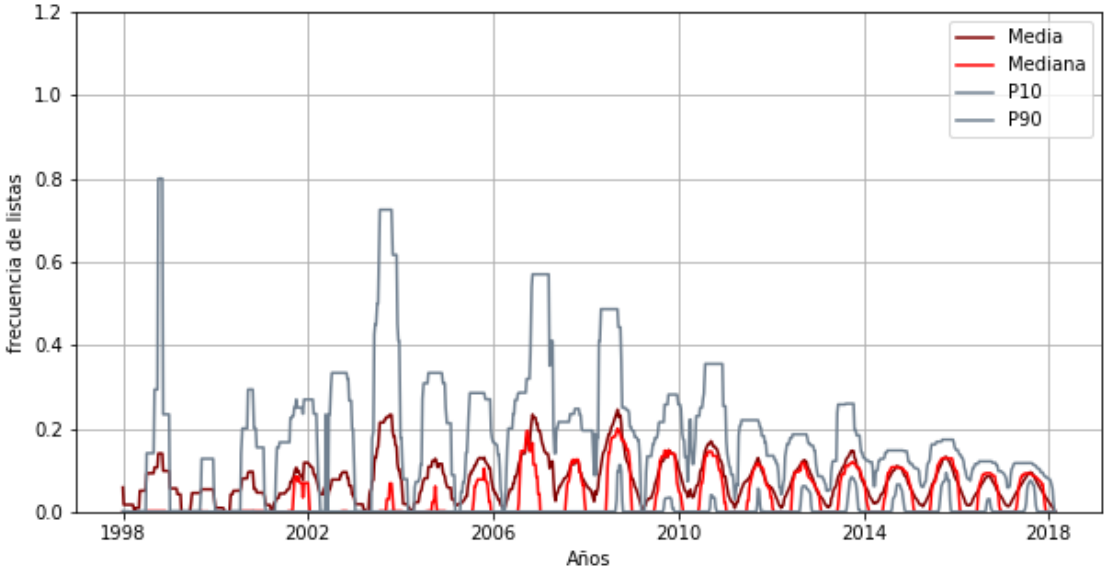


Figura 8. Medidas de localización para la serie de datos.

Yo no caí en cuenta en el momento que me mostraste esto en decirte que agregaras la serie. Pero podrías hacer el ejercicio, empecemos por mensual…. nos sirve para el resto del semestre ver como nos queda… me vas mostrando ;)

Índices de dispersión

Existen diversos métodos para hallar las medidas de dispersión alrededor del valor central de los datos. En este caso utilizaremos la desviación estándar y el rango intercuartil (IQR). La desviación estándar es una medida que no es robusta ni resistente, debido a que toma la diferencia entre cada uno de los datos respecto a la media, sin tomar en cuenta si estas son "outliers"; por el contrario el rango intercuartil no toma en cuenta el 25% de mis datos al inicio y al final, siendo una medida más resiste y robusta.

En la figura 9 se observa que ambas medidas tienen un comportamiento muy similar con ligeras variaciones. La desviación estándar al principio de la serie ocupaba un rango de variación muy alta, esto debido a que eran pocas las listas totales ingresadas, entonces la frecuencia de observación de la especie era superior, dando como resultado valores que fluctuaban entre 0 y 0.35. En años más recientes con la inclusión de un mayor número de listas totales por semana ha producido que la frecuencia de las listas donde se observa la especie sea inferior y fluctué entre 0 y 0.10 , y que sea 0 solamente en la época migratoria. EI IQR muestra rangos de variación más altos que fluctúan entre 0 y 0.40 al inicio de la serie, y de 0 a 0.15 en los últimos años, precisamente porque no toma en cuenta los "outliers".

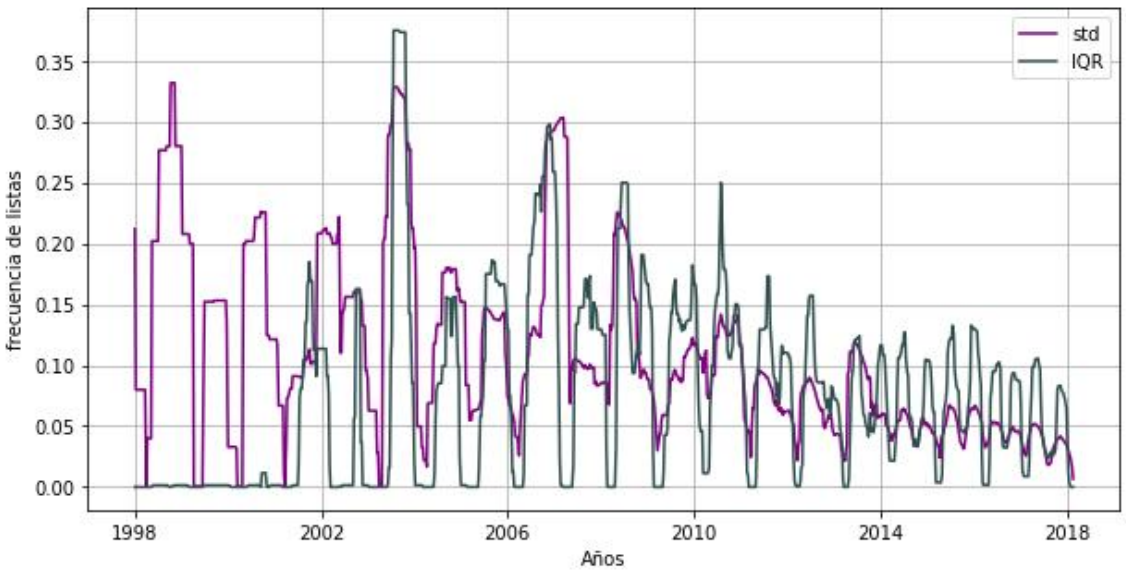


Figura 9. Medidas de dispersión de los datos.

Índices de dispersión

La medida de dispersión de *Yule Kendall* es uno de los descriptores de la distribución de los datos. Cuando esta medida es superior a 0, se puede inferir que nuestros datos tienen una mayor frecuencia hacía la izquierda y si el índice es inferior a 0, tienen una distribución a la derecha del conjunto de datos.

En la figura 10 se observa que el grueso de los datos se agrupa por encima de 0, es decir que la distribución de los datos se encuentra distribuida hacía la izquierda, y en una menor proporción hay datos hacía la derecha, por lo que se puede concluir que la serie no tiene simetría.

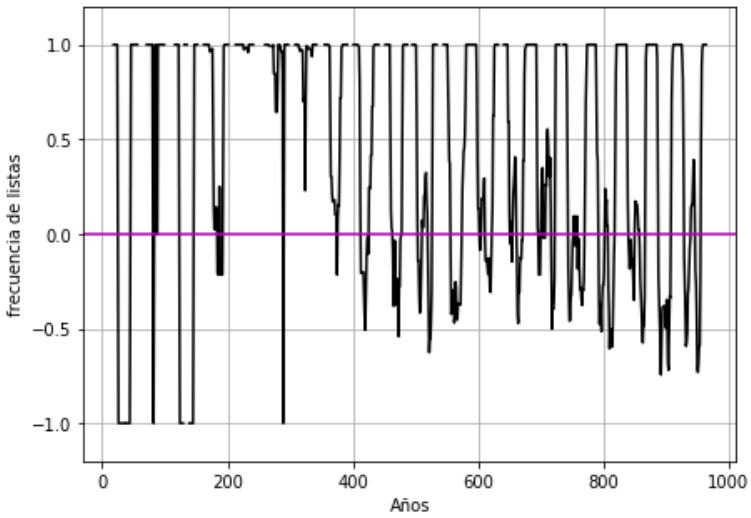
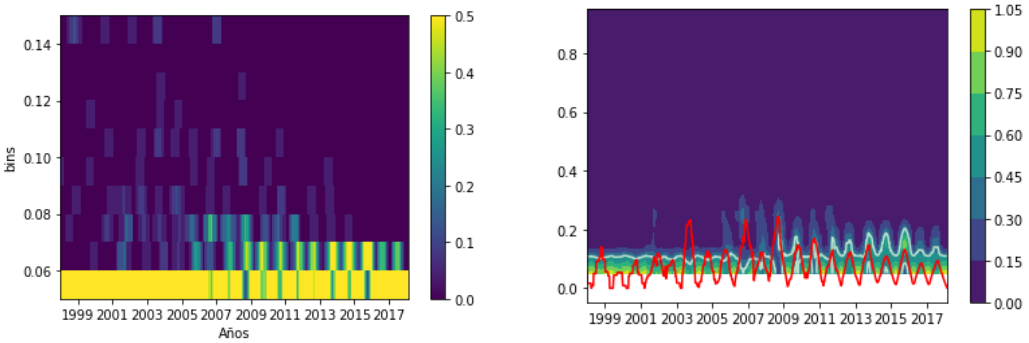


Figura 10. Indice de Yule Kendall para la serie de datos.

4. *Son estacionarios los histogramas?*

En la figura 11 se observa cómo se comportan todos los histogramas de la ventana móvil que se va moviendo a través de la serie de tiempo. Se observa que estos no se mantienen estables alrededor de un valor, sino que fluctúan y su frecuencia en los primeros rangos de distribución va aumentando hacía el final de la serie. Así mismo la media móvil de los datos no se mantiene alrededor de un valor, sino que va fluctuando entre 0 y 0.24, por lo que se puede decir que la serie de datos de Cardellina canadensis no es estacionaria.

Cuál es la diferencia entre estos dos?



No entiendo esa franja blanca en el histograma

Figura 11. Grafica de la media móvil de la serie.

6. *Existe tendencia en su serie y en los percentiles?*

La función de Mann Kendall tiene como objetivo identificar si existe una tendencia en los datos y dependiendo de la hipótesis nula propuesta, si esta tendencia es positiva o negativa, es decir si los datos tienen tenencia creciente o decreciente.

Para ello, primero se propuso la hipótesis nula de que los datos no tenían una tendencia, como se muestra a continuación:

Ho: no hay tendencia

Una vez establecida la hipótesis nula, el primer paso es determinar a través del estadístico de prueba (S) el signo de la diferencia entre Xj y Xk en la ecuación (1) que en el caso de los datos de la serie resulto ser mayor a 0.

$$S = \sum_{k-1}^{n-1} \sum_{j-k+1}^{n} \mathrm{sgn}(x_j - x_k) \quad (1)$$

Por lo que se aplica la formula (2) para establecer el valor de Z, se toma el caso de S>0.

$$\text{Si } S > 0 \qquad Z_{MK} = \frac{S-1}{\sqrt{V(S)}}$$

$$\text{Si } S = 0 \qquad Z_{MK} = 0$$

$$\text{Si } S < 0 \qquad Z_{MK} = \frac{S+1}{\sqrt{V(S)}}$$

En la tabla 3 se describen los valores de z obtenidos para la media y para los percentiles más importantes (P25,P50 y P75). Por ejemplo para el caso de la media, el valor de z es de 1.66, de tal forma que necesito buscar un z en la distribución gaussiana que sea menor a mi z para poder rechazar la hipótesis nula con un nivel de significancia acorde.

En el caso de mis datos el nivel de significancia de la distribución gaussiana para mi z es de 0.9515, debo buscar un z de distribución normal que sea inferior al de mi serie con el fin de rechazar al hipótesis nula.

Se debe tener en cuenta que en este caso mi hipótesis nula es que no hay tendencia en los datos, por ello debo evaluar el z de la distribución gaussiana entre 2 ($\propto/2$) porque tengo probabilidades tanto de que la tendencia sea negativa o de que sea positiva, por ello tiene dos colas. Finalmente entonces decido rechazar la hipótesis nula con un nivel de significancia de (0.95202/2)=0.47601. De la misma manera se aplica para todos los percentiles cuyo nivel de significancia es de 1, decido entonces rechazar la Ho de tendencia nula para cada percentil con un nivel de significancia de 0.5.

Tabla 2. Tabla con valores del test de Man Kendall para la serie móvil.

test de mann kendall para las ventanas móviles				
	Z	S	var	Nivel de significancia
Media	1.66	16727	100937752	0.95202
Percentil 25	9.67	73739	58144767	1
Percentil 50	10.21	96522	89289996	1
Percentil 75	9.55	95662	100205050	1

Conclusiones

La Serie de datos de Cardellina canadensis de frecuencia de listas observadas sobre listas totales, parece finalmente no ser el mejor descriptor de la población en Colombia porque es muy suceptible a las listas vacías que existen en el inicio de la serie de datos. Es necesario buscar datos que representen de mejor manera si la población de la especie ha aumentado o disminuido en Colombia y si su época migratoria se ha visto afectada por el cambio en alguna variable climática.

El histograma de la serie y la función de distribución de probabilidad muestran que la serie esta dominada por un gran número de datos con valor 0, lo cual se atribuye a los pocos datos que se podían registrar antes del 2000 por la poca observación de la especie y porque la plataforma Ebird se creó en el 2002, antes de dicho año las observaciones del ave no se registraron con frecuencia en la plataforma.

La serie es no estacionaria de acuerdo con el análisis de la media móvil en una ventana de análisis de 24, que representa el número de datos para los meses de migración de la especie, es necesario analizar con otro método si el número de individuos ha aumentado o disminuido para cumplir con el segundo objetivo.

Las medidas de dispersión muestran que la serie de datos tiene una alta variabilidad al inicio de la serie pero ha ido estabilizándose con el paso del tiempo y con el aumento del numero de registros en la plataforma.

El resultado del test de Mann Kendall indica que si existe una tendencia en la serie de datos, es necesario explorar con otro tipo de datos y determinar con mayor certeza si al tendencia es positiva o negativa y finalmente a que se puede atribuir este fenómeno.