TRABAJO DE GRADO
Opción Seminario-Diplomado.

Aplicación de Ciencia de Datos para determinar la Eficiencia de los Métodos de Muestreo para el Orden Hemíptera (Insecta) por medio de la Metodología CRISP-DM

Corporación Universitaria Remington
Facultad de Ingenierías
Ingeniería de Sistemas

Aaron Ryan Grady
Ivonne Castaño Osorio
John Edisson Amortegui Granada
Opción de Trabajo de grado Seminario
2024

Dedicatoria

Dedico este trabajo de grado a mi novia que me ha acompañado desde el principio de la carrera, y quien ha sido mi soporte incondicional y mi guía en mis procesos académicos.

Agradecimientos

Agradezco a mi novia por todos los aportes y conocimientos que me brindó para la realización de este trabajo, además de su constante asesoramiento y acompañamiento.

Tabla de contenidos

Resumen

La ciencia de datos es un campo interdisciplinario que emplea métodos estadísticos, algoritmos y tecnologías para transformar grandes volúmenes de datos en información valiosa y accionable, mejorando la toma de decisiones y creando nuevas oportunidades en prácticamente todas las áreas y disciplinas, convirtiéndose así en una herramienta fundamental y poderosa dentro de cualquier organización. Los proyectos basados en la ciencia de datos pueden beneficiarse a partir de las metodologías de procesos empleadas, ya que éstas determinan el éxito de la investigación. Modelos de procesos como CRISP-DM pueden ser útiles y mejorados con enfoques ágiles, por lo cual es una metodología popular en la práctica debido a que es fácil de comprender, estructurada, confiable, de uso común e independiente de la industria. Teniendo en cuenta lo anterior, se propuso implementar el método CRISP-DM para analizar una base de datos de libre acceso, con el propósito de extraer información relevante a partir de un set de datos grande y complejo y contestar a una pregunta de investigación relacionada con dicho set. El presente trabajo se realizó a partir de la revisión de una base de datos del SIB Colombia y del GBIF, sobre la diversidad del orden Hemíptera (insectos) presentes en la Ecoreserva ASA La Guarupaya de Acacías, Meta. El análisis de los datos se enfocó en determinar cuál de los métodos de muestreo empleados en el estudio fue el más eficiente para la recolección, con respecto a las cantidades encontradas para cada hemíptero y el tiempo empleado por cada método. Se encontró que la eficiencia de la red de golpeteo fue mucho mejor que la de la trampa Malaise, con unos valores de 213.85 hemípteros por hora y de 0.52 hemípteros por hora respectivamente.

Palabras clave: Ciencia de Datos, Big Data, Metodología CRISP-DM, Recolección de Hemípteros, Eficiencia de Muestreo.

Pregunta orientadora de la búsqueda

Los insectos pertenecientes al orden Hemíptera comprenden organismos como los chinches, las cigarras, los salivazos, los áfidos, las escamas y la famosa mosca blanca, entre otros (Bonilla *et al*. 2023). Son herbívoros en su mayoría y están especializados en succionar los tejidos vegetales por medio de un aparato bucal modificado de tipo chupador, conocido como estilete, el cual introducen a modo de jeringa en la planta (Zhang, 2011).

A pesar de ser un grupo de insectos de gran importancia por ser reconocidos como plagas agrícolas significativas, la diversidad de especies y sus hábitos de vida complejos plantean desafíos complejos para realizar un muestreo eficaz y representativo (Schuh & Slater, 1995 citado por Sarwar, 2020). Por lo tanto, el desarrollo de estrategias alternativas para su captura, cuantificación e identificación es crucial si se pretenden mitigar los efectos económicos negativos que estos ejercen en la agricultura.

Para ello, es indispensable evaluar y comparar a partir de datos ya publicados cuál sería el método de muestreo para hemípteros que representa una mejor efectividad a la hora de recolectarlos en diferentes hábitats, bajo diversos contextos y para distintas especies. El objetivo es desarrollar un marco metodológico que permita tanto a los agricultores como a los técnicos en manejo integrado de plagas, tomar decisiones informadas sobre los métodos de muestreo más eficaces y sostenibles.

Partiendo de esto, este proyecto busca obtener información relevante y utilizable a partir de una base de datos existente y de libre acceso por medio de las fases que tiene la metodología CRISP-DM, la cual se aplicará para entender de la manera más oportuna y

óptima posible cuál es el método de muestreo ideal para el orden Hemíptera respecto a la eficiencia, la cual se basa en el tiempo empleado para la recolección, la cantidad de insectos recolectados y el hábitat en que estaban presentes.

Metodología de búsqueda de la información

El método CRISP-DM surgió en 1996 y fue desarrollado por las empresas DAIMLER-BENS Y SPSS (Schröer *et al.* 2021). Este método es un modelo de proceso independiente y ha sido indispensable para los avances en el tema de minería de datos, por lo cual fue concebido con el objetivo de brindar un apoyo de consultoría para suplir las necesidades de los clientes que solicitaban estos servicios (Wirth & Hipp, 2000).

El método se basa en una serie de procesos (seis fases iterativas) que inician desde el entendimiento del negocio hasta la implementación de la solución a una necesidad particular del cliente. La tabla 1 describe brevemente la idea principal, las tareas y el resultado de estas fases, según la guía del usuario de CRISP-DM.

Tabla 1. Descripciones del modelo de proceso CRISP-DM.

FASE	BREVE DESCRIPCIÓN
Bussines Understanding	Se analizan, desde la perspectiva empresarial, los objetivos más relevantes que pueden servir para una correcta solución a la necesidad del cliente.
Analytic Approach	Teniendo los objetivos claros, este proceso se enfoca en definir el tipo de análisis que se aplicará en el método para lograr con éxito los objetivos propuestos.
Data Requirements	Se identifica la información que se necesitará para realizar específicamente el tipo de análisis que se escogió para aplicar el método.
Data Collection	Se determina la fuente de la que se obtendrán los datos y se evalúa si estos son suficientes para aplicar el método o si, al contrario, se

	debe adicionar más información para complementar y poder así aplicar el método de forma eficiente.
Data Understanding	Se eligen los campos más relevantes de la fuente de información que se utilizarán como variables para aplicar el modelo con el tipo de análisis definido.
Data Preparation	Este paso se divide en varios subprocesos donde los datos se limpian, agrupan y transforman con el fin de llevar a cabo un correcto análisis de la información.
Modeling	En este proceso se define la técnica de modelado, se documenta dicha técnica, se definen las restricciones o los requerimientos con los que se va a trabajar el modelo, y se testea la aplicación de éste para determinar si está aplicado correctamente según todos los requerimientos que se definieron para el entrenamiento del mismo.
Evaluation	En el proceso de evaluation se parte del correcto modelado de los datos para analizar los resultados y determinar una lista de posibles acciones, lo cual permite tomar una decisión que dé solución a los objetivos propuestos.
Deployment	En el deployment se redacta un informe final que será presentado al cliente sobre los resultados y las alternativas realizables para dar solución a los objetivos propuestos en el bussines understanding. Esta tarea toma los resultados de la evaluación y determina una estrategia de implementación.
Feedback	El feedback es un proceso importante que permite hacer ajustes al modelo, según los aportes que los usuarios puedan dar una vez el proceso de deployment esté concluido. Lo anterior es con el fin de pulir detalles para llegar a resolver los objetivos propuestos de forma oportuna y acertada.

Fuente: Chapman *et al.* 2000; IBM Corporation, 2011.

Aplicación de la Metodología CRISP-DM

Entendimiento del negocio

Por medio de la recolección directa de los insectos (hemípteros en este caso), es posible realizar una identificación taxonómica de cada uno, además de tener claridad sobre cuál es su distribución geográfica. Para ello, se emplean diferentes métodos de muestreo como la jama (Red Entomológica), la trampa de luz, la trampa Malaise y la red de golpeteo, entre otros.

Cada método varía según su forma de uso y según el tiempo requerido para ejecutarse, al igual que la cantidad de veces que se determina su uso (eventos de muestreo). Es por esto que es importante identificar cuál es el método de muestreo más efectivo para aumentar significativamente la cantidad de hemípteros recolectados, teniéndose en cuenta los diferentes hábitats en los que se trabaja, siendo estos bosques de galerías, plantaciones forestales y coberturas agroecológicas principalmente.

Al cumplirse este objetivo, se aumenta la tasa de recolección de hemípteros según el mejor método demostrado, y también se esperaría que cada evento de captura tenga un mejor rendimiento. Adicionalmente, se pretende que el mejor método de muestreo sea útil para los tres hábitats, con el fin de que éste exhiba el mejor rendimiento multihábitat. Por último, también es fundamental que se dé una reducción en el tiempo de muestreo empleado, donde haya una relación inversa entre el tiempo empleado y la cantidad de insectos capturados: a menor tiempo de muestreo, mayor cantidad de hemípteros recolectados.

Analytic approach

El analytic approach determina el enfoque descriptivo para alcanzar el objetivo propuesto, definiendo las relaciones entre las variables con análisis de regresiones, covarianzas, correlación de Pearson, entre otros, además de un análisis descriptivo por medio de tablas dinámicas y gráficos de barras que expongan los resultados de estas variables relacionadas y cómo se comportan en sus diferentes definiciones para dar respuesta al objetivo planteado.

Data requirements

Los requerimientos de los datos para dar respuesta al objetivo planteado son los siguientes:

- Variables numéricas tipo entero y decimal para medir resultados individuales de cada método de muestreo en cada uno de sus registros.
- Variables tipo texto para identificar el nombre del método de muestreo y el nombre del hábitat y poder relacionar los registros de forma correcta para su posterior evaluación.

Data collection

El origen de los datos proviene de los muestreos realizados entre los meses de noviembre del 2022 y mayo del 2023 en la Ecoreserva ASA La Guarupaya (Acacías, Meta). A partir de estos datos se debieron implementar otros campos para crear las

fórmulas que permitieron la medición individual de cada registro con respecto a la eficiencia y a la cantidad de eventos de muestreo. Esta información no la proporcionó la base de datos, ya que ésta tuvo como propósito exclusivamente la identificación de los hemípteros presentes en los tres hábitats determinados por los investigadores del proyecto.

Entendimiento de los datos

Se utilizó una base de datos que fue creada a partir de los datos obtenidos en la Ecoreseva ASA La Guarupaya (Acacías, Meta), donde se recolectaron diferentes tipos de insectos pertenecientes al orden Hemíptera con diferentes métodos de muestreo (figura 1). Esta labor fue llevada a cabo por un grupo de biólogos y la base de datos contiene varios campos donde se especifica la identificación taxonómica del hemíptero, la persona que hizo la identificación, el hábitat donde fue recolectado, la fecha de recolección, el método de muestreo que se empleó, la institución a la cual pertenece el proyecto, entre otros. Se seleccionaron y describieron los campos más relevantes que se identificaron para el proceso de muestreó, los cuales se definieron así:

SamplingProtocol: Es el nombre del método de muestreo empleado para la recolección de los hemípteros.

SamplingEffort: Es el tiempo de muestreo empleado en cada evento para la recolección de los hemípteros.

Habitat: Es la zona/cobertura vegetal donde se llevó a cabo el muestreo para la recolección de los hemípteros.

IndividualCount: Es la cantidad de hemípteros recolectados para cada método de muestreo.

Database Hemiptera.csv

Origen de archivo: 1252: Europeo occidental (Windows) | Delimitador: Punto y coma | Detección del tipo de datos: Basado en las primeras 200 filas

occurrenceID	basisOfRecord	institutionCode	collectionCode	catalogNumber	
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262904	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262905	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262906	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262907	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-261856	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.			
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262910	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-261859	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262911	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-261860	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262912	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262913	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262914	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262915	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262916	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-261861	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-261862	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-261863	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-262917	https://scientific-c
IAVH:ESPECIMENPRESERVADO:BIOMONITOREO-HEMIP...	PreservedSpecimen	Ecopetrol S.A.	IAvH-E	IAvH-E-261864	https://scientific-c

Figura 1. Base de datos de los hemípteros recolectados en la Ecoreserva ASA La Guarupaya. Fuente: Ecopetrol. S.A. & Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, 2024.

Preparación de los datos

Se implementó la herramienta Excel para realizar la preparación de los datos, por medio de la cual se efectuó la conversión de la base de datos en un .csv y se refinaron los datos para facilitar su limpieza y transformación (Figura 2).

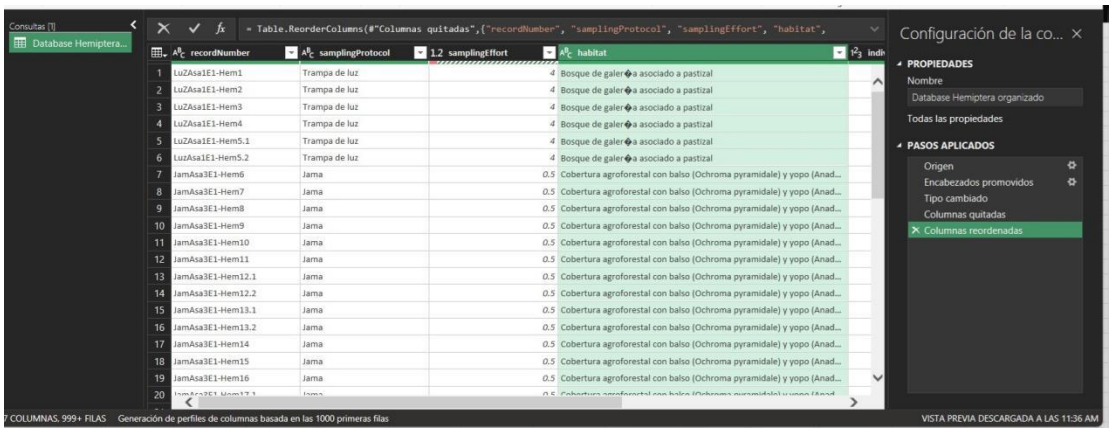

Figura 2. Transformación de los datos por medio de la herramienta de Excel. Fuente: Autoría propia.

Se observa la interfaz donde fueron transformados los datos y donde se eliminaron algunos campos para dejar solamente aquellos que se consideraron necesarios para continuar con el análisis (Figura 3).

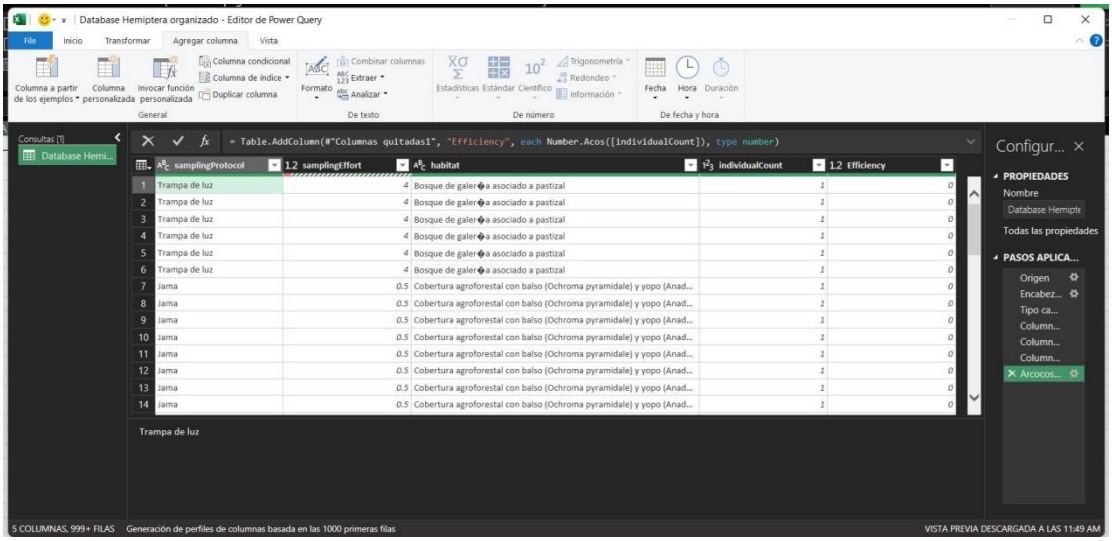

Figura 3. Eliminación de las columnas innecesarias para el análisis de los datos. Fuente: Autoría propia.

Los datos necesarios se organizaron y se adicionó la columna *eficiencia*, la cual permitió determinar qué tan eficiente fue cada método de muestreo con respecto al tiempo de muestreo empleado y la cantidad de hemípteros recolectados. Se le asignó un valor 0 para su posterior transformación (figura 4).

La transformación final de los datos incluyó los siguientes pasos:

1. Se crearon todas las columnas necesarias para el correcto análisis y la correcta aplicación del método CRISP, y se cambian todos los nombres de las columnas a español para un entendimiento mejor de cada campo y cada registro. Las columnas creadas fueron Método Muestreo (representa el tipo de trampa empleada para la recolecta); Hábitat (cobertura vegetal donde se realizó el muestreo); CódigoHábitat (código asignado a cada hábitat para identificarlos númericamente); TiempoMuestreo (el tiempo empleado en realizar la recolección); CantidadMuestreos (identifica cuántas veces se utilizó el método de muestreo en todos los hábitats); CantidadColecta (representa la cantidad de individuos recolectados por cada método y para cada hábitat) y finalmente la Eficiencia (mide el rendimiento de cada registro con respecto al tiempo de muestreo y la cantidad recolectada de hemípteros).
2. Se individualizó el tiempo de muestreo, ya que se tenía un valor total para todos los registros y esto podía generar un análisis incorrecto. Para solucionarlo, se multiplicó el tiempo de evento de muestreo por la cantidad de muestreos totales y luego se dividió por la totalidad de registros según su método de muestreo. Lo anterior dio como resultado un tiempo individual para cada registro.

3. Se individualizaron la cantidad de muestreos al multiplicar la cantidad de muestreos por método y por los tres hábitats indicados. El resultado de esto dio un total de muestreos a nivel global, el cual se dividió por la cantidad de registros según el método de muestreo. Se consiguió finalmente una cantidad de muestreos individuales para cada registro.
4. Se calculó la eficiencia de cada registro dividiendo el tiempo de muestreo sobre la cantidad de colecta, dando un resultado de eficiencia individual para cada registro.
5. Se cambió el tipo de dato de cada campo, donde al TiempoMuestreo, Eficiencia, CantidadMuestreos se les asignó un tipo de dato decimal, mientras que a las columnas MétodoMuestreo y Hábitat se les asignó el tipo de dato texto y, por último, CódigoHábitat y CantidadColecta tuvieron asignado el tipo de dato entero.
6. Los datos fueron todos transformados y se guardaron en un documento en Excel.

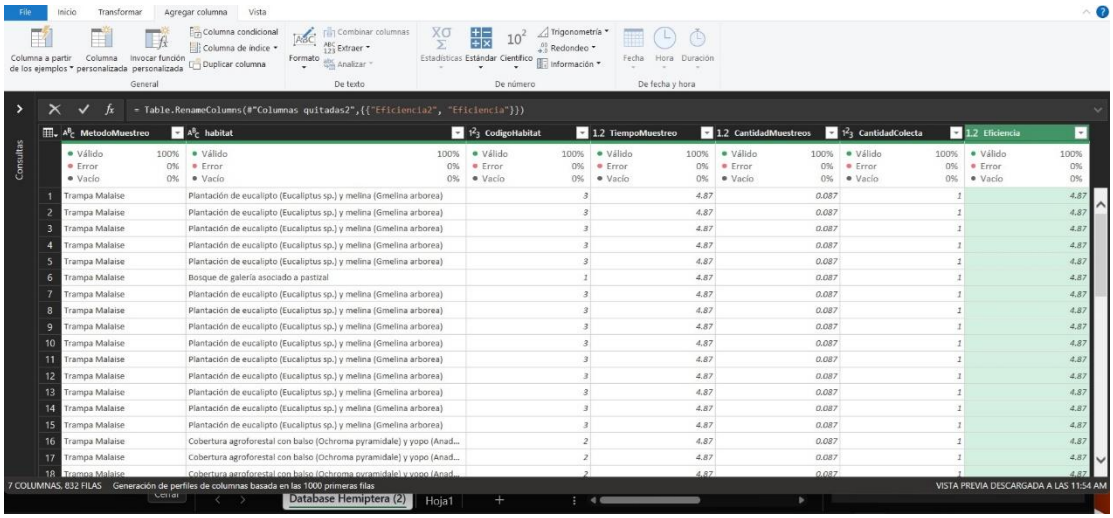

Figura 4. Transformación final de los datos y adición de la columna *eficiencia*. Fuente: Autoría propia.

Modelado

Análisis estadístico

Para realizar el modelado en el análisis estadístico, se usaron los campos TiempoMuestreo, CantidadMuestreo, CantidadColecta y Eficiencia, ya que estos son los campos con valores numéricos relevantes.

Estadística descriptiva

Se obvió en este análisis el código del hábitat, ya que sólo representa una clasificación del hábitat y no un resultado individual por registro.

En las estadísticas descriptivas en el campo tiempo de muestreo, se encontró una media de 2.82, una desviación estándar de 2.399, lo cual es óptima debido a que el valor mínimo es de 0.011 y el máximo de 4.87. Respecto a la curtosis, se observó que tiene un comportamiento estable con su valor -1.899. El coeficiente de asimetría mostró un sesgo mínimo al lado negativo del -0.32. El rango es de 4.87 con su valor mínimo de 0.011 y su valor máximo de 4.87 y un nivel de confianza de 95% (Figura 5).

CodigoHabitat		*TiempoMuestreo*		*CantidadMuestreo*		*CantidadColecta*		*Eficiencia*	
Media	2.025270758	Media	2.82933694	Media	0.079440433	Media	2.452466907	Media	2.41452492
Error típico	0.027992776	Error típico	0.08324211	Error típico	0.000308368	Error típico	0.182005002	Error típico	0.08109133
Mediana	2	Mediana	4.87	Mediana	0.087	Mediana	1	Mediana	1.62333333
Moda	2	Moda	4.87	Moda	0.087	Moda	1	Moda	4.87
Desviación estándar	0.806949719	Desviación estándar	2.39962608	Desviación estándar	0.008889333	Desviación estándar	5.246671055	Desviación estándar	2.33762536
Varianza de la muestra	0.651167848	Varianza de la muestra	5.75820533	Varianza de la muestra	7.90202E-05	Varianza de la muestra	27.52755716	Varianza de la muestra	5.46449233
Curtosis	-1.462656505	Curtosis	-1.89904025	Curtosis	-1.899040248	Curtosis	51.11277992	Curtosis	-1.93765402
Coeficiente de asimetría	-0.045939899	Coeficiente de asimetría	-0.32486314	Coeficiente de asimetría	-0.324863139	Coeficiente de asimetría	6.343890312	Coeficiente de asimetría	0.04770476
Rango	2	Rango	4.859	Rango	0.018	Rango	60	Rango	4.86973171
Mínimo	1	Mínimo	0.011	Mínimo	0.069	Mínimo	1	Mínimo	0.00026829
Máximo	3	Máximo	4.87	Máximo	0.087	Máximo	61	Máximo	4.87
Suma	1683	Suma	2351.179	Suma	66.015	Suma	2038	Suma	2006.47021
Cuenta	831	Cuenta	831	Cuenta	831	Cuenta	831	Cuenta	831
Nivel de confianza(95.0%)	0.054944955	Nivel de confianza(95.0%)	0.16338979	Nivel de confianza(95.0%)	0.000605272	Nivel de confianza(95.0%)	0.357244195	Nivel de confianza(95.0%)	0.15916818

Figura 5. Estadística descriptiva del análisis. Fuente: Autoría propia.

Coeficiente de correlación de Pearson

Los resultados arrojados por la correlación de Pearson parten de la variable dependiente que es eficiencia comparada con las demás variables independientes (figura 6). Los resultados fueron:

- En cuanto a la correlación entre la Eficiencia y el CódigoHábitat es de -0.07, indicando que hay una relación negativa muy débil.
- Entre la Eficiencia y el TiempoMuestreo, la correlación fue de 0.87, indicando esto que hay una buena correlación positiva.
- Entre la Eficiencia y la CantidadMuestreos, el valor fue de 0.87, por lo cual hay una buena correlación positiva.
- Entre la Eficiencia y la CantidadColecta, la correlación fue de -0.25, por lo tanto, se tuvo una correlación negativa débil.

		CodigoHabitat	TiempoMuestreo	CantidadMuestreos	CantidadColecta	Eficiencia
	3		*4.87*	*0.087*	*1*	*4.87*
CodigoHabitat	3	1				
TiempoMuestreo	4.87	-0.064034656	1			
CantidadMuestreos	0.087	-0.064034656	1	1		
CantidadColecta	1	0.058194384	0.01623316	0.01623316	1	
Eficiencia	4.87	-0.073775866	0.876085808	0.876085808	-0.252425099	1

Figura 6. Coeficiente de correlación de Pearson. Fuente: Autoría propia.

Covarianza

Se identificó la covarianza entre TiempoMuestreo y Eficiencia que es de 4.9, mientras que las demás variables en cuanto a valores están muy por debajo y como tal no tuvieron un puntaje de covarianza óptimo. En este análisis se evidenció que el tiempo de muestreo tiene mucha relación con la eficiencia (Figura 7).

		CodigoHabitat	TiempoMuestreo	CantidadMuestreos	CantidadColecta	Eficiencia
		3	*4.87*	*0.087*	*1*	*4.87*
CodigoHabitat	3	0.650384253				
TiempoMuestreo	4.87	-0.12384606	5.751276081			
CantidadMuestreos	0.087	-0.000458784	0.021305406	7.89252E-05		
CantidadColecta	1	0.246086877	0.204130336	0.000756194	27.49443134	
Eficiencia	4.87	-0.138999349	4.908425436	0.018183095	-3.092205514	5.45791653

Figura 7. Covarianza del análisis. Fuente: Autoría propia.

Regresión

La estadística de regresión (Figura 8) arrojó los siguientes resultados:

- El coeficiente de correlación múltiple fue de 0.96, lo cual indica que las variables entre ellas sí influyen como tal a la variable Y de eficiencia, en cuanto al método de recolección más indicado en todos los hábitats.
- El coeficiente de determinación R^2 fue de 0.92, lo cual señala que las variables sí explican el hecho de la eficiencia en el método de muestreo.
- R^2 ajustado fue óptimo, debido a que presentó un valor de 0.92 igual que el coeficiente de determinación R^2. Por tanto, todas las variables son indispensables para medir la eficiencia del método de muestreo.

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.960199658
Coeficiente de determinación R^2	0.921983384
R^2 ajustado	0.920491183
Error típico	0.940690744
Observaciones	831

Figura 8. Estadística de regresión para el análisis de los datos. Fuente: Autoría propia.

Varianza

Se visualizó un estadístico f de 2400 y un valor crítico de f de 0. Se evidenció que entre los dos valores hay una diferencia grande y el valor de f está por encima del umbral del valor crítico de f. Por lo tanto, se concluye que el modelo es óptimo con un nivel de confianza del 97 % y sirve para identificar cual es el método de muestreo más eficiente para la recolección de hemípteros en los hábitats mencionados.

En el error típico se pudo apreciar que cada variable tuvo un valor muy pequeño, teniendo el valor más alto para la cantidad de muestreos con 1.42. Esto sugiere que existe un buen error típico para todas las variables.

A partir del Estadístico T, se pudo observar que la variable tiempo de muestreo está muy alejada del valor 0 como referencia y que la variable cantidad de muestreo fue de 2.93. Las otras dos variables tuvieron valores negativos, entendiéndose que las que más tienen relación son la cantidad de muestreo y el tiempo de muestreo.

La probabilidad es óptima, ya que los valores de probabilidad son muy pequeños y por ende las variables independientes tienen un mayor efecto en la variable dependiente. Se encontró que el valor más pequeño es 8.8425E-255 de la variable tiempo de muestreo. Todos los valores anteriores están representados en la figura 9.

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*romedio de los cuadrado*	*F*	*Valor crítico de F*
Regresión	4	8648.389411	2162.097353	2443.326487	0
Residuos	827	731.8115365	0.884899077		
Total	831	9380.200947			

Figura 9. Análisis de la varianza. Fuente: Autoría propia.

Evaluación

Se utilizó la herramienta de tablas dinámicas de Excel para poder agrupar los resultados de los datos por hábitat, tipo de muestreo, cantidad de colecta y eficiencia y poder tener un análisis más completo que responda apropiadamente al objetivo propuesto en la primera etapa de la metodología CRISP-DM. También se recurrió a una gráfica de barras para interpretar los datos arrojados por la tabla dinámica.

Deployment

Esta fase del proceso no sé empleo porque no aplica para el presente trabajo.

Feedback

Esta fase del proceso tampoco fue aplicada para este análisis, ya que no hace parte de los objetivos.

Sustentación teórica de la pregunta

A partir de la información consignada en la tabla dinámica de la figura 10, se analiza que el método de la red de golpeteo tiene un tiempo de muestreo total de 3,83 horas y que la suma total de la cantidad de eventos de muestreo fue de 24 y la cantidad recolectada de hemípteros fue de 821. Para este método, la suma de la eficiencia total fue de 3,21. En cuanto a la trampa Malaise, se obtuvo que el tiempo de muestreo fue de 2.352 horas y que la suma total de la cantidad de eventos de muestreo fue de 42 y la cantidad recolectada de hemípteros fue de 1.218. La suma de la eficiencia en este caso fue de 2008,12.

Respecto a la eficiencia total, se obtuvo un valor de 213,85 hemípteros por hora para la red de golpeteo y una eficiencia significativa de 112,55 para el hábitat de bosque de galería; de 276,94 para la cobertura agroforestal y de 204,94 para la plantación forestal de eucalipto y melina. Con respecto a la trampa Malaise, se logró una eficiencia en todas las áreas de 0,52 hemípteros por hora y al sectorizar los resultados, se encontró que la eficiencia significativa fue de 0,48 para el bosque de galería, de 0,42 para la cobertura agroforestal y de 0,63 para la plantación forestal de eucalipto y melina.

Cabe resaltar que entre menor sea la suma de las eficiencias, mejor será la eficiencia del método de muestreo, ya que la suma de las eficiencias hace referencia a la eficiencia individual. En el caso de la eficiencia total, ésta representa la eficiencia por horas y entre mayor sea su valor, mayor es la eficiencia del método de muestreo.

Etiquetas de fila	Suma de TiempoMuestreo	Suma de CantidadMuestreos	Suma de CantidadColecta	Suma de Eficiencia	Eficiencia Total
Red de golpeteo	**3.839**	**24.081**	**821**	**3.216317661**	**213.857775**
Bosque de galería asociado a pastizal	0.924	5.796	104	0.8569	112.554112
Cobertura agroforestal con balso (Ochroma pyramidale	1.661	10.419	460	1.355275908	276.941601
Plantación de eucalipto (Eucaliptus sp.) y melina (Gmel	1.254	7.866	257	1.004141753	204.944178
Trampa Malaise	**2352.21**	**42.021**	**1218**	**2008.123889**	**0.517811**
Bosque de galería asociado a pastizal	857.12	15.312	415	736.6276936	0.484179
Cobertura agroforestal con balso (Ochroma pyramidale	676.93	12.093	284	594.9491795	0.419541
Plantación de eucalipto (Eucaliptus sp.) y melina (Gmel	818.16	14.616	519	676.5470159	0.634351
Total general	**2356.049**	**66.102**	**2039**	**2011.340207**	**214.375586**

Figura 10. Tabla dinámica para el análisis de las eficiencias. Fuente: Autoría propia.

Se realizó un gráfico de barras (gráfico 1), donde se logró apreciar que las barras verdes correspondientes a la cantidad de colecta presentan diferencias al comparar los resultados para cada hábitat y método de muestreo. Se observó también que la red de golpeteo presentó más registros para la cobertura agroforestal, mientras que la trampa Malaise tuvo más registros para la plantación forestal de eucalipto y melina.

Las barras azul oscuro exhibieron una diferencia abismal entre los dos métodos y se aprecia que el método de golpeteo requiere que muy poco tiempo para la recolección.

Las barras naranjas mostraron la cantidad de muestreos por zona y se observó que los resultados son muy similares, con una leve diferencia a favor del método de golpeteo porque tiene menos cantidad de muestreos.

Las barras azul clara señalaron la eficiencia individual de cada colecta y se apreció que la que tiene un menor valor para todos los hábitats fue el método de golpeteo, siendo el método que menos tiempo requirió para capturar hemípteros por hora.

Finalmente, las barras violetas representaron la eficiencia total en horas y claramente se evidenció una gran diferencia a favor de la red de golpeteo, ya que captura entre 276 a 112 hemípteros por hora, a diferencia de la trampa Malaise, la cual captura 0.5 hemípteros por hora en los diferentes hábitats.

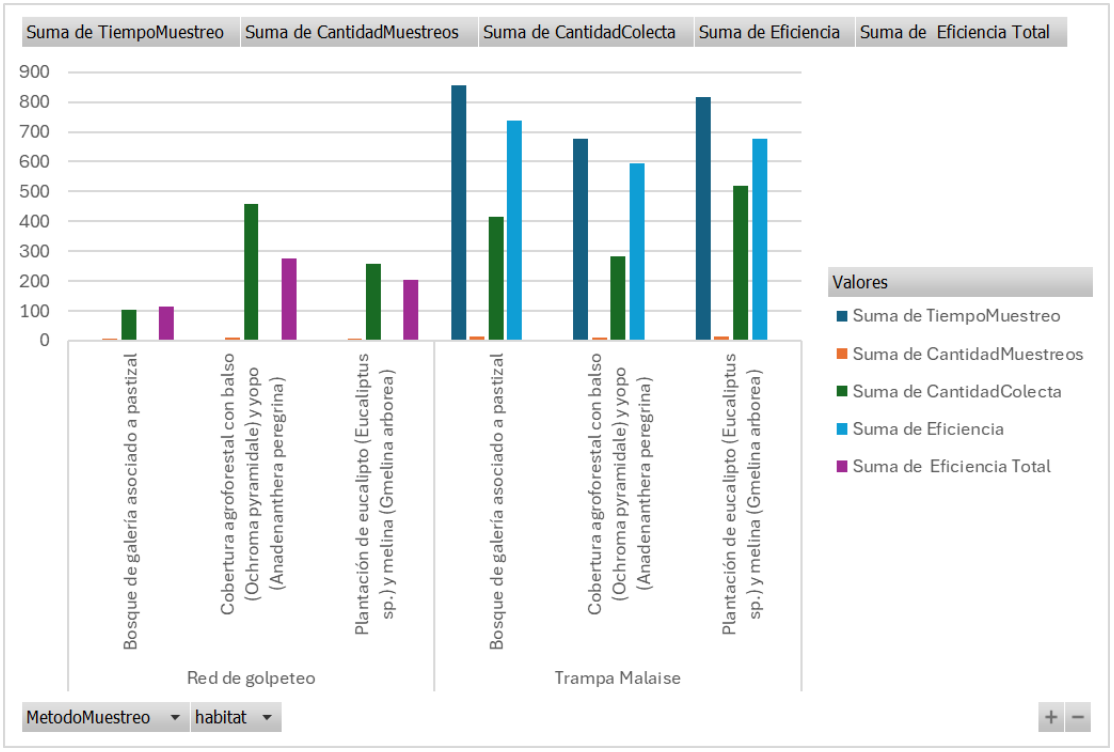

Gráfico 1. Comparación de las múltiples variables por medio de un gráfico de barras.

Fuente: Autoría propia.

.

Conclusiones

En términos de eficiencia, la red de golpeteo superó significativamente a la trampa Malaise para el muestreo de hemípteros, con una tasa de recolección de aproximadamente 213,85 individuos por hora en comparación con los 0,52 hemípteros por hora recolectados por la trampa Malaise. Para estudios que requieren una recolección intensiva, rápida y efectiva de este orden de insectos, la red de golpeteo es claramente la opción más acertada.

Los resultados obtenidos tienen sentido, si se analiza que la red de golpeteo es un método activo y por ende va a ser más efectivo al momento de recolectar insectos por hora, ya que el muestreador está yendo a buscar y capturar a los hemípteros, haciendo todo el esfuerzo de muestreo por su cuenta durante el tiempo establecido para el uso de este método. En contraste, la trampa Malaise es un método pasivo que requiere que el insecto llegue hasta ella y caiga en el contenedor con alcohol, lo cual está fuera del control del muestreador, ya que no puede intervenir para que la recolecta se haga en un menor tiempo.

La literatura sugiere que la red de golpeteo es mejor para los insectos de hábitos arbóreos, los cuales no vuelan constantemente. Por el contrario, la trampa Malaise es ideal para atrapar insectos voladores (Montgomery *et al.* 2021). Dentro del orden Hemíptera, hay muchas especies que se encuentran más establecidas en el follaje, mientras que otras son bastante voladoras y están constantemente desplazándose (Schmidt *et al.* 2019). Es importante establecer qué subgrupos de hemípteros se pretenden muestrear para hacer una escogencia asertiva del método de muestreo, u optar

por una combinación de métodos para tener una cobertura más amplia de las especies presentes en un área determinada.

Lista de referencias

Bonilla-Páez, M.M., Uribe, S.I., Forero, I.D. & González, M.A. (2024). Hemípteros de la Ecoreserva ASA la Guarupaya (Acacías, Meta). Instituto de investigación de Recursos Biológicos Alexander von Humboldt. Bogotá D.C., Colombia. 80 pp.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0. Step-by-step data mining guide.

Ecopetrol. S.A., Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (2024). Biomonitoreo de Hemiptera para la Ecoreserva ASA La Guarupaya (Acacías-Meta) - Convenio Ecoreservas. Ecopetrol S.A.. Occurrence dataset https://doi.org/10.15472/g2urco.

IBM Corporation (2011). IBM SPSS Modeler CRISP-DM Guide. IBM Director of Licensing - IBM Corporation. 46 pp.

Montgomery, G.A., Belitz, M.W., Guralnick, R.P. & Tingley, M.W. (2021). Standards and Best Practices for Monitoring and Benchmarking Insects. Front. Ecol. Evol. 8:579193.

Sarwar, M. (2020). Chapter 27 - Insects as transport devices of plant viruses. Editor(s): Awasthi, L.P. Applied Plant Virology, Academic Press. 381-402 pp.

Schmidt, O., Schmidt, S., Hauser, C., Hausmann, A., Vu, L. V. (2019). Using Malaise traps for collecting Lepidoptera (Insecta), with notes on the preparation of Macrolepidoptera from ethanol. Biodivers. Data J., 7:e32192.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). *A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, 181, 526–534.*

Schuh, R.T. & Slater, J.A. (1995). True Bugs of the World (Hemiptera: Heteroptera). Classification and Natural History. Cornell University Press, New York. 348 pp.

Wirth, R. & Hipp, J. (2000). "CRISP-DM: Towards a Standard Process Model for Data Mining." Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (4), 29–39 pp.

Zhang, Z.Q. (2011). Phylum Arthropoda von Siebold, 1848. In: Zhang, Z.Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. Zootaxa, 3148: 99–103.