CONVENIO INTERADMINISTRATIVO No. 21-095 (264 de 2021 ANH) ENTRE LA AGENCIA NACIONAL DE HIDROCARBUROS – ANH Y EL INSTITUTO DE INVESTIGACIÓN DE RECURSOS BIOLÓGICOS ALEXANDER VON HUMBOLDT - INSTITUTO HUMBOLDT

Producto N. 7

Flujos de trabajo y scripts para el manejo y análisis de datos

Diciembre de 2021

EQUIPO DE TRABAJO

Equipo central

Subdirección Proyectos Especiales y Servicios Científicos

Francisco José Gómez – Subdirector

Diana Díaz – Gerente del proyecto

Juanita Valdivieso – Gerente junior

Adriana Torres – Asistente de gerencia

Diana Herrera – Asistente de gerencia

Subdirección de Investigaciones

Sergio Vargas – Coordinador técnico

Adriana Restrepo – Coordinadora en campo

Francisco Nieto – Especialista SIG

Nicolas Corral – Gestión de la información

Análisis de datos

Susana Velásquez – Investigadora Programa de Evaluación y Monitoreo (PEM)

Bibiana Gómez – Líder línea de Análisis y Modelamiento (PEM)

José Manuel Ochoa – Coordinador PEM

Microorganismos

Coordinador de grupo

José Vladimir Sandoval Sierra – Investigador Programa de Ciencias Básicas de la Biodiversidad (PCBB)

Investigadores

Eduardo Tovar Luque – Colecciones Biológicas (CB)

Paola Montoya Valencia – PCBB

Yuliana del Pilar Castañeda Molina – PCBB

Camilo Andrés Quiroga González – PCBB

Luis Miguel Leyton Ramos – CB

Sandra Patricia Medina Saiz – PCBB

Alejandro Salazar Villegas - PCBB

Mailyn Adriana González Herrera – Líder Línea de Gestión de Recursos Genéticos -PCBB

Hidrobiológicos

Coordinador de grupo

Juan Carlos Quijano-Tristancho – Investigador PEM

Investigadores

Nédiker Stiven González Castillo – Contratista

Alejandro Villarreal Grisales – Contratista

Cristhian Camilo Castillo Ávila - Contratista

Insectos

Coordinador de grupo

Arturo González Alvarado – Investigador CB

Indiana Cristobal Ríos – Investigador CB

Investigadores

Mariposas

Cindy Flautero - Contratista

Juliana González - Contratista

Yenny Correa - Contratista

Hormigas

Luisa Arcila - CB

Gavy Mercado - Contratista

Laura Velasquez - Contratista

Cindy Quevedo - Contratista

Escarabajos

Mauricio Cobos - Contratista

David Vanegas - Contratista

Daniel Silva - Contratista

Colémbolos

Adriana Ramos – CB

Dayssy Duarte - Contratista

Nataly Torres - Contratista

Natalia Frye - Contratista

Anfibios y reptiles

Coordinador de grupo

Julián Andrés Rojas Morales – Investigador CB

Investigadores

Alejandra María Salazar Guzmán – CB

Viviana Cartagena Otálvaro - Contratista

Diego Alzate Estrada - Contratista

Aves

Coordinadora de grupo

Daniela Gómez Giraldo – Investigadora CB

Nattaly Tejeiro Mahecha – CB

Nelson Camilo Gonzáles Infante - Contratista

Sebastián Giraldo Dávila - Contratista

Yemay Toro López - Contratista

Santiago Lugo Enciso - Contratista

Investigadores

Mamíferos

Coordinadores de grupo

Andrés Julián Lozano Flórez – Investigador CB

Alejandra Niño Reyes – Investigadora CB

Investigadores

Diana Katherine Pérez Gómez - Contratista

Catalina Cárdenas González - Contratista

Javier Alejandro Salas Gordillo - Contratista

Yuli Fernanda Tique Bernal - Contratista

Nathalia Moreno Niño - Contratista

Ingrith Yuliani Mejía Fontecha - Contratista

Sonidos

Coordinadora de grupo

Daniela Martínez Medina – Investigadora PEM

Investigadores

Juan Sebastián Ulloa - PEM

Alexandra Buitrago Cardona - Contratista

Diego Alejandro Gómez Morales - Contratista

Santiago Ruiz Guzmán - Contratista

Flora

Coordinadores de grupo

Nelson Ricardo Salinas Garzón – Investigador CB

Sandra Milena Urbano Apraez – Investigadora CB

Investigadores

Angélica Guzmán Guzmán - Contratista

Angélica Ramírez Albarracín - Contratista

Daniel Arturo Franco Rodríguez - Contratista

Harold Giovanny Zambrano Ávila - Contratista

Lady Katherin Arango Gómez - Contratista

Oscar Iván Gómez Runcería - Contratista

Sebastián Alejandro Molano Cavieles - Contratista

Peces

Coordinadores de grupo

Daniel David Gutiérrez – CB

Investigadores

Daniela Bedoya Giraldo – CB

Magda Susana Bernal Sierra – Contratista

José Luis Poveda Cuellar – Contratista

Gobernanza

Emmerson Pastas – Investigador Progama de Ciencias Sociales y Saberes de la Biodiversidad

Ana Maria Roldán – Líder de la línea de investigación en Gobernanza y Equidad

1 DESCRIPCIÓN GENERAL DEL FLUJO DE ANÁLISIS

En el marco del Convenio 21-095 cuyo objeto es: “Aunar esfuerzos técnicos, financieros, jurídicos y administrativos para realizar la segunda fase del levantamiento de la línea base general de los ecosistemas y la biodiversidad para las áreas priorizadas de proyectos de hidrocarburos en la cuenca Valle Medio del Magdalena” se presenta a continuación el producto N.7 correspondiente al flujo de trabajo y scripts para el análisis de los datos.

El siguiente documento describe de forma general los flujos de análisis utilizados durante la fase de aguas altas-medias para la caracterización de la biodiversidad en el área regional del proyecto. Estos flujos comienzan con los registros biológicos, los archivos de eventos de muestro y las variables ambientales por unidad muestral. La validación de los dos primeros es realizada por la infraestructura institucional de datos (I2D) siguiendo los protocolos establecidos para registros biológicos en formato DarwinCore y los protocolos institucionales.

Los procedimientos acá descritos fueron implementados de acuerdo a las recomendaciones y flujos analíticos descritos en Legendre y Legendre (2012) que son implementados en el software R (R Core Team -2020- versión 4.0.0) según Borcard et al. (2018) y se encuentran en el repositorio institucional en GitHub en el siguiente link (Script de análisis de diversidad Fase I). La figura 1 describe de forma general el flujo de análisis.

“El proceso comienza con la curaduría adicional de los tres archivos básicos: los archivos de eventos de muestro, los de registros de muestreo y las variables ambientales. Esta curaduría básicamente verifica que los nombres de las columnas estén homologadas entre archivos, especialmente las columnas que describen los eventos de muestreo, las unidades muestrales, el esfuerzo de muestreo, los protocolos de muestreo y la identidad de las especies o grupos taxonómicos focales. Una vez los tres archivos estén curados se genera una “dataframe” para cada tipo de información y comienza el flujo analítico. Con las dataframe de registros se estiman las matrices de abundancia y presencia/ausencia. Todos los análisis que se describen a continuación se hacen tanto con matrices de abundancia como con matrices de presencia/ausencia (exceptuando las curvas de Rank-abundancia). Las funciones incluidas en el script tienen la posibilidad de estimar matrices con base en la suma de los individuos detectados, o en caso que el muestreo no garantice la independencia de los conteos de individuos, el máximo del número de individuos detectados.

Paso 1: Determinación de las unidades muestrales sin registros. Este procedimiento identifica las unidades muestrales y los eventos de muestreo que las componen en donde no se encontraron registros biológicos. En pasos posteriores del análisis estos puntos se

consideran ceros estructurales y se incorporan en las gráficas de diversidad versus variables ambientales.

Paso 2. Determinación del esfuerzo de muestreo por unidad muestral, estandarizado por protocolo de muestreo. Para cada uno de los protocolos de muestreo usado en los análisis, se estima el valor estandarizado (entre 0 y 1) del esfuerzo de muestreo para cada una de las unidades muestrales. Estos valores entran posteriormente en los modelos lineales con variables continuas. Las estimaciones de diversidad usualmente se corrigen por esfuerzo de muestreo dado los eventos de muestreo en cada unidad muestreada o el número de individuos muestreados, pero no necesariamente por el tiempo o la distancia recorrida en cada unidad de muestreo.

Paso 3. Estimación de diversidad por factor determinante utilizando técnicas de interpolación y extrapolación del paquete iNEXT. Las estimaciones de diversidad utilizaron la estimación de diversidad verdadera general. Primero se hicieron las estimaciones por método de muestreo y dentro de cada método de muestreo por alguna categoría involucrada en el diseño (Orden del drenaje o Cobertura) y por área de influencia de cada una de las plataformas. Para estas estimaciones se siguió la aproximación de Chao et al. (2014) y el paquete iNEXT (Hsieh et al. 2020) en R (R Core Team -2020- versión 4.0.0). Dependiendo de la estimación se utilizaron los datos de incidencia de especies, o los datos de abundancia. Cada una de estas estimaciones presenta tres tipos de curvas: 1) curva tipo 1 que presenta la diversidad estimada en función del tamaño muestral. 2) curva tipo 2 que presenta la cobertura de muestreo con respecto al tamaño muestral. 3) curva tipo 3 que presenta las estimaciones de diversidad en función de la cobertura del muestreo. En todos los casos se estimaron la riqueza (q=0), el índice de shannon (q=1) y el índice de Simpson (q=2). Estas estimaciones se hicieron a escala regional, por área de cada plataforma, o para cada uno de los niveles de alguna variable determinada (Orden del drenaje, cobertura, tipo de suelo etc.). Las salidas de este paso incluyen las curvas de acumulación de cada indicador de diversidad, las tablas correspondientes a las estimaciones asintóticas, las gráficas resumen por unidad muestral o evento de muestreo o por variable categórica

Paso 4. Con las matrices de diversidad se estiman las curvas de Rank-Abundancia para variables categóricas de interés (áreas de la plataforma, cobertura, tipo de suelo, etc). La salida de este paso incluye curvas de rank-abundancia por categoría y tablas con la jerarquía de cada especie usando su abundancia relativa.

Paso 5. Con las tablas de registros se realiza un análisis de ordenamiento no paramétrico. Este componente contiene varios análisis que buscan describir la relación entre la estructura

de las comunidades entre ellas y la correlación entre esta ubicación y la abundancia de algunas especies o algunas variables ambientales. A continuación se describen las diferentes salidas gráficas generadas NMDS por método. Corresponde a un análisis de ordenamiento no paramétrico en donde se representan las unidades muestrales agrupadas por el factor de variación usado en el diseño y por el método de muestreo. En las gráficas se indica el nivel de estrés alcanzado en el análisis lo que indica qué tan bien la ordenación refleja las relaciones originales entre las unidades muestrales. Los valores bajos son mejores

Paso 6. Estimación de la importancia de las especies en el ordenamiento. Relación post-hoc de la abundancia de las especies en la ordenación. Este análisis ayuda a evaluar qué especies pueden estar contribuyendo a la diferenciación entre unidades muestrales. Las salidas gráficas adicionan al NMDS anterior los vectores de correlación de las especies. Solo se muestran las especies con una correlación superior al 0,05. Este valor de p se indica en el título de la gráfica, así como el método y el grupo biológico. Como salida también se incluyen las tablas completas de la correlación entre el vector de abundancia de las especies y los vectores de la ordenación.

Paso 7. Estimación de la correspondencia entre la variabilidad ambiental y la variabilidad biótica. En esta serie de análisis se exploran las diferencias ambientales entre las unidades de muestreo y la importancia de algunas variables para estructurar las comunidades (Tabla 1). Se priorizaron variables representando presiones antrópicas o variables que reflejen un recurso para las comunidades. Estos análisis comienzan con un análisis de componentes principales sobre las variables ambientales en las unidades muestrales, seguido de un *hierarchical clustering* para determinar los grupos ambientales. Posteriormente estos grupos se grafican en las gráficas de NMDS anteriormente estimadas. De igual forma se corre un análisis post-hoc para evaluar la correlación de variables ambientales con la ordenación de las comunidades bióticas. Las salidas de este análisis incluyen los NMDS anteriores con vectores de variables ambientales significativos, las tablas de correlación de variables ambientales, las gráficas de PCA así como el resumen del análisis, y la gráfica del *hierarchical clustering*.

Paso 8. El último paso es un análisis de redundancia donde se incorporan en el análisis de ordenamiento las variables importantes detectadas en el paso anterior con el fin de determinar la importancia de forma cuantitativa y comparar sus importancias relativas. Para esto se hace un análisis de redundancia que parte de seleccionar las variables ambientales a incluir. Las salidas de este paso incluyen las gráficas de RDA con los vectores ambientales

incluidos, y un archivo con el resumen del procedimiento estadístico”.(Instituto Humboldt 2021)

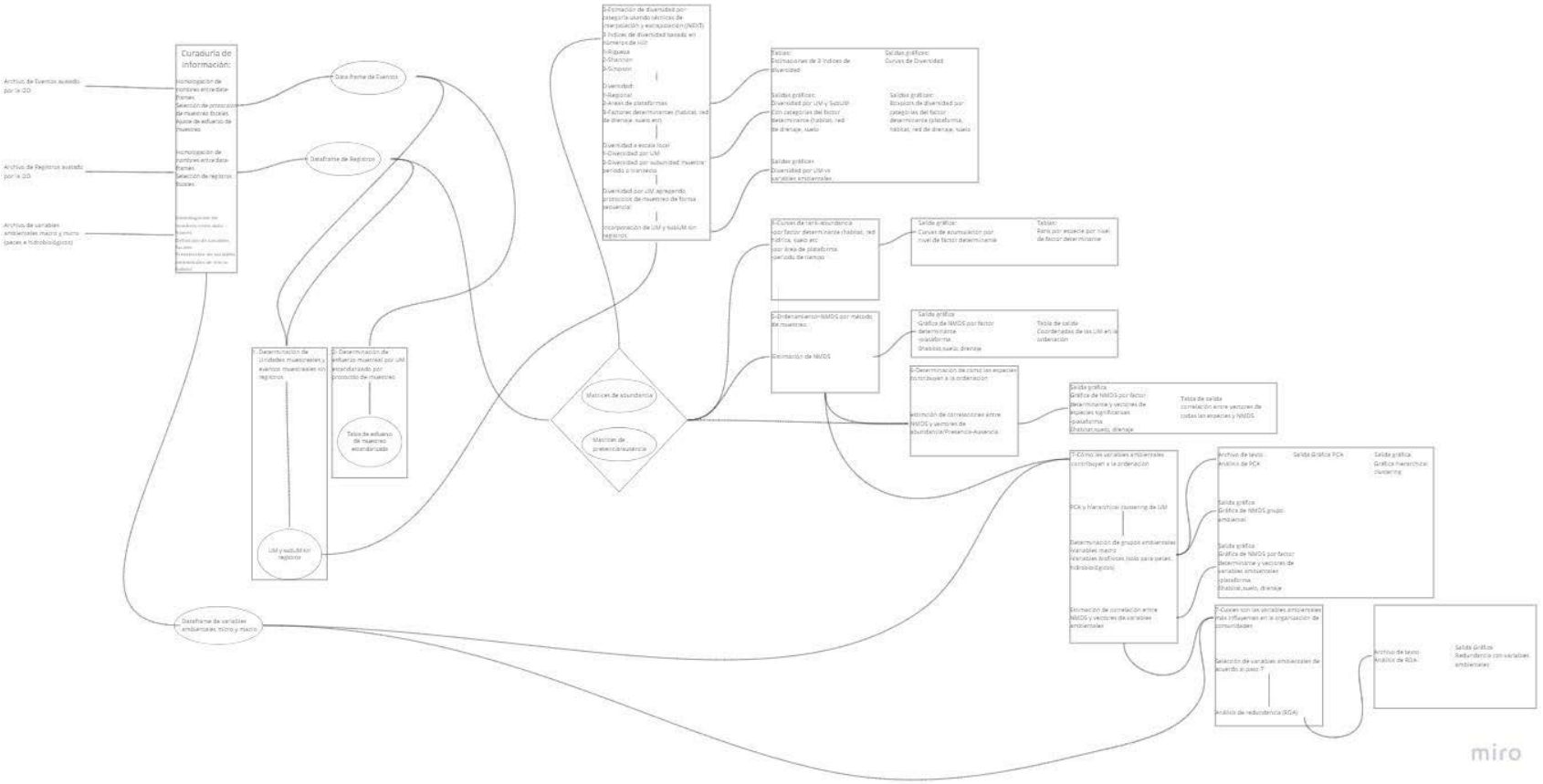

Figura 1. Flujo general de análisis. Los paso se presentan con más detalle en las tres figuras siguientes

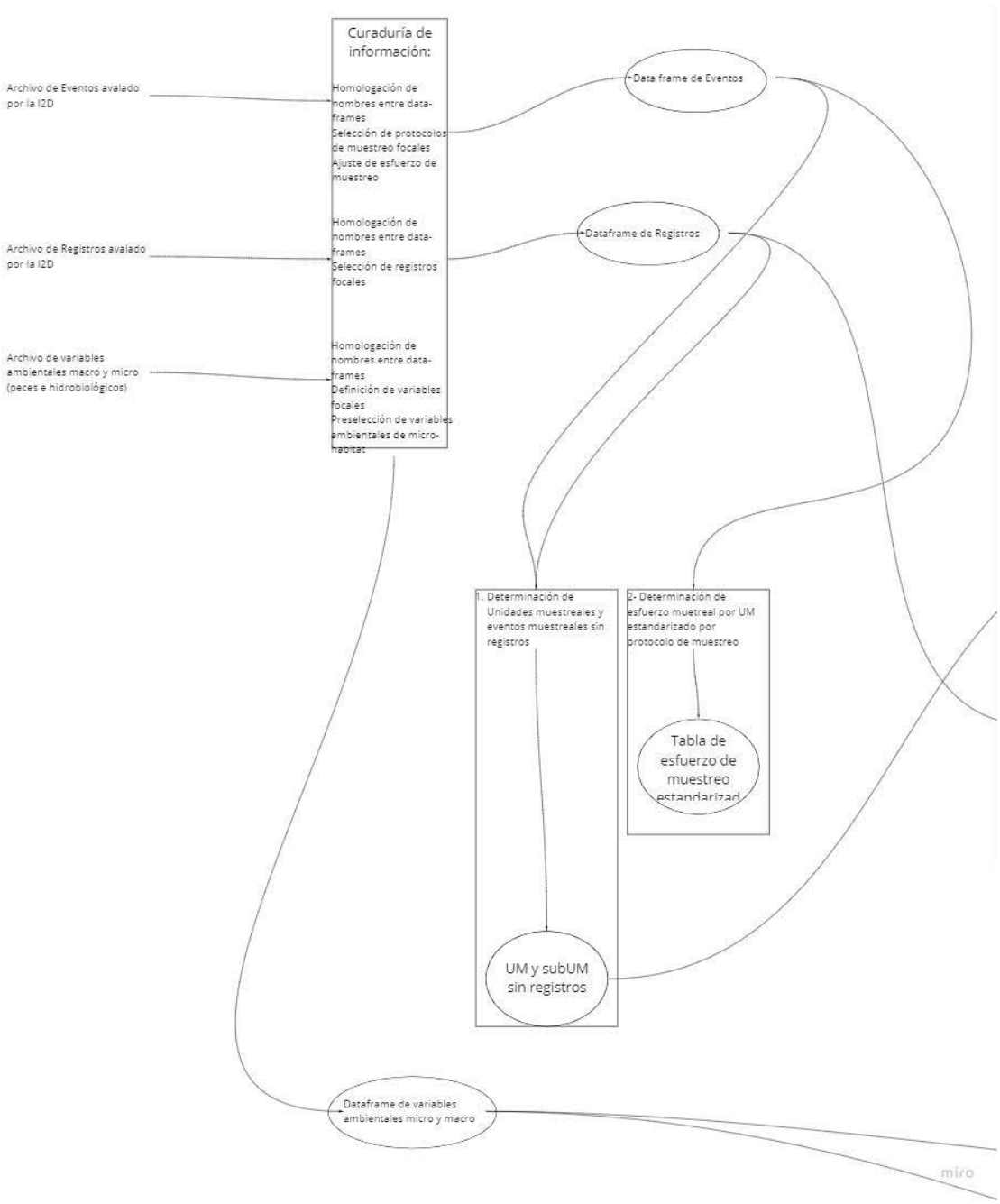

Figura 2. Detalle 1 del flujo de trabajo

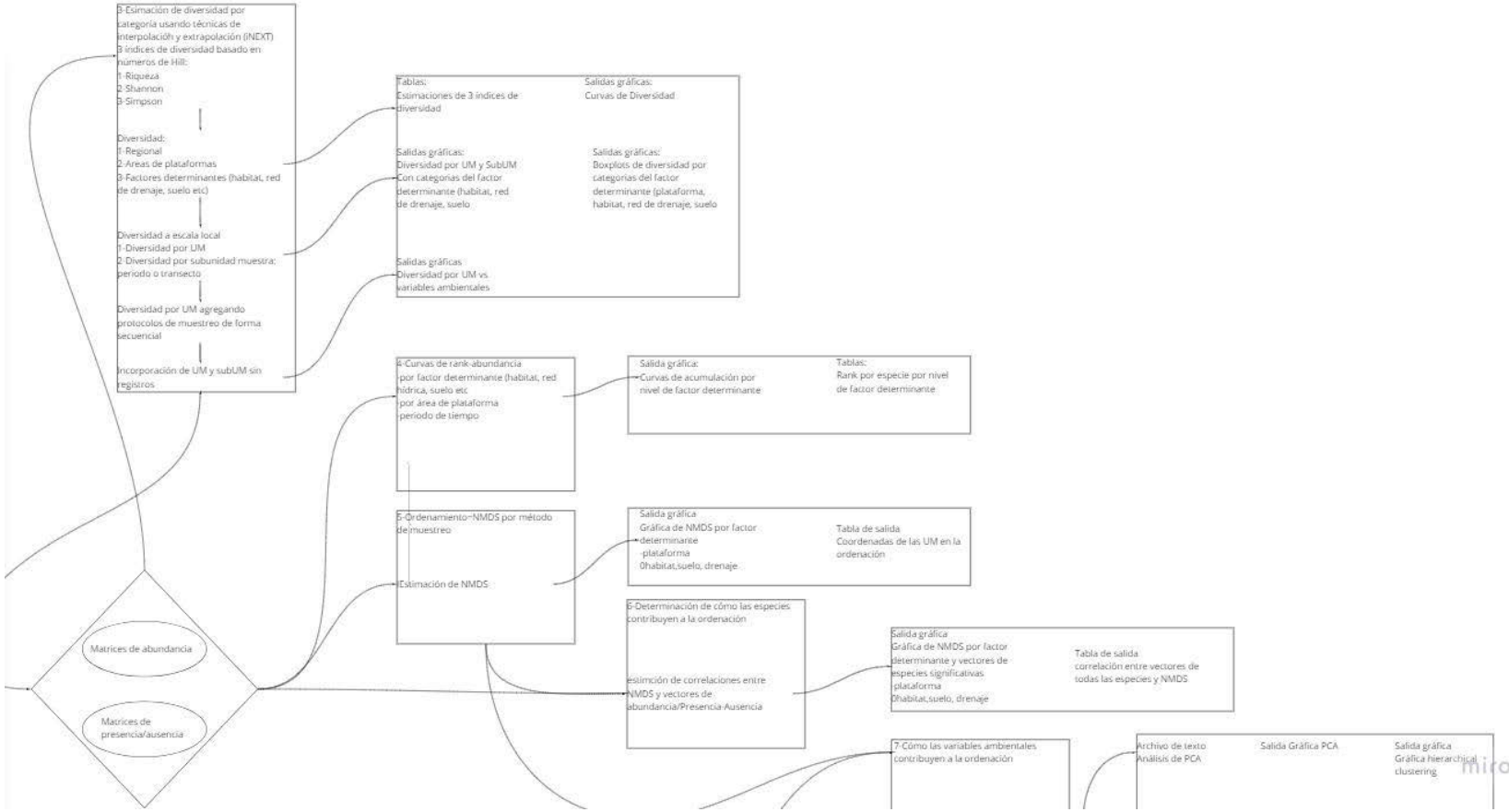

Figura 3. Detalle 2 del flujo de trabajo.

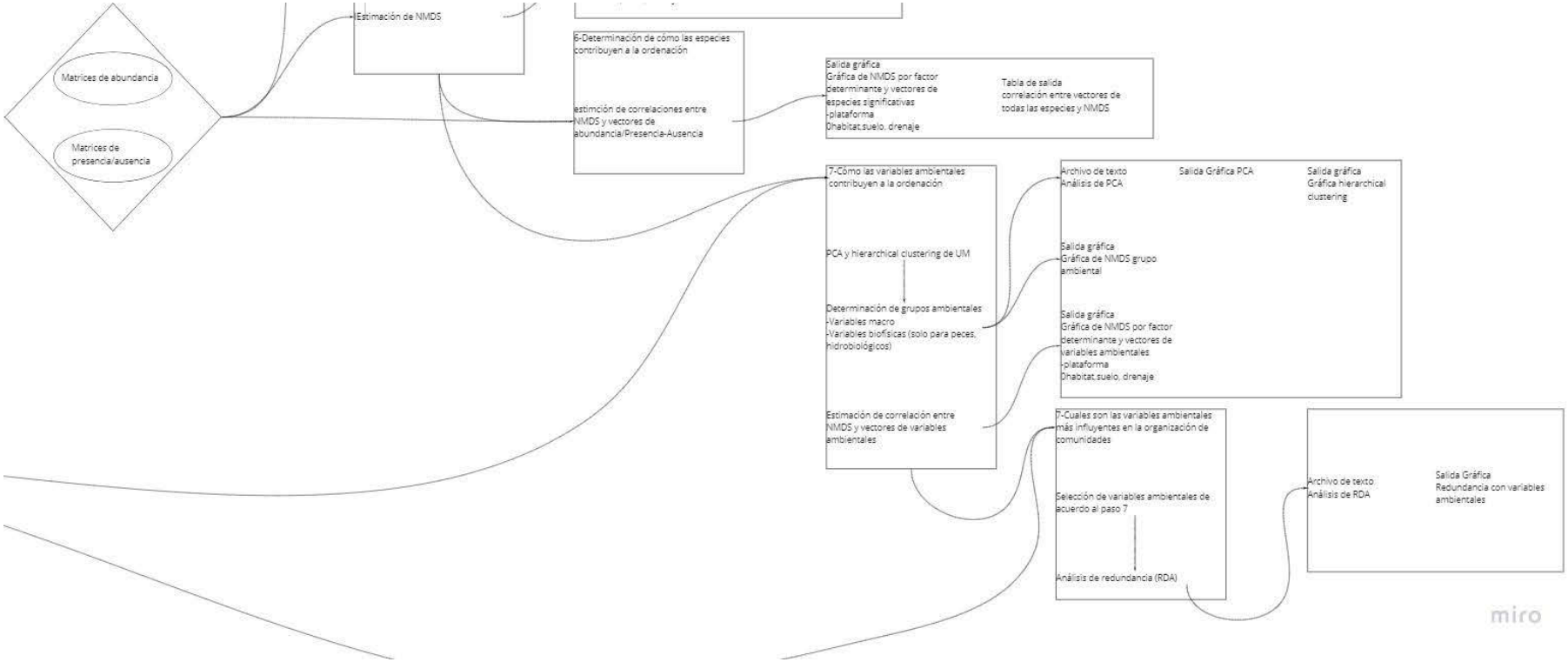

Figura 4. Detalle 4 del flujo de trabajo

Tabla 1. Variables ambientales incluidas en los análisis, Instituto Humboldt (2021)

Nomenclatura	Variable	Nomenclatura	Variable
Dis_CP	Distancia a centros poblados	Dis_Pozo	Distancia a pozo activo o inactivo
Dis_Pozoact	Distancia a pozo activo	Dis_Ferroc	Distancia a ferrocarril
Dist_Kale	Distancia al punto probable de instalación e la plataforma Kale	Dist_Plater	Distancia al punto probable de ubicación de la plataforma Platero
DisBosque	Distancia al parche de bosque más cercano	Dis_cobNat	Distancia al parche de cobertura natural más cercano
Dis_Oleodu	Distancia al oleoducto más cercano	Dis_Cienag	Distancia a ciénaga
Dis_MGSG	Distancia a grandes ríos, Magdalena o Sogamoso	Dis_Dre345	Distancia a drenajes con orden mayor a 3
UCSuelo	Unidad de Suelo	Cobertura	Cobertura según interpretación del 2020
Tam_Parch	Tamaño del parche	Plataf	Zona asignada por plataforma
Dist_PlatEf	Distancia a la plataforma asignada	Orden	Orden del drenaje
Dist_ViasPri	Distancia a vías primarias	Dist_ViaSec	Distancia a Vías secundarias

2 Productos adicionales

Tablas de composición de especies: Presenta un resumen de la composición encontrada a nivel regional por especie. Se utilizan los campos incluidos en la tabla del pool regional de especies y se listan para cada especie en qué coberturas, orden, plataforma fue encontrada.

Tablas de vacíos de información: Identifica las especies que fueron detectadas dentro del pool regional de especies, y las que no estaban representadas en ese pool (si existen).

3 Referencias

Borcard, D., Gillet, F., & Legendre, P. (2018). Numerical ecology with R (2nd ed.). Springer International Publishing.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. & Ellison, A.M. (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs, 84, 45–67.

Instituto de investigación de recursos biológicos alexander von Humboldt - Instituto Humboldt. (2021) Producto 3. Informe con los resultados Del evento de muestreo de aguas altas-medias de: plantas, mamíferos, anfibios, reptiles, aves, colémbolos, himenópteros terrestres, lepidópteros diurnos, coleópteros escarabeidos y melolóntidos, e insectos estridulantes, microorganismos, ictiofauna, macrófitas, macroinvertebrados, perifiton, fitoplancton y zooplancton en el área denominada Guane-Kalypso y Platero.

Legendre, P., and Legendre, L. (2012), Numerical Ecology (3rd. English Ed.), Amsterdam: Elsevier.

T. C. Hsieh, K. H. Ma and Anne Chao. 2020 iNEXT: iNterpolation and EXTrapolation for species diversity. R package version 2.0.20 URL: http://chao.stat.nthu.edu.tw/wordpress/software-download/.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.