Proyecto: Apoyo para la limpieza, depuración y concatenación de bases de datos de monitoreo de arrecifes del Nodo del Pacífico Tropical Este de la Red Global de Monitoreo de Arrecifes

Andrea Arriaga, Manuel Olán, Juan Pablo Quimbayo

2024-09-26

Contents

Fuente de la información: Eastern Tropical Pacific Node (ETP-Node) y colaboradores

Misión: El Nodo del Pacífico Oriental Tropical tiene como misión promover un entendimiento integral de la biodiversidad regional a través de la colaboración multidisciplinaria e internacional. Reuniendo investigadores de siete países desde México hasta Ecuador, con especialidades en ecología de arrecifes y cambio climático, buscamos generar conocimiento científico de alta calidad que sirva de base para la conservación y el manejo sustentable en la región.

Acerca de la base de datos de bentos

La base de datos de **bentos** es un registro detallado de las especies que habitan el fondo marino en diversas zonas arrecifales del Pacífico Tropical Oriental. Este registro de organismos bentónicos se realiza mediante varios métodos, descritos a continuación:

Identificación de los organismos: La identificación se lleva a cabo en campo, alcanzando el nivel más específico que el observador pueda determinar. En muchos casos, se logra identificar hasta el nivel de especie; sin embargo, en otras ocasiones, la identificación se limita a niveles más generales, como género o familia. Por lo tanto, el término "taxa" se refiere a la identidad del organismo observado, independientemente del nivel de identificación alcanzado. Esta columna puede incluir nombres de familias, géneros, especies, entre otros.

Durante todo el proceso, se mantendrá la información al nivel más detallado posible. Si es necesario agrupar especies en categorías más generales, se añadirán columnas adicionales para estas categorías, preservando siempre los datos disponibles del nivel más específico identificado en campo.

Métodos de cuantificación

Punto de intersección (cadena): Este método utiliza una cadena que se coloca sobre el fondo marino en línea recta, siguiendo el contorno de la superficie. Se registra la especie que entra en contacto directo con la cadena en intervalos de 15 cm. Tanto la longitud de la cadena como los intervalos de conteo pueden variar, y estos detalles deben especificarse en los métodos; de lo contrario, es necesario consultar al investigador. En ocasiones, también se registra la profundidad inicial y final, o una profundidad promedio. A partir del conteo de los puntos ocupados por cada especie, se estima el porcentaje de cobertura basado en el número total de puntos.

Punto de intersección (cuadrícula): En este método se emplea una cuadrícula de 1 m^2, dividida en segmentos de 10 cm con cuerda. La cuadrícula se coloca de manera consecutiva a lo largo de 10 metros, formando un transecto. Generalmente, los transectos tienen una longitud de 10 metros, aunque esta medida puede variar. En un sitio, suelen realizarse entre 1 y 9 transectos. Es posible registrar la profundidad de cada transecto o anotar una profundidad promedio para varios transectos en un mismo sitio. El tamaño de la cuadrícula debe estar especificado en los metadatos; de no ser así, se asume que es de 1 m^2. Se cuenta el número de puntos ocupados por cada especie, y el porcentaje de cobertura se calcula con base en el número total de puntos.

Porcentaje de cobertura: Este método también utiliza una cuadrícula de 1 m^2 dividida en segmentos de 10 cm. La cuadrícula se coloca de manera consecutiva a lo largo de 10 metros, formando un transecto. Aunque la longitud de los transectos generalmente es de 10 metros, puede variar. En cada sitio, se suelen realizar entre 1 y 9 transectos. Se puede registrar la profundidad de cada transecto individualmente o calcular una profundidad promedio para varios transectos en un sitio. El tamaño de la cuadrícula debe encontrarse en los metadatos; si no se especifica, se asume que es de 1 m^2. En cada cuadrícula, se estima la cantidad de cuadros ocupados por cada taxa, obteniendo así un porcentaje de cobertura con una precisión de hasta 0.25.

Formatos de hojas de datos:

Formato ancho: En este formato, cada especie ocupa una línea, y cada cuadrícula se representa como una columna. Un conjunto de columnas (cuadrículas) constituye un transecto. Es necesario convertir los datos a un formato largo para su análisis.

Formato largo: En este formato, cada fila representa una especie dentro de una cuadrícula específica en un transecto. Los valores registrados son los porcentajes de cobertura. Las especies se listan en las filas, repitiéndose tantas veces como aparezcan en distintas cuadrículas. Se deben eliminar las filas correspondientes a especies que no estuvieron presentes en una cuadrícula, manteniendo la información sobre el número de cuadrícula, el número de transecto y la profundidad. Los valores pueden ser conteos o porcentajes de cobertura.

Consideraciones generales:

- Verificar si existen densidades (valores entre 0 y 1) que deban ser convertidas a porcentajes de cobertura. Cualquier dato que se desvíe de los valores esperados debe ser señalado para su revisión.
- Identifica y gestiona adecuadamente los datos faltantes. En caso de que falten registros esenciales, considera si es posible imputarlos o si se debe dejar constancia de su ausencia en la base de datos.

Georreferenciación de los datos: Asegúrate de que los nombres de los sitios y los transectos sean consistentes en todas las hojas de cálculo. Cualquier discrepancia en los nombres puede generar errores en la vinculación de datos y en la georreferenciación.

Importante notar que en las coordenadas de los datos crudos se puede haberse cambiado el " por" con el fin de no generar problemas con el codigo.

Instrucciones para analistas de datos

Las personas interesadas en participar en la curaduría de datos para apoyar al GCRMN-ETPN deben seguir los siguientes pasos:

1. Limpiar las bases de datos crudas "Benthos_Data/Data_curation_example/data_raw/raw_data1.xlsx" y "Benthos_Data/Data_curation_example/data_raw/raw_data1.xlsx" de manera que coincida en formato y estructura con la base de datos de referencia "Benthos_Data/Data_curation_example/data_intermediate/da
2. Unir las dos bases de datos crudas.
3. Verificar que las coordenadas está en grados decimales, de lo contarrio transformarlas
4. Guardar la base de datos limpia y ordenada en formato .rds. La base de datos final corregida debe incluir las siguientes columnas, los niveles que puede tener cada columna se muestran entre paréntesis:

- **Region:** (GCal, SouthMex, Nicaragua, ElSalvador, CostaRica, Panama, ColombiaCont, Ecuador-Cont, Galapagos, Clipperton, Malpelo, CocoIs, Revillagigedos, ChileCont, ChileIslands)
- **Country:** (Mexico, Nicaragua, ElSalvador, CostaRica, Panama, Colombia, Ecuador, Chile, Francia)
- **Location**
- **Site:** (Eliminar todos los caracteres en español y reemplazar los espacios con "-")
- **Site_id:** (Las primeras tres letras de la región, país, ubicación + nombre completo del sitio, sep="_")
- **Day:** (Números con dos dígitos)
- **Month:** (Números con dos dígitos)
- **Year:** (Cuatro dígitos, por ejemplo, 2011)
- **Observer:** (Nombres originales de los datos crudos, evitando cualquier carácter en español)

- **Dataset_id:** (Nombre original de los datos crudos + rango de años + proveedor, sep="_", por ejemplo, juanjose_2001_2010_alvaradojj)
- **Sampling_Method:** (Intercept_point, Quadrant, Photoquadrant)
- **Depth:** (metros)
- **Transect:** Numerico entre 1 y 10
- **Size_transect:** (metros)
- **Transect_id:** (Site_id + Day + Month + Year + Observer + Depth + columna Transect)
- **Quadrant:** Numerico entre 1 y 50
- **Size_quadrant:** (cm)
- **Temp:** (Celsius)
- **Functional_group:** (Protocolo ETP Benthos - Guía de especies - Juan Jose Alvarado). Descrito en el documento Benthos_Data/Data_curation_example/benthos_functional_groups.csv
- **Taxa:** (Nivel taxonómico más específico posible). Si no se muestra explícita en los datos crudos, se debe generar una columna llamada "Taxa" con los mismos valores del Functional group.
- **Value:** (Valor porcentual o de punto de intersección)

Documentar los siguientes casos específicos:

- Método faltante o valores fuera de rango. Documentar cualquier caso donde falte información sobre el método utilizado o donde los valores de porcentaje de cobertura o conteos se desvíen de los rangos esperados.
- Inconsistencias entre bases de datos. Registra cualquier fila en la base de datos de bentos que no tenga coincidencias en la base de datos de sitios, lo que impida agregar las coordenadas geográficas correspondientes.
- Errores identificados. Señala cualquier otro error que el investigador pueda identificar, como profundidades mayores a 35 m, temperaturas fuera del rango de 17-33 °C, cuadrantes o transectos con valores superiores a 10, años anteriores a 2000, o especies que sean terrestres o que no pertenezcan a la región. Estos pasos asegurarán que los datos estén correctamente formateados y documentados para su posterior análisis.

Las columnas que deben estar completas son:

- Functional_group
- Month
- Year

- Lat
- Long
- Sampling_Method

Todas las demás pueden tener valores vacíos y no suele ser un problema.

Ante cualquier duda, no dude en comunicarse con Andrea Arriaga o Manuel Olán al correo:

Bases de datos crudas

Los datos crudos se ven asi

```
## tibble [340 x 29] (S3: tbl_df/tbl/data.frame)
##  $ YEAR            : chr [1:340] "year1" "year1" "year1" "year1" ...
##  $ DATE            : POSIXct[1:340], format: "1998-11-14" "1998-11-14" ...
##  $ SITE            : chr [1:340] "sitio1" "sitio1" "sitio1" "sitio1" ...
##  $ DEPTH           : chr [1:340] "DEEP" "DEEP" "DEEP" "DEEP" ...
##  $ TRANSECT        : num [1:340] 1 2 3 4 5 1 2 3 4 5 ...
##  $ % ROCK          : num [1:340] 0 0 2.5 0 0 2.7 0 1.6 0 0 ...
##  $ % BOUL          : num [1:340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ % RUBB          : num [1:340] 2.4 0 0.4 0 0 0 0 0 0 0 ...
##  $ % SAND          : num [1:340] 22.5 0 0 0 0 0 0 0 0 0 ...
##  $ % DCOR          : num [1:340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ % BRAN          : num [1:340] 11.2 41.4 44.6 66.2 78.6 25.4 65.1 46.2 46.7 73.6 ...
##  $ % MASS          : num [1:340] 16.7 29.9 27.2 12.2 1.3 0 0 0 0 0 ...
##  $ % ENCO          : num [1:340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ % FALG          : num [1:340] 0 0 1.8 0 0 0 0 0 0 0 ...
##  $ % TALG          : num [1:340] 43.5 25.2 15.9 17.8 17.2 63.8 29.4 45.7 52.9 24.4 ...
##  $ % CYAN          : num [1:340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ % CALG          : num [1:340] 0 0 0 0 0 0 0 0 0.4 0.1 ...
##  $ % EALG          : num [1:340] 3.7 3.5 7.6 3.8 3 8.1 5.5 6.4 0 1.8 ...
##  $ % OTHR          : num [1:340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ...20           : logi [1:340] NA NA NA NA NA NA ...
##  $ % Gardineroseris: num [1:340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ % Pavona        : num [1:340] 16.7 29.9 27.2 12.2 1.3 0 0 0 0 0 ...
##  $ % Pocillopora   : num [1:340] 11.2 41.4 44.6 66.2 78.6 25.4 65.1 46.2 46.7 73.6 ...
##  $ % Psammocora    : num [1:340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ lat             : chr [1:340] NA NA NA NA ...
##  $ long            : chr [1:340] NA NA NA NA ...
##  $ Dataset_id      : chr [1:340] "raw_data1_example_year1_year17_zapataf" "raw_data1_example_year1_y
##  $ MONTH           : chr [1:340] "11" "11" "11" "11" ...
##  $ DAY             : chr [1:340] "14" "14" "14" "14" ...

## tibble [221 x 28] (S3: tbl_df/tbl/data.frame)
##  $ Year            : chr [1:221] "year18" "year18" "year18" "year18" ...
##  $ Reef            : chr [1:221] "ReefA" "ReefA" "ReefA" "ReefA" ...
##  $ Site            : chr [1:221] "Site8" "Site8" "Site8" "Site8" ...
##  $ Depth           : chr [1:221] "Deep" "Deep" "Deep" "Deep" ...
##  $ Reef_zone       : chr [1:221] "Outer" "Outer" "Outer" "Outer" ...
```

```
##  $ Transect        : num [1:221] 1 2 3 4 5 1 2 3 4 5 ...
##  $ ROCK            : num [1:221] 0 0 0 0 0 0 0 0 0 0 ...
##  $ RUBB            : num [1:221] 2.2 1.3 0 0 19.7 0 0 0 0 0 ...
##  $ SAND            : num [1:221] 22.1 1.9 5.1 0 0 0 0 0 0 0 ...
##  $ DCOR            : num [1:221] 1.1 0.9 0 0 1.2 0 0 0.4 0 0 ...
##  $ BRAN_Total      : num [1:221] 35.4 38.2 38.1 72.9 45.7 30.4 13.3 29.1 65.9 26.9 ...
##  $ MASS_Total      : num [1:221] 13.3 14.8 12.9 14.8 0.8 0 0 0.9 0 1.7 ...
##  $ FALG            : num [1:221] 0 0 0 0 0 0 0 0 0 0 ...
##  $ TALG            : num [1:221] 13.3 25.8 38.3 7.8 30.8 24.2 7.9 29.1 12.3 31.5 ...
##  $ CYAN            : num [1:221] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CALG            : num [1:221] 0 0 0 0 0 0 0 0 0 0 ...
##  $ EALG            : num [1:221] 12.7 17.1 5.7 4.6 1.8 45.4 78.8 40.5 21.8 39.8 ...
##  $ ENSP            : num [1:221] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ...20           : logi [1:221] NA NA NA NA NA NA ...
##  $ Pavona          : num [1:221] 13.3 14.8 12.9 14.8 0.8 0 0 0.9 0 1.7 ...
##  $ Pocillopora     : num [1:221] 35.4 38.2 38.1 72.9 45.7 29.8 13.1 28.1 65.9 26.2 ...
##  $ Psammocora      : num [1:221] 0 0 0 0 0 0.6 0.3 1 0 0.7 ...
##  $ Total live coral: num [1:221] 48.7 53 51 87.7 46.5 30.4 13.4 30 65.9 28.6 ...
##  $ Dataset_id      : chr [1:221] "raw_data2_example_year18_year21_zapataf" "raw_data2_example_year18
##  $ Month           : chr [1:221] "11" "11" "11" "11" ...
##  $ Day             : chr [1:221] "15" "15" "15" "15" ...
##  $ lat             : chr [1:221] NA NA NA NA ...
##  $ long            : chr [1:221] NA NA NA NA ...
```

Base de datos limpia (referencia)

Los datos limpios se ven asi

```
## 'data.frame':    7973 obs. of  22 variables:
##  $ Region          : chr  "ColombiaCont" "ColombiaCont" "ColombiaCont" "ColombiaCont" ...
##  $ Country         : chr  "Colombia" "Colombia" "Colombia" "Colombia" ...
##  $ Location        : chr  "LocationA" "LocationA" "LocationA" "LocationA" ...
##  $ Lat             : num  2.57 2.57 2.57 2.57 2.57 ...
##  $ Long            : num  -78.1 -78.1 -78.1 -78.1 -78.1 ...
##  $ Site            : chr  "siteA" "siteA" "siteA" "siteA" ...
##  $ Site_id         : chr  "Col_Col_Loc_siteA" "Col_Col_Loc_siteA" "Col_Col_Loc_siteA" "Col_Col_Loc_s
##  $ Day             : chr  "14" "14" "14" "14" ...
##  $ Month           : chr  "11" "11" "11" "11" ...
##  $ Year            : chr  "year1" "year1" "year1" "year1" ...
##  $ Observer        : chr  "NA" "NA" "NA" "NA" ...
##  $ Dataset_id      : chr  "data_intermediate_example_year1_year21_zapataf" "data_intermediate_exampl
##  $ Sampling_method : chr  "intercept_point" "intercept_point" "intercept_point" "intercept_point" ..
##  $ Depth           : num  4 4 4 4 4 0.5 0.5 0.5 0.5 0.5 ...
##  $ Transect        : num  1 2 3 4 5 1 2 3 4 5 ...
##  $ Size_transect   : num  30 30 30 30 30 30 30 30 30 30 ...
##  $ Transect_id     : chr  "Col_Col_Loc_siteA_14_11_year1_NA_4_1" "Col_Col_Loc_siteA_14_11_year1_NA_4_
##  $ Quadrant        : chr  "NA" "NA" "NA" "NA" ...
##  $ Size_quadrant   : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ Temp            : chr  "NA" "NA" "NA" "NA" ...
```

```
##  $ Functional_group: chr  "roca" "roca" "roca" "roca" ...
##  $ Value           : num  0 0 2.5 0 0 2.7 0 1.6 0 0 ...
```

Es importante aclarar que tanto en este ejemplo como en el trabajo con las bases de datos reales es normal que haya información faltante. Parte del trabajo del analista es encontrar los valores indispensables que estén faltando y solicitarlos de forma ordenada y sistemática.