Protocolo para ejecutar el *script* de R para la verificación geográfica de registros biológicos en Colombia, Venezuela, Panamá, Ecuador, Brasil y Perú.

Versión 2.0 Abril de 2013

Laboratorio de Biogeografía y Bio-acustica (LABB)

Instituto de Investigación de Recursos Biológicos
Alexander von Humboldt.
Bogotá. Colombia

Cítese como:

Bello, C., O. Ramos., A.C. Moreno-Ramírez., J. Velázquez-Tibatá., M.C. Londoño-Murcia. 2012. Protocolo para ejecutar el script de R para la verificación geográfica de registros biológicos en Colombia, Venezuela, Panamá, Ecuador, Brasil y Perú. Laboratorio de Biogeografía y Bio-acustica (LABB). Instituto de Investigación de Recursos Biológicos Alexander von Humboldt. Bogotá D.C., Colombia. 23 p.

Idioma: Español

RESUMEN

Las deficiencias en los datos biológicos afecta la interpretación y por lo tanto la fiabilidad en el resultado de los análisis de biodiversidad, por esta razón al momento de usar la información se debe verificar su contenido, mediante un proceso analítico que permita evaluar calidad y entender los errores y sesgos con el fin de generar conclusiones contundentes acerca de las tendencias reales de la diversidad biológica.

El presente protocolo realiza la verificación de las coordenadas geográficas de los registros de las bases de datos biológicas para 6 países: **COLOMBIA, VENEZUELA, BRASIL, ECUADOR, PERÚ Y PANAMÁ .** Esta verificación se realiza en 5 grandes pasos, donde se prueba que la ubicación de la coordenada dentro del país, la consistencia en el departamento y municipio y la ubicación del registro , la ubicación del registro en cascos urbano, la inconsistencia del registro en cuanto a sus valor de altitud y la presencia de datos duplicados. El resultado es una base de datos depurada con los registros que tienen consistencia geográfica y varios set de datos con los registros que presentan errores en su ubicación (un set de datos por cada error).

El código para desarrollar esta verificación fue desarrollado en el lenguaje de programación R, y esta disponible en línea junto con los archivos necesarios para realizar la verificación de información.

CONTENIDO

1. INTRODUCCIÓN

Las bases de datos de registros biológicos provenientes de colecciones y museos son, en muchos casos, el único recurso documentado de registros biológico. La información provenientes de estas bases de datos tiene muchas aplicaciones incluyendo procesos de mapeo de distribución pasada, actual y futura de especies, estudios sobre tendencias y patrones de biodiversidad, evaluación de acciones de conservación y seguimiento del estado de conservación de la biodiversidad, entre otros (Margules et al. 2007, Boakes et al. 2010).

La disponibilidad de estos datos en línea, la información ambiental geográficamente explícita y el desarrollo de tecnologías computacionales para su análisis, han permitido potencializar el uso de registros biológicos en procesos de planificación de la conservación (Graham et al. 2004) generando respuestas y apoyando la toma de decisiones frente al estado y continua pérdida de la diversidad mundial. Sin embargo, si bien esta información es útil, no está exenta de problemas de calidad, bien sea por errores u omisiones en la información (Chapman 2005a). En términos generales los registros presentan tres tipos de error(Graham et al. 2004).:

- error en la identidad taxonómica,
- sesgo en la distribución de colecciones
- sesgos asociados a la presencia y ausencia de registros

Los sesgos geográficos se presentan por errores en la localización de la coordenada geográfica o bien por factores causantes de sesgos en su distribución como la cercanía a ciertas localidades como carreteras, parque naturales, ríos etc… lo que genera problemas de auto-correlación espacial. (Phillips et al. 2006, Boakes et al. 2010, Garcia Márquez et al. 2012).

Estas deficiencias en los datos afecta la interpretación y por lo tanto la fiabilidad en el resultado de los análisis de biodiversidad, por esta razón al momento de usar la información se debe verificar su contenido, mediante un proceso analítico que permita evaluar calidad y entender los errores y sesgos con el fin de generar conclusiones contundentes acerca de las tendencias reales de la diversidad biológica (Boakes et al. 2010).

Para ello, desde el área de sistematización de colecciones biológicas y desde el Laboratorio de Biogeografía Aplicada y Bioacústica (LABB) del Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Colombia, se han desarrollado estas rutinas para que la comunidad científica puede verificar la calidad de los registros biológicos, con el fin de tomar decisiones para la conservación con información de calidad.

2. VERIFICACIÓN GEOGRÁFICA DE REGISTROS BIOLÓGICO.

La metodología propuesta para hacer el análisis de verificación y vacíos geográficos de las base de datos de registros biológicos consta de 4 grandes pasos y se producen 18 subconjuntos de información que permiten la depuración de cada una de las inconsistencias (Figura 1).

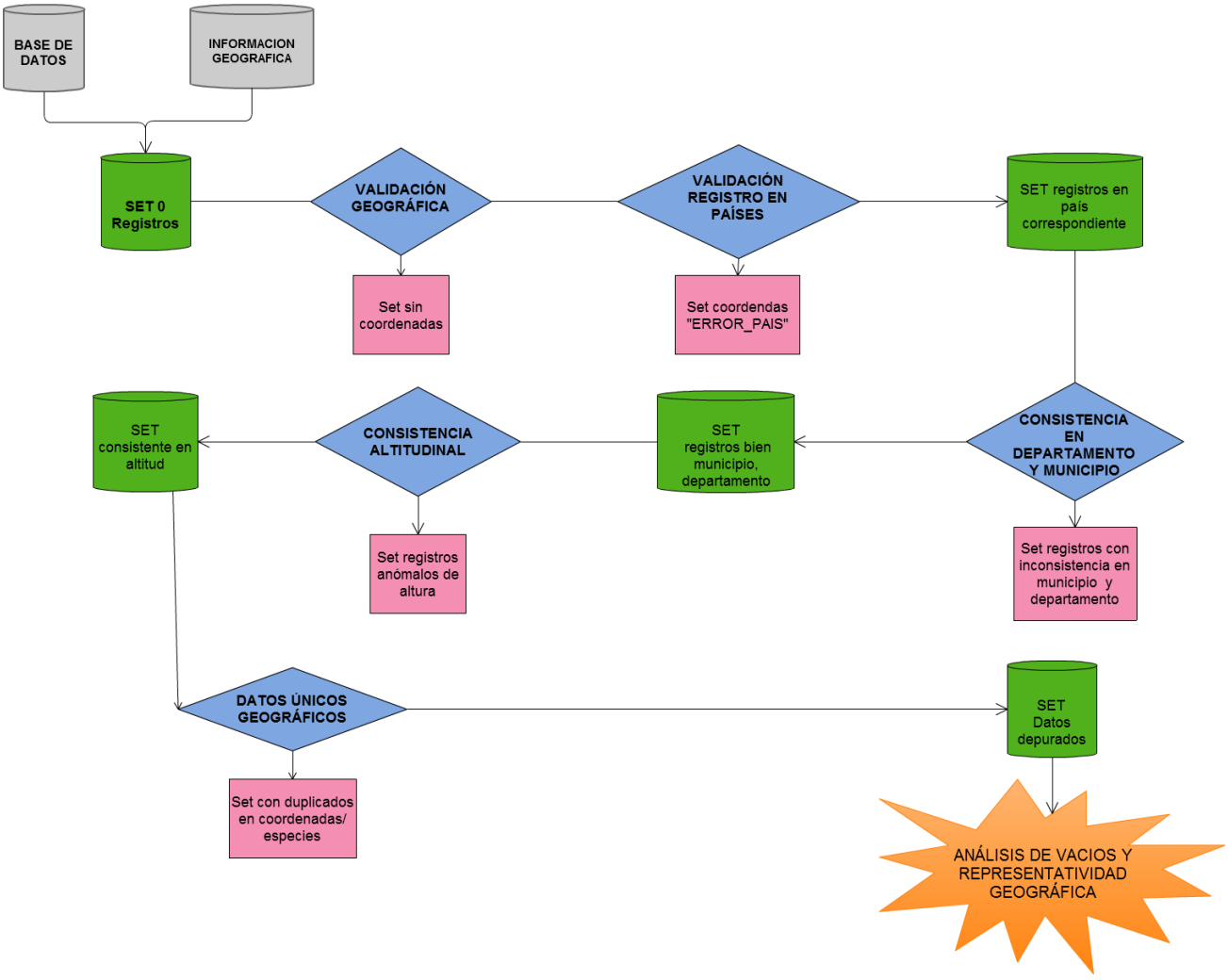

Figura 1. Diagrama metodológico del proceso de verificación geográfica

a. *Validación información geográfica:* Se descartan los registros que no posean latitud y longitud en grados decimales.

b. *Consistencia en datos de ubicación:* Se realiza una sobre posición espacial entre los puntos de los registros y las capas de países, municipios, departamentos y cascos urbanos. Con esta unión se verifica que la información que tiene originalmente el

registro en los campos de país, departamento y municipio corresponda con su ubicación espacial. De la misma forma se identifican los registros que caen en los cascos urbanos para evitar introducir registros que usen la cabecera municipal como georeferencia aunque el sitio de colecta se encuentre fuera de esta, o para evitar introducir información proveniente de decomisos, zoológicos y jardines botánicos. Según la pregunta de su investigación será pertinente o no incorporar aquellos registros que se encuentren dentro de los cascos urbanos, sin embargo para conservación *in situ* es conveniente no incorporarlos a no ser que la especie se encuentre distribuida naturalmente en áreas urbanas (e.g. aves de humedales).

c. *Consistencia altitudinal:* Para cada registro se realiza una extracción de los valores de altitud con base en el raster de altitud. A partir de las aturas registradas para cada registro se hace un análisis de datos extremos por especie. Si se encuentra algún registro cuya altitud aparece como un dato extremo (3.5 veces la desviación de la medianas) según el método de Z score se remueve dicho registro (Iglewicz 1993).

d. *Eliminación de duplicados:* Se eliminan los registros de la misma especie duplicados en la misma celda de 1km2, siendo esto práctico para procesos de modelamiento de distribución de especies.

3. INFORMACIÓN REQUERIDA

La información requerida de entrada es:

a) Base de datos con registros biológicos a analizar
b) Carpeta descomprimida “Informacion Geográfica”
c) Códigos para ejecutar en R

Base de datos con registros biológicos:

La información de registros biológicos puede ir en cualquiera de los siguientes formatos: mysql (.sql**), archivo de texto (.txt), separado por tabulaciones**. Esta información debe contener obligatoriamente, como mínimo la información pertinente a los campos definidos en la Tabla 1. El nombre de los campos y su formato debe ser idéntico al mencionado en la Tabla 1 y para los campos vacíos se debe poner “NA”. Se pueden incluir otros campos que el investigador quiera mantener en su base de datos pero no serán usados para la verificación.

Tabla 1. Descripción de los campos que debe contener la base de datos que va a ser sometida a la verificación geográfica. **El nombre de las columnas debe ser exactamente como aparece en la tabla**

Campos obligatorios	Nombre del campo en la base de datos	Tipo de campo	Descripción	Ejemplo del Campo
ID del registro	**ID**	Varchar (carácter)	Identificar único del registro	*IAvH15246*
Nombre de la especie	**Nombre**	Varchar (carácter)	El nombre de la especie debe tener separación entre género y epíteto usando el símbolo guión bajo (_)	*Atelopus_nicefori*
País	**pais**	Varchar (carácter)	Ponga la sigla según corresponda CO=Colombia EC=Ecuador PE=Perú VE=Venezuela PA=Panamá BR=Brasil	CO
Departamento	**departamento**	Varchar (carácter) Mayúscula	Describe el departamento del registro	BOYACÁ
Municipio	**municipio**	Varchar (carácter) Mayúscula	Describe Municipio del registro	TUNJA
Latitud	**latitud**	Double (Grado Decimal)	Describe latitud donde se registró la especie	4.99
Longitud	**longitud**	Double (Grado Decimal)	Describe la longitud donde se registró la especie	-75.45
Fecha_inicial	**fecha_inicial**	Varchar (carácter)	Describe la fecha del registro	2011/08/08
Localidad	**Localidad**	Varchar (carácter)	Describe la localidad del registro	San José de Maipures

Información geográfica:

Las capas geográficas de departamentos, municipios, cascos urbanos y altitud se encuentran incorporadas en la rutina para Colombia como área de estudio, utilizando la cartografía oficial generada por el IGAC (2012) a escala 1:100000. **Para acceder a ella solo es necesario descomprimir la carpeta " Información geográfica"**

Se cuenta con cartografía de Departamentos y municipios para los 6 países. Adicionalmente para Colombia se cuenta con un conjunto de 6 shapefiles históricos que

reconstruyen los cambios en la delimitación de los municipios desde 1964 hasta el 2011 (años 1964, 1973, 1985, 1993, 2003, 2011) (IGAC, *s.f.*).Todas las capas cartográficas deben tener coordenadas en grados decimales WGS84. Si el investigador desea realizar la rutina para otra área de estudio diferente debe suministrar la información cartográfica mencionada en formato shapefile y re direccionarla en el código.

4. INSTALACIÓN DE PROGRAMAS Y ARCHIVOS REQUERIDOS:

Instalación de R:

Para poder desarrollar éste protocolo debe instalar en su computador el lenguaje R que puede descargar de http://www.r-project.org/ y preferiblemente alguna interfaz gráfica de usuario para R como R studio que puede descargar de www.**rstudio**.com/.

Guardar Archivos para ejecución del código.

Descomprima la carpeta "VERIFICACION GEOGRAFICA.rar" en la capeta "Mis documentos". Es muy importante que la ruta donde se descomprime la carpeta sea "C:\Users\GIC 9\Documents\" para que el cargado de información funcione.

También **descomprima la carpeta " Infromacion geográfica.rar"** en la carpeta que acaba de descomprimir

Además **guarde su archivo de base de datos en la carpeta que acaba de descomprimir**, esta base de datos debe contener los campos especificados en la tabla1 y según la descripción del apartado "*Base de datos con registros biológicos"* de la sección anterior. La carpeta descomprimida debe quedar como en la figura 2.

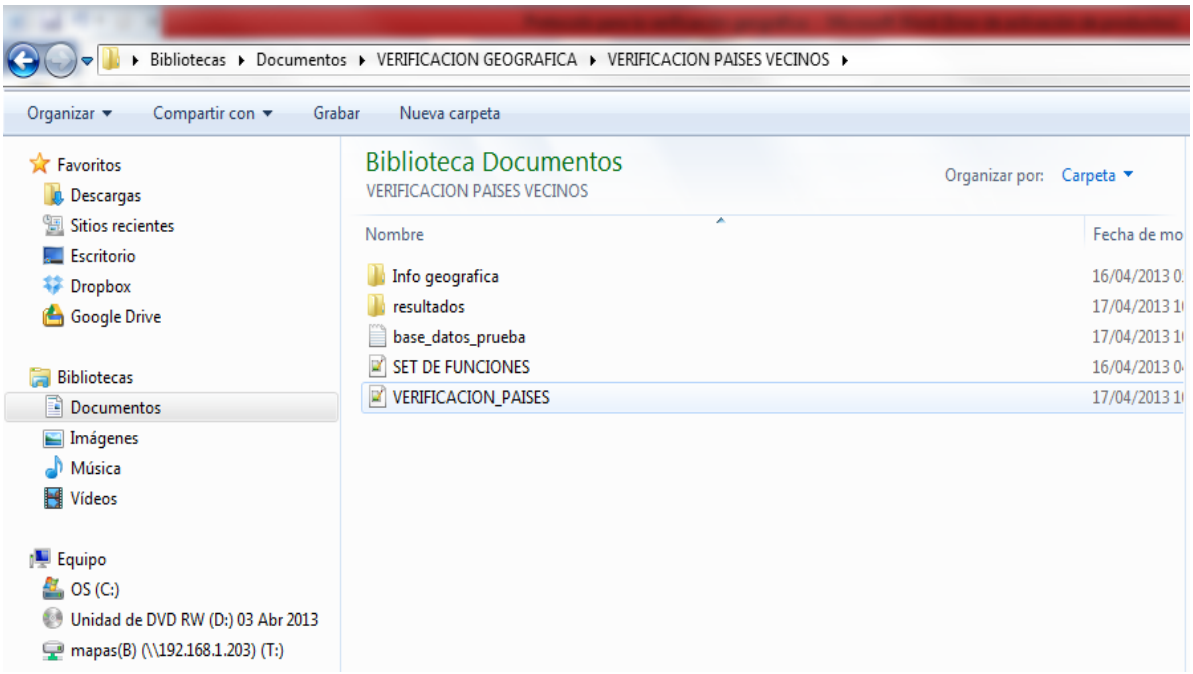

Figura 2. Imagen carpeta con los archivos para ejecutar la verificación geográfica.

5. EJECUCIÓN DEL CÓDIGO PARA DEPURACIÓN

5.1. Abrir el archivo.

Para ejecutar el código siga esta ionstrucciones y apoyese en el video ilustrativo.

Abra el archivo " 1. VERIFICACION_PAISES" en Rstudio. A partir de este momento usted ejecutara la verificación geográfica desde el R instalado en su computador.

En la figura 3 , se observa la forma en que se abre el archivo en R studio. R studio tiene la pantalla dividida en 4 secciones. La superior izquierda (1) muestra el código, la superior derecha (2) va mostrando las variables que se van creando al ejecutar el código, la inferior derecha (3) muestra los paquetes que tiene instalados y chuleados aparecen los paquetes que están activos, y la inferior izquierda (4) muestra la consola de R donde se van ejecutando cada una de las líneas.

Para poder ejecutar una línea de código ponga su cursos sobre la línea que desea ejecutar en ventana superior derecha y seleccione la **opción "Run"** o simplemente digite las teclas **" Ctrl+Enter[1]"**

[1] Este comando funciona solo para Windows.

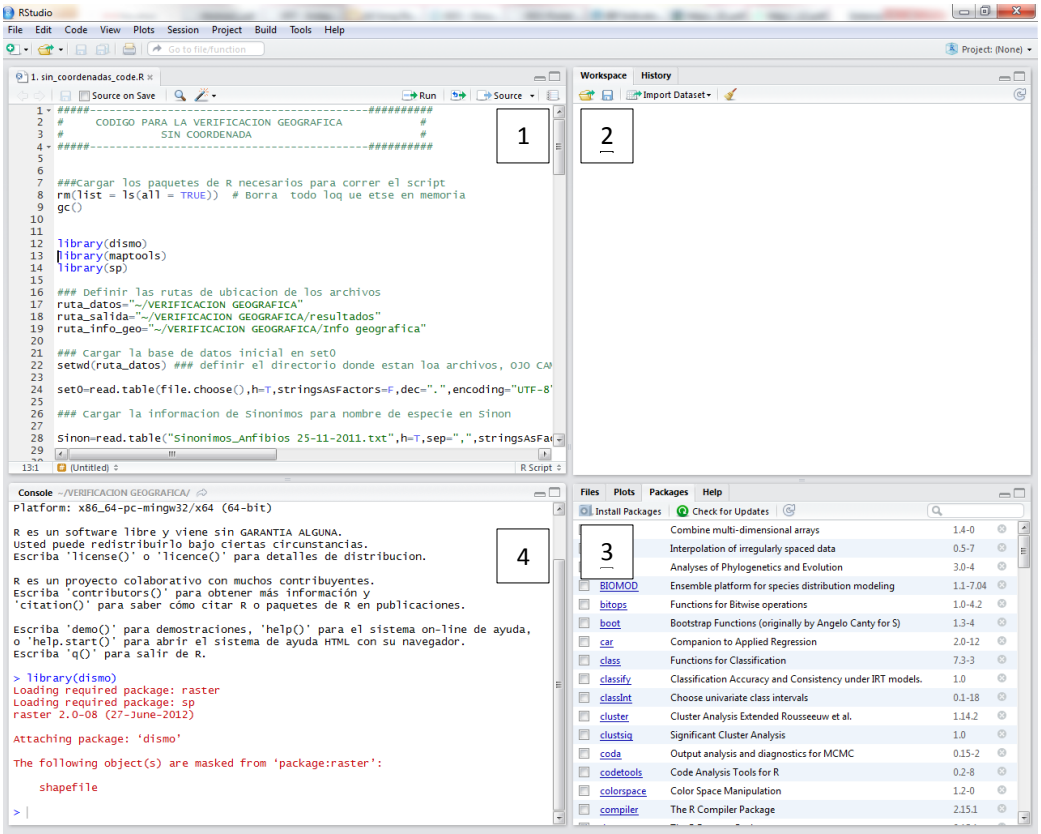

Figura 3. Imagen de Rstudio con el código abierto, se especifican las 4 ventanas que componen R studio

5.2. Cargar los paquetes

A partir de la línea 12 se cargan los 5 paquetes necesarios para ejecutar el código. Si usted aun no los tiene descargado puede darle install package en la ventana 3 (Figura 3) digitar el nombre del paquete y descargarlos. Recuerde verificar que el paquete quedo en la lista de paquetes y que está activo (chuliado) .

```
# ####### 1. CARGAR LOS PAQUTES DE R NECESARIOS ---------------
library(dismo)
library(maptools)
library(sp)
library(maps)
library("svDialogs")
```

5.3. Cargar las funciones

Ejecute las lineas 20-24, Estas líneas define la ruta donde se encuentra el archivo “SET DE FUNCIONES.R” . Verifique que después de correr las líneas de código aparezcan 7

funciones en el workspace llamadas: CARGAR DATOS, VERIFICACION_PAISES, corroboración, corroboración_dep, graficos, info_geografica y rutas (Figura 4)

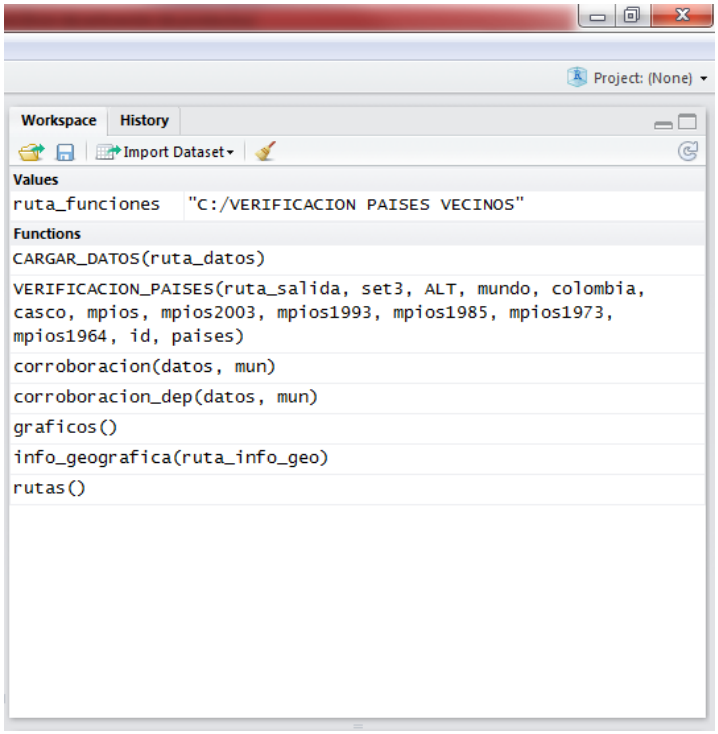

Figura 4. Set de funciones cargadas

5.4. Definir rutas de trabajo de los archivos.

Simplemente ejecute la línea :

```
rutas()
```

Esta línea desplegá 4 ventanas emergente consecutivas (Figura 5) donde se le pide que especifique la ruta donde se encuentran sus datos a analizar, la información geográfica y la ruta donde desea sus resultados.

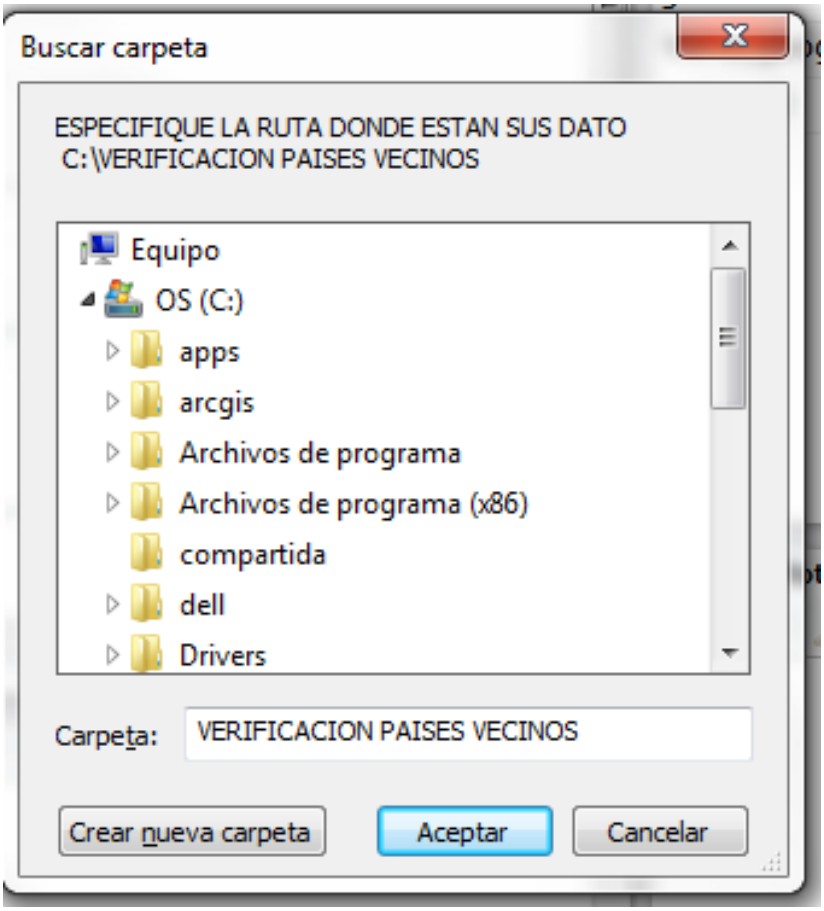

Figura 5. Ejemplo ventana emergente donde se pide la ruta para la ubicación de sus datos, información geográfica y resultados

5.5. Cargar datos

Ejecute las líneas:

```
# ### 4. CARGAR DATOS ----------------------------------------------
###      Cargar      datos      de      informacion      geografica:
info_geografica(ruta_info_geo)

# #### 5. ARREGLAR TABLA ------------------------------------------
### verificacion taxonomica ###
CARGAR_DATOS(ruta_datos)
```

El numeral 4 te carga las capas geografías Altura (ALT), casco urbano (casco), identificador de celda (id), municipios Colombia (mpios, mpios1964-2003), países con sus límites administrativos (países). Verifica que en el workspace aparezcan las capas como se observa en la figura 6.

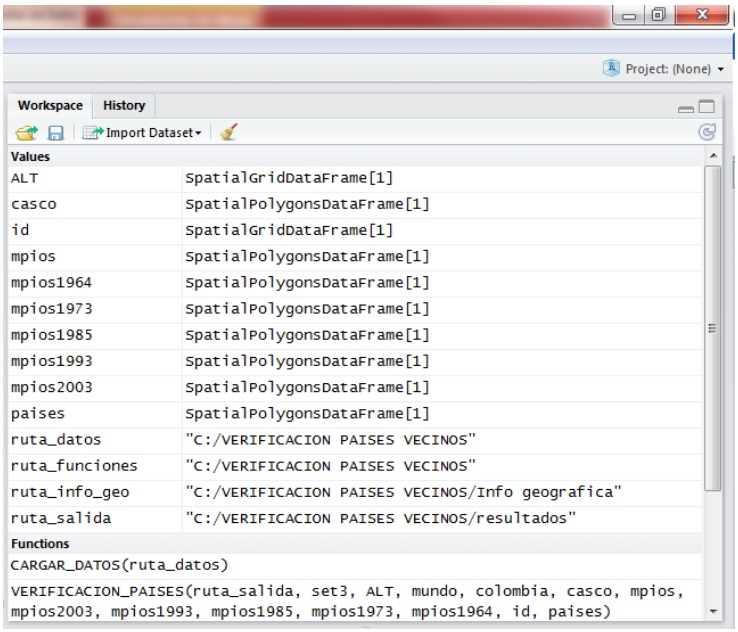

Figura 6. Imagen de capas geográficas cargadas

El numeral 5 te carga la base de datos a analizar. Se despliegan mensajes de aviso donde debes responder si o no según corresponda el caso.

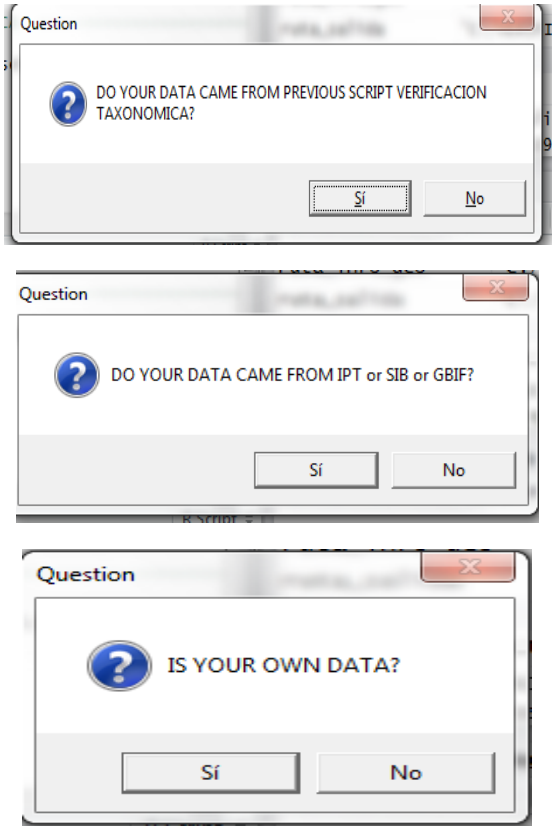

Figura 7. Selección de sus tipos de datos

Posteriormente se abrirá una ventana donde debe seleccionar su archivo de datos. Si sus datos viene de verificación taxonómica selecciones el objeto de R resultante (figura 8)

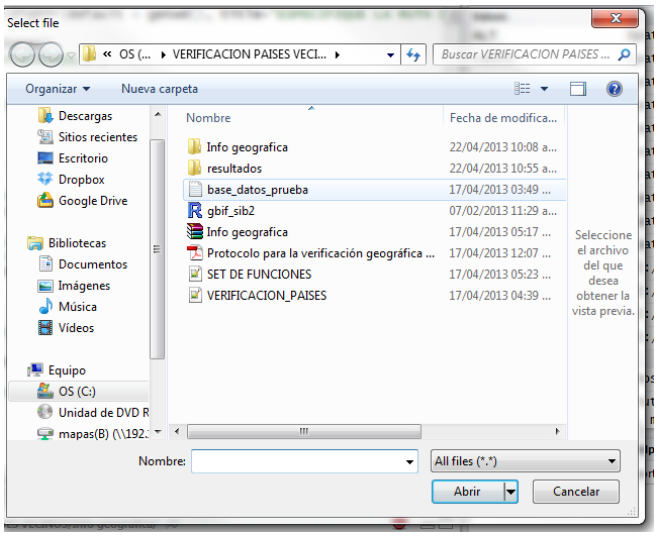

Figura 8. Ventana para selección de datos

Verifique la naturaleza de sus datos en el resumen impreso en la consola, verifique que latitud y longitud sean numéricos de no serlo habrá sus datos en Excel y modifique el tipo de celda de esas columnas.

Recuerde que de ser un archivo de IPT o SIB o GBIF o propio debe estar en formato .txt delimitado por tabulaciones, los espacios en blanco deben decir “NA” y el separador de decimal es “.”.

5.6. Ejecutar verificación geográfica

Ejecute las líneas:

```
# # ##### 6. HACER LA VERIFICACION GEOGRAFICA  --------------------

RESULTADOS=VERIFICACION_PAISES(ruta_salida,set3,ALT,mundo,colombia,casco,
mpios,mpios2003,mpios1993,mpios1985,mpios1973,mpios1964,id,paises)
```

Este proceso pude demorar varios minutos dependiendo de la cantidad de datos de su archivo al final se observan los archivos resultantes en la carpeta resultados o en la ruta donde pidió sus resultados.

5.7. Genere gráficos informes

Corra las líneas restantes que generaran las siguientes graficas resumen

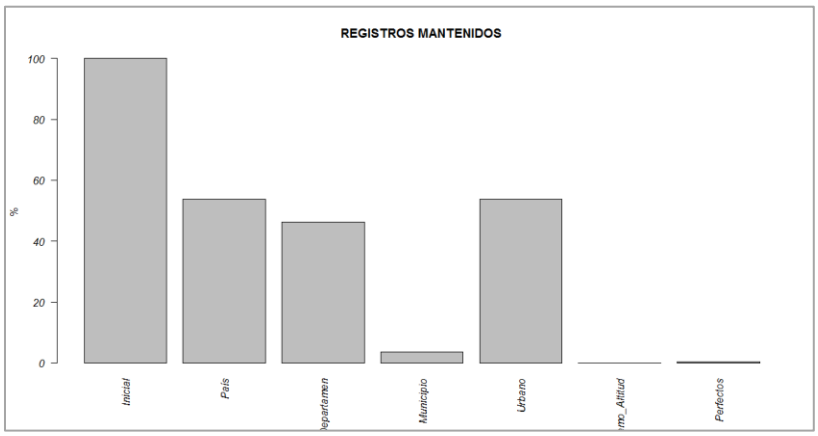

Figura 9. Muestra el resumen de los registros que pasaron cada paso

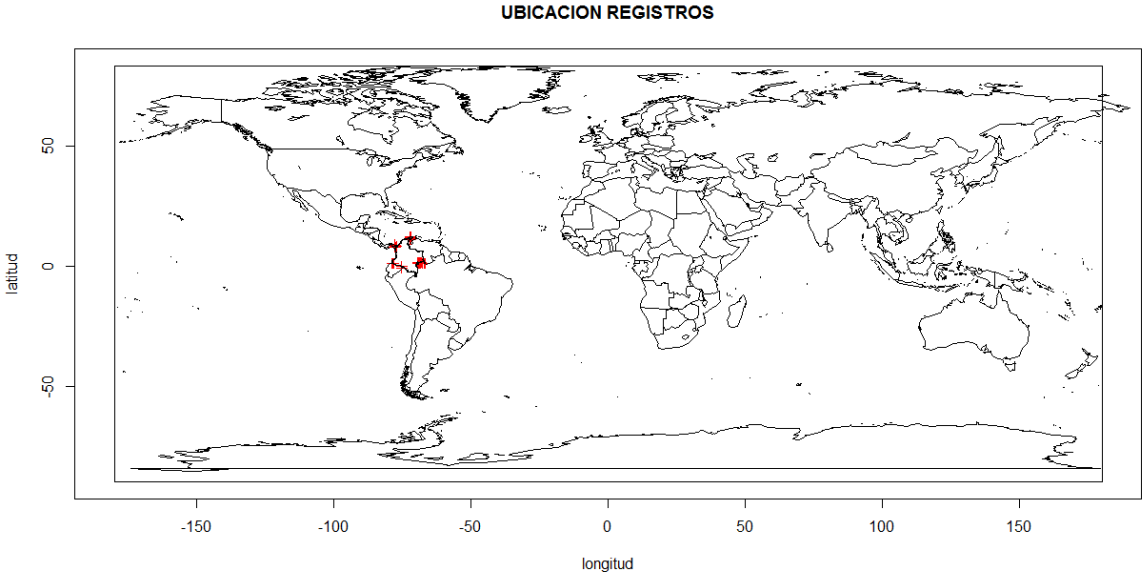

Figura 10. Muestra la ubicación inicial de los registros

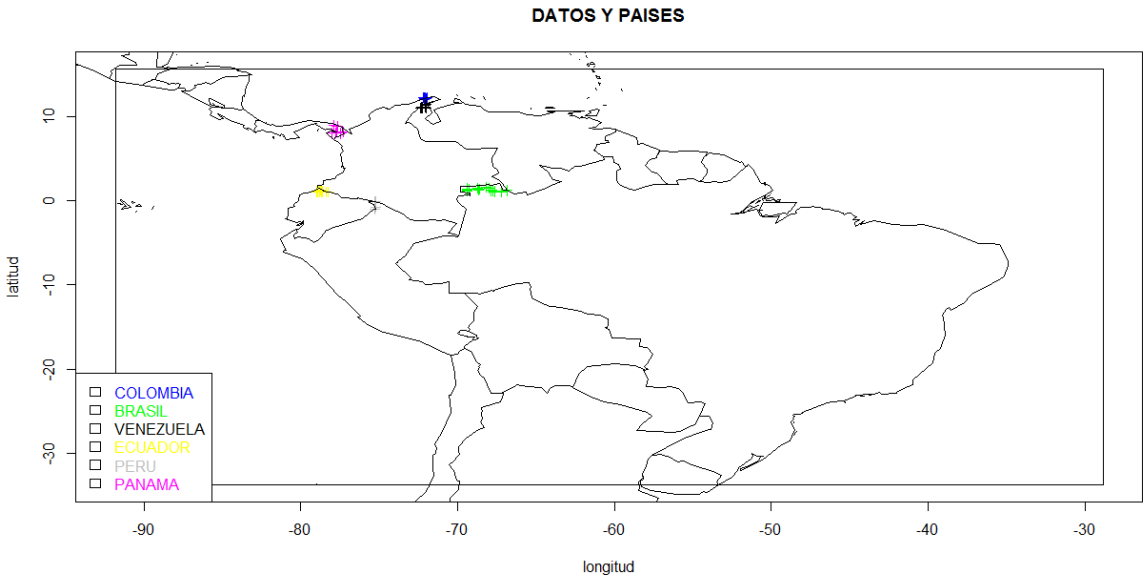

Figura 11. Muestra la ubicación inicial de los registros según el país

6. CONTACTO.

Si algún proceso de la ejecución del código tuvo algún inconveniente puede contactarnos al correo lbello@humboldt.org.co con mucho gusto ayudaremos a solucionar sus inconvenientes.

7. LICENCIA

8. BIBLIOGRAFÍA

Boakes, E. H., P. J. K. McGowan, R. A. Fuller, D. Chang-qing, N. E. Clark, K. O'Connor, and G. M. Mace. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS biology **8**:e1000385.

Chapman, A. D. 2005a. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Garcia Márquez, J. R. G., C. F. Dormann, J. H. Sommer, M. Schmidt, A. Thiombiano, S. S. Da, C. Chatelain, S. Dressler, and W. Barthlott. 2012. A methodological framework to quantify the spatial quality of biological databases. Biodiversity and Ecology **4**:25-39.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution **19**:497-503.

IGAC (s.f.).Limite municipal de los años 1964, 1973, 1985, 1993, 2003, 2011. Bogotá. Colombia.

IGAC .2012. Base cartográfica integrada a escala 1:100000. Bogotá. Colombia.

Iglewicz, B. H., D. . 1993. How to Detect and Handle Outliers.

Margules, C., S. Sarkar, and C. Margules. 2007. Systematic conservation planning. Cambridge University Press.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling **190**:231-259.