Instituto de Capacitación en Clima y Salud - Mercosur

Control de Calidad de los Datos

Rémi COUSIN, DL
IRI

Objetivos

- Asegurarse de que las variables climáticas y sanitarias, y los dominios temporales y espaciales estén bien definidos
- Validar la calidad de las bases de datos al aplicar análisis preliminares utilizando una notación de las retículas apropiada a la Data Library del IRI

Plan

- Introducción
- Cargar un archivo Excel
- Añadir Metadatos (variables independientes y dependientes)
- Diagnosticar Problemas Comunes en los Datos (coherencia y uniformidad)
- Resumen

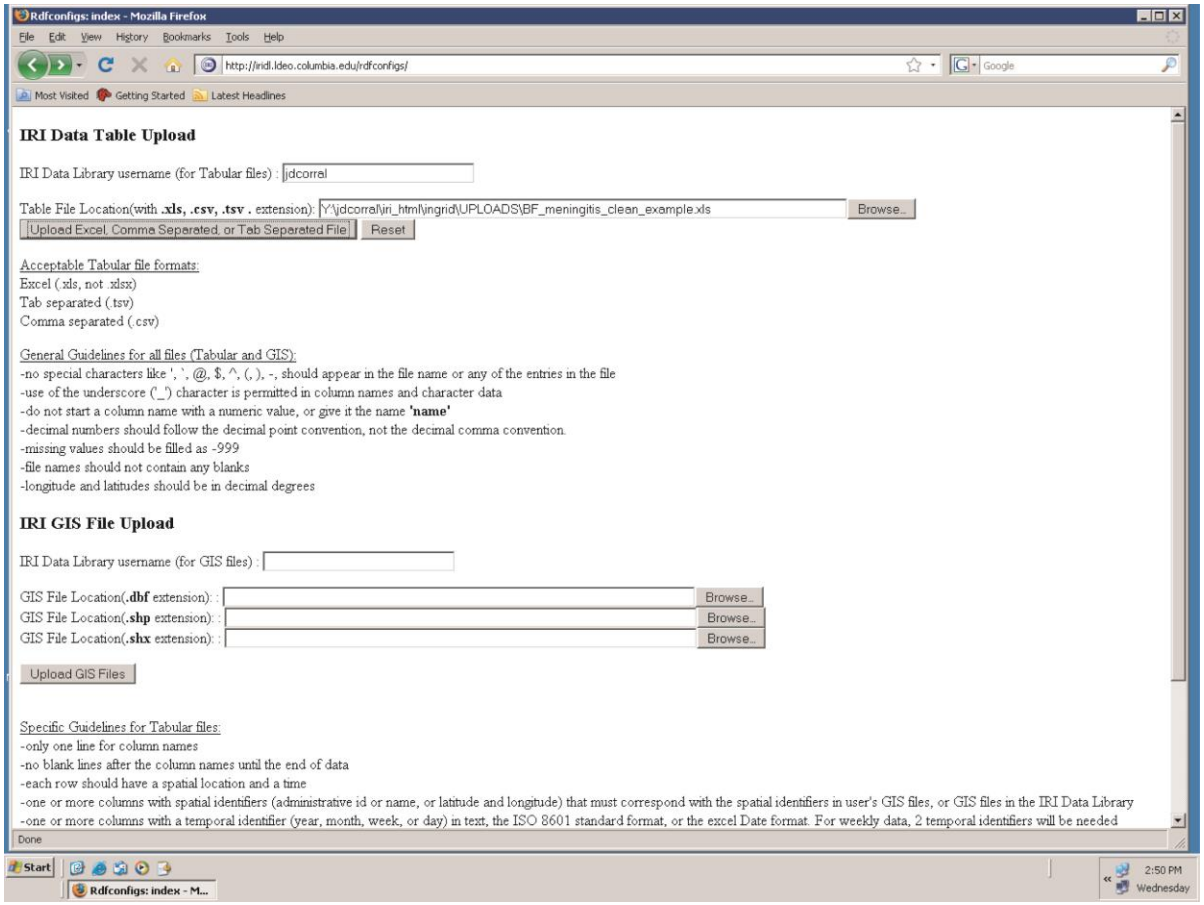
Datos de Malaria

Cargamento

- Localizar archivo Excel o shapefile
- Explorar y exponer la estructura de la tabla
- Corregir/mejorar las descripciones de los datos

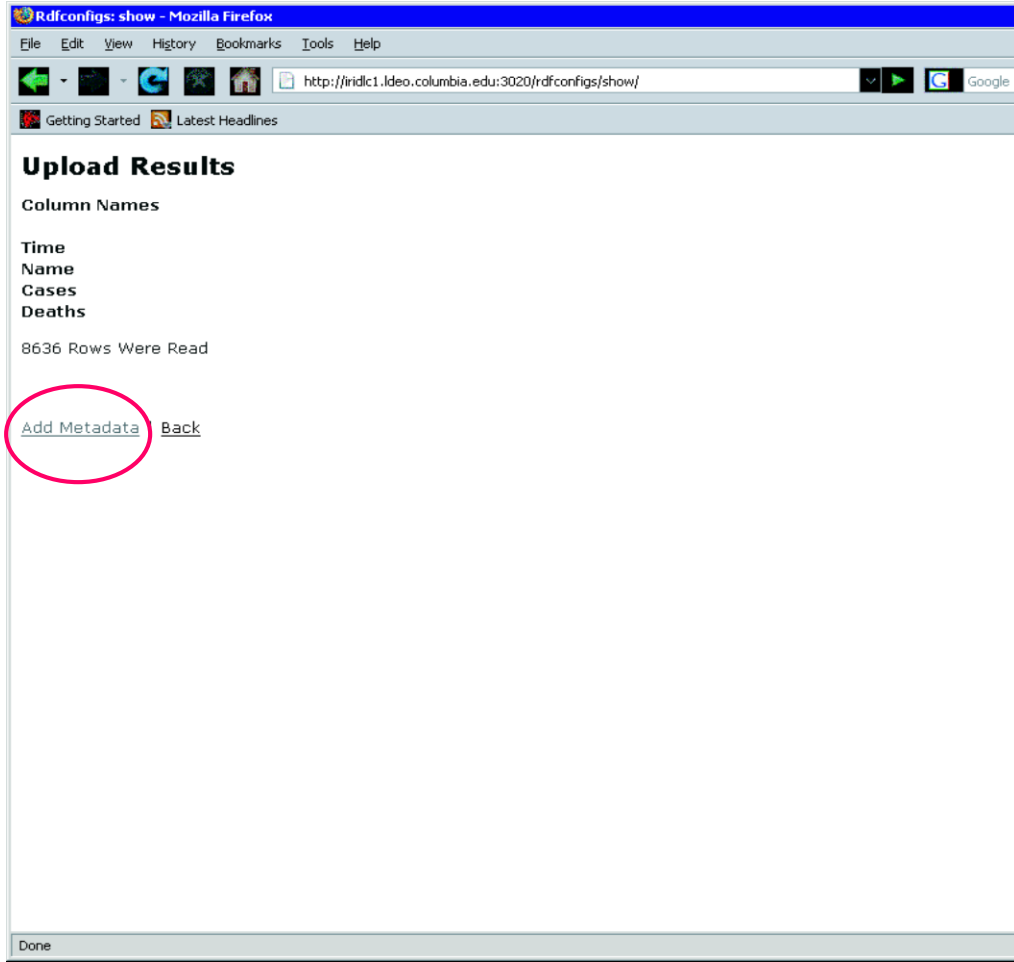
Cargar un Archivo Excel

http://iridl.ldeo.columbia.edu/rdfconfigs/

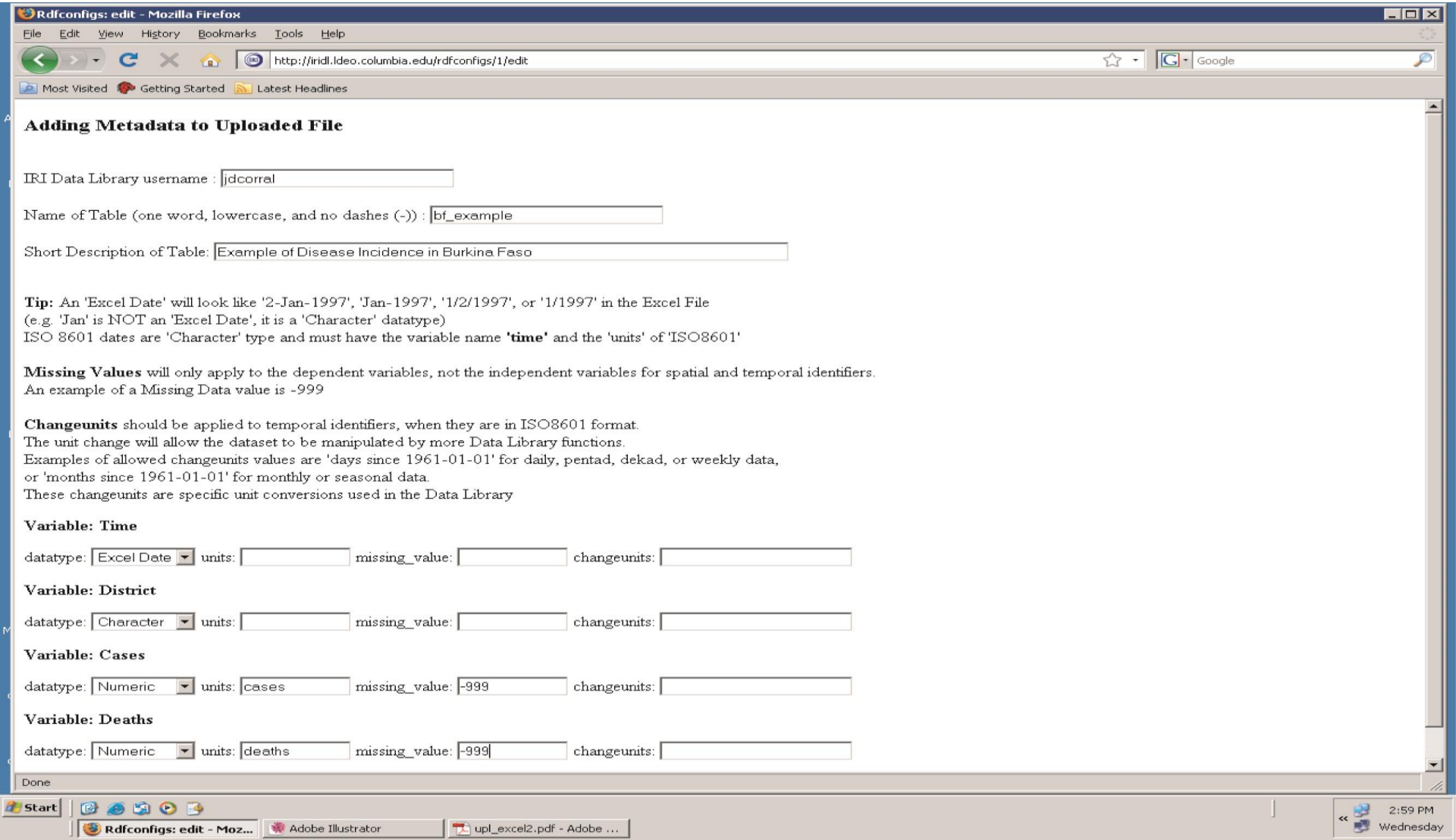


Respuesta al Cargar del Archivo Excel

El botón ‘Add Metadata’ permite de continuar hasta la página siguiente donde se añade información a propósito de la base de datos.



Añadir más Metadatos



Página para Metadatos

Esta página permite añadir más información. La parte superior permite añadir una descripción de la nueva base de datos. Información para cada columna puede ser añadida también: asegurarse de que las columnas estén reconocidas como fechas o valores numerales es particularmente útil.

Variables Independientes: Tiempo y Espacio

Al final los datos serán considerados como dependientes del tiempo y del espacio (a lo menos). Seguramente el archivo Excel tiene información a propósito del tiempo y un tipo de indicación espacial (como distrito o estado).

Tiempo

Idealmente, una de las columnas indica el tiempo, con un formato de fecha Excel estándar o ISO 8601. En este caso, la columna está indicada como una variable independiente, y la herramienta de cargamento extrae una retícula de tiempo ordenada. Columnas múltiples para describir el tiempo es aceptable también (año, mes, día)

Describir el Tiempo Precisamente

- Empiezo y fin, es decir que **January 2011** indica el mes entero, o que **1 January 2011** indica el día entero
- Intervalo de tiempo (días, semanas, meses, años)
- No es siempre posible de definir precisamente
- Ejemplo: ‘weekly sea surface temperature data’

Espacio

Idealmente, una de las columnas del archivo Excel indica una entidad espacial, es decir que cada valor de los datos tiene un identificador espacial único. En este caso, la columna está indicada como una variable independiente, y la herramienta de cargamento extrae una retícula para el dominio espacial.

Variables Dependientes

Las variables dependientes son los datos que analizar, y pueden ser dadas en columnas en el archivo Excel. En este caso, la página ‘Add Metadata’ lista las columnas, y información descriptiva adicional puede ser añadida.

Diagnosticar Problemas Comunes en los Datos

Los nombres de las entidades espaciales deben ser coherentes con sus mismos y con sus homólogos en los shapefiles.

Este tipo de problemas puede ser detectado fácilmente después de cargar una versión preliminar de los datos, y pues corregido.

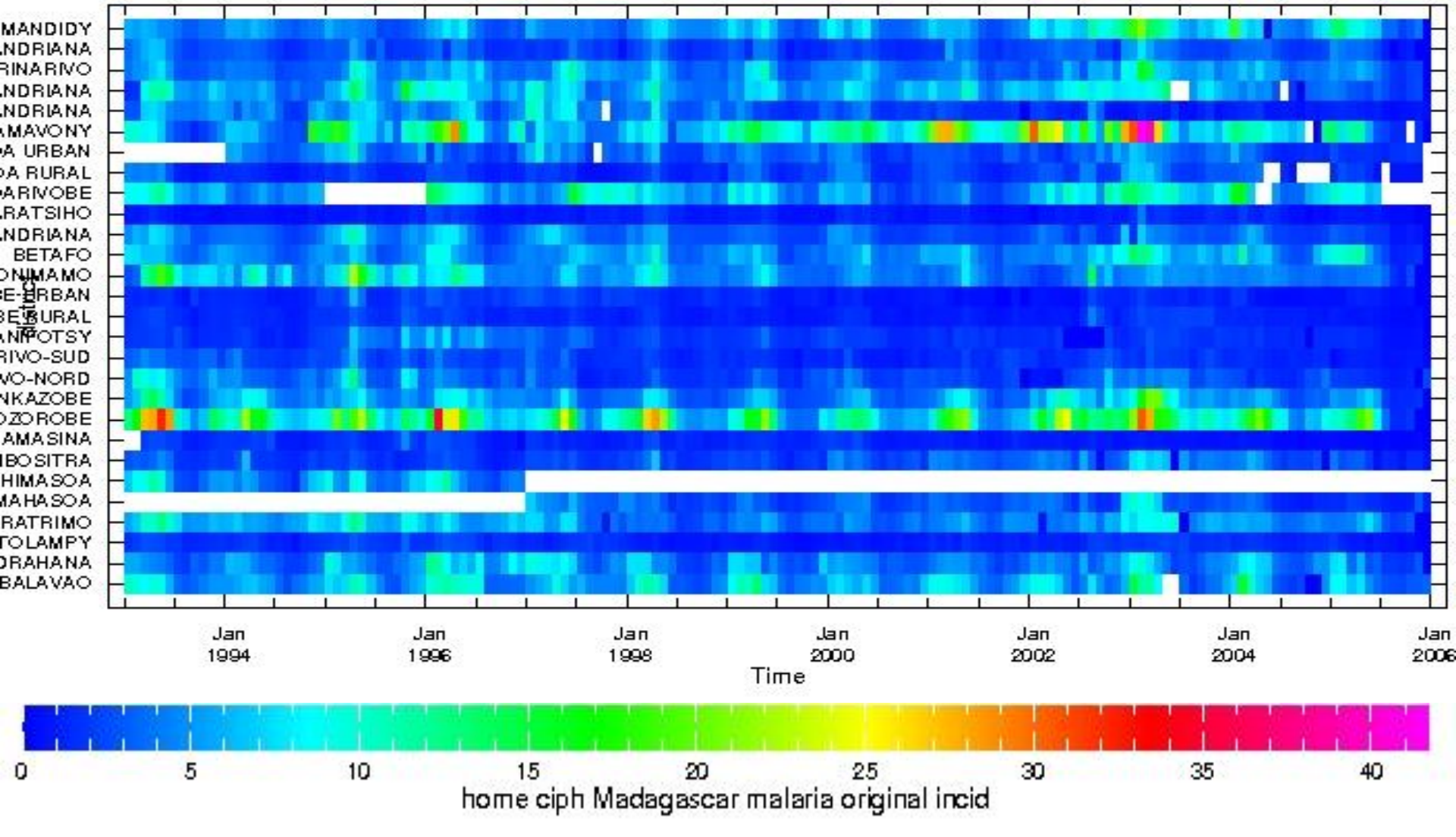
Examinación Inicial de los Datos

Por ejemplo, considere la incidencia de malaria en los Altiplanos de Madagascar para 1993-2005. Esta base de datos está accesible en la Data Library aquí:

expert

home .ciph .Madagascar .malaria .original

Examinación Grafica Inicial de los Datos



Examinación Grafica Inicial de los Datos

El grafico muestra incidencia como color según el tiempo (eje X) y los distritos (eje Y). **Blanco** indica datos que faltan, y parece evidente que Ambohimahasoa y Ambohimasoa son un mismo distrito disjunto.

Sugiere que ambos nombres representan el mismo lugar y que, a un momento, un nombre diferente fue utilizado. Al consultar el productor de estos datos, se verifica que es lo que pasó.

Examinación de la Revisión de los Datos

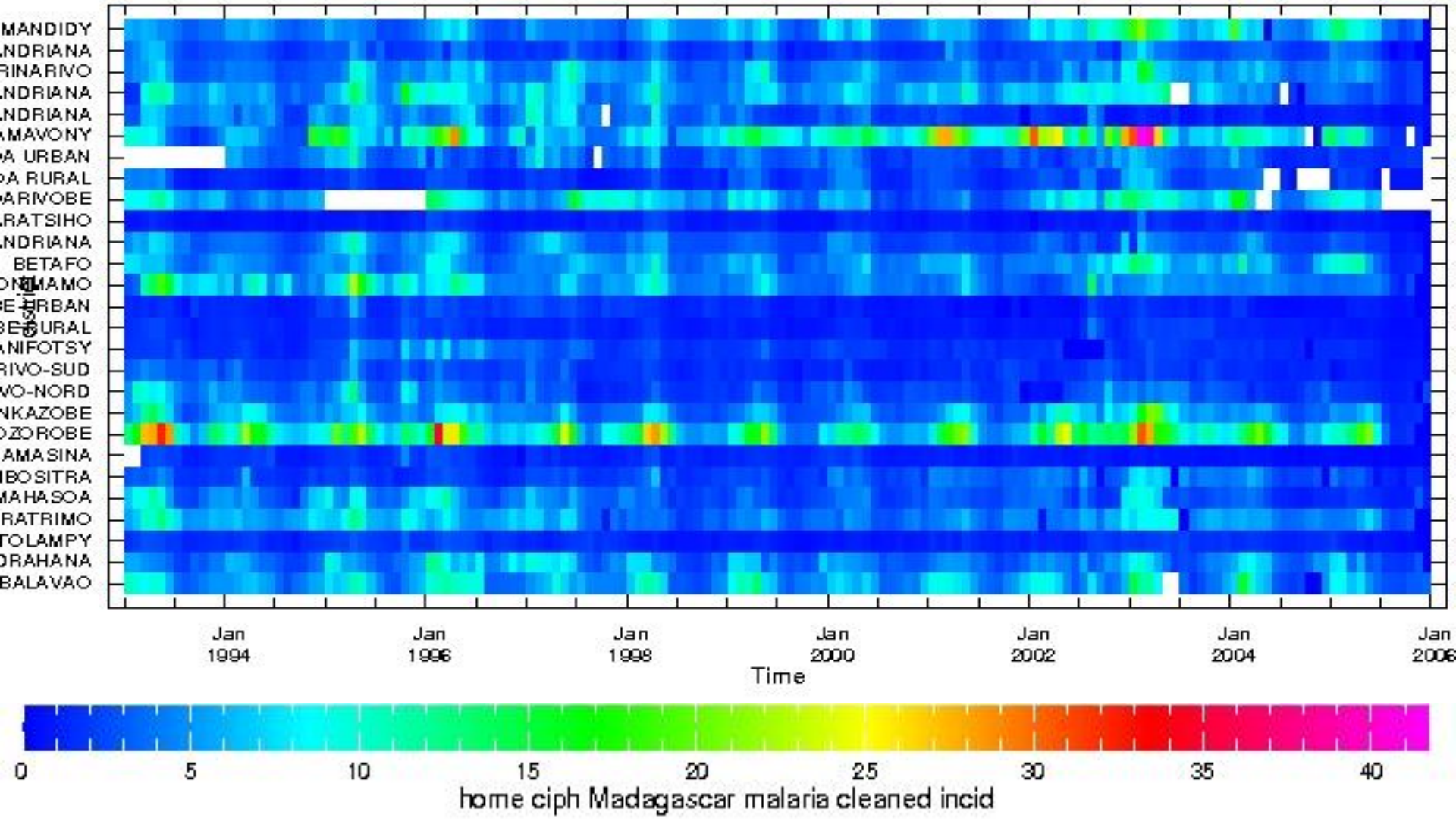
La revisión de los datos está cargada con un nuevo nombre:

expert

home .ciph .Madagascar .malaria .cleaned

Los datos que faltaban que eran sospechosos desaparecieron.

Examinación Grafica de la Revisión de los Datos



Geo-locación

Ahora que esta coherente, se puede analizar la base de datos. Sin embargo, se quiere visualizar los datos de incidencia en su locación geográfica. Para geo-localizar los datos, explore las bases de datos de la Data Library en **SOURCES .Features**.

SOURCES .Features .Political .Madagascar .Districts

use_as_grid

Una de las variables de esta base de datos es ***District geometry*** (the_geom), que son las geometrías de cada distrito.

En este base de datos, las variables dependen de 'codefiv', un código numeral para cada distrito. Una variable *FIV name* (nomfiv) corresponde a los nombres de los distritos de los datos cargados.

Se puede cambiar la variable the_geom tal que depende de 'nomfiv' y dar esta nueva retícula el mismo nombre que ella tiene en la base de datos cargados, es decir 'district', con la función **use_as_grid**:

SOURCES .Features .Political .Madagascar .Districts
nomfiv /district use_as_grid

add_variable

Por fin, hay que añadir la variable the_geom a la base de datos cargados utilizando la función add_variable para referenciarla. En **Expert Mode** se traduce:

```
home .ciph .Madagascar .malaria .cleaned
SOURCES .Features .Political .Madagascar .Districts
nomfiv /district use_as_grid .the_geom add_variable
```

Verificación de la Coherencia Espacial

Asegúrese que ninguno distrito falta. Para comparar los distritos de los datos de incidencia con los de la variable the_geom, utilice la función SAMPLE_MISSING que resulta en la variable incid restringida a los distritos que faltan en la variable the_geom.

incid the_geom[district]SAMPLE_MISSING

Resultados

Idealmente, ninguno distrito de la variable incid falta en la variable the_geom. Pero en este caso, si… Quedan siete distritos que existen para la variable incid pero no para la variable the_geom:

(ANTANANARIVO-NORD) (ANTANANARIVO-SUD) (ANTSIRABE RURAL) (ANTSIRABE URBAN) (FENOARIVOBE) (FIANARANTSOA RURAL) (FIANARANTSOA URBAN)

Consulte con el Productor de los Datos

Una nueva consultación con el productor de los datos nos indica la correspondencia entre los nombres que se quedan sin geometría para la variable incid y nombres de distritos en la base de datos de geometrías.

Malaria District	Feature District
Antananarivo-Nord	ANTANANARIVO-AVARADRANO
Antananarivo-Sud	ANTANANARIVO-ATSIMONDRANO
Antsirabe Rural	ANTSIRABE II
Antsirabe Urban	ANTSIRABE I
Fenoarivobe	FENOARIVO-AFOVOANY
Fianarantsoa Rural	FIANARANTSOA II
Fianarantsoa Urban	FIANARANTSOA I

Version Corregida

Una versión corregida de la base de datos está llamada geolocated y está verificada de nuevo con la función SAMPLE_MISSING

home .ciph .Madagascar .malaria .geolocated

incid the_geom[district]SAMPLE_MISSING

Ahora, la incidencia y la cuenta de casos dependen del tiempo y de los distritos, y las geometrías de los distritos fueron incluidas en la base de datos.

Resumen

Describir precisamente el tiempo y la referencia espacial de los datos:

- Simplifica los análisis que siguen
- Permite aplicar funciones más sofisticadas
- Permite comparar con otras bases de datos instantáneamente