VÍNCULOS
JULIO - DICIEMBRE DE 2011
VOLUMEN 8 ■ NÚMERO 2

Análisis de distribución de datos de biodiversidad usando técnicas de minería de datos

Distribution analysis of biodiversity data using data mining techniques

Héctor Arturo Flórez Fernández*
Diego Fernando Roa**
Nathalia Garcés***

Fecha de recepción: Agosto 10 de 2011
Fecha de aceptación: Octubre 25 de 2011

Resumen

La administración de información entre diferentes actores en una comunidad virtual es un gran reto para los investigadores hoy en día. En comunidades de biólogos es necesario compartir la información de diferentes laboratorios con el fin de realizar análisis de distribución, monitoreo o predicción de comportamientos de varias especies. Este artículo presenta una arquitectura enfocada en realizar análisis de distribuciones de especies teniendo en cuenta un ambiente heterogéneo y distribuido de la información.

Palabras clave

Comunidades virtuales, análisis de información, biodiversidad, sistema distribuido, peer to peer, minería de datos, clustering.

* Ingeniero electrónico e ingeniero de sistemas de la Universidad El Bosque, magister en Ciencias de la Información y las Comunicaciones de la Universidad Distrital Francisco José de Caldas, especialista en Alta gerencia y Magíster en Gestión de Organizaciones de la Universidad Militar Nueva Granada y estudiante de doctorado en Ingeniería en la Universidad de los Andes. Docente Universidad Distrital Francisco José de Caldas. Correo electrónico: ha.florez39@uniandes.edu.co.

** Ingeniero de sistemas y computación y estudiante de maestría en Ingeniería de sistemas y computación de la Universidad de los Andes Correo electrónico: df.roa34@uniandes.edu.co

*** Ingeniero de sistemas y computación y estudiante de maestría en Ingeniería de sistemas y computación de la Universidad de los Andes Correo electrónico: n.garces26@uniandes.edu.co

Abstract

The information management between different actors in a virtual community is a great challenge for researchers today. In communities of biologists, it is necessary to share information from different laboratories to undertake analysis of distribution, monitoring or predicting behavior of various species. This paper presents an architecture focused on the analysis of species distributions having in consideration a heterogeneous and distributed environment of the information.

Keywords

Virtual Community, information analysis, biodiversity, distributed system, peer to peer, data mining, clustering.

Introducción

El desarrollo y evolución de la web 2.0 ha permitido el crecimiento exponencial de blogs, portales, redes sociales (entre otros). Estos medios permiten el intercambio de información estructurada y no estructurada entre grupos de personas con un objetivo en común.

Los biólogos por ejemplo, usan estas herramientas para monitorear especies, predecir comportamientos de los ecosistemas, y entender las relaciones entre diferentes tipos de datos, cómo los morfológicos, climáticos, oceanográficos y de población, que son actualizados constantemente por personas de la misma comunidad virtual o por entes gubernamentales que controlan este tipo de información.

Estos tipos de análisis pueden ser basados en información pública, es decir, cualquier persona suscrita a la comunidad virtual puede actualizar y modificar la información presente en un portal, o por el contrario, pueden existir entidades definidas que validan la información y de las cuales se tiene un alto grado de confianza sobre los datos expuestos, de forma que los resultados pueden ser validados y corroborados por expertos en el área.

Aunque esta información provenga de sitios "confiables", es importante resaltar la calidad de los datos con los que se trabaja en biología. Secuencias de ADN, monitoreo de epidemias o datos geo referenciados de distribución de especies, son ejemplos de tipos de información que es insertada manualmente en diferentes bases de datos. Por esta razón, se encuentran problemas de consistencia de los datos, falta de estándares de nombramiento y el desarrollo de una buena arquitectura de información.

Adicionalmente, la información presente en estas comunidades virtuales muy pocas veces se encuentra centralizada, debido a los grandes volúmenes de información que se presentan. En consecuencia, diferentes tecnologías como Grid, Peer to peer (P2P) o Clúster son usadas para almacenar los datos.

Este artículo presenta una arquitectura para realizar análisis de datos de biodiversidad, teniendo en cuenta un ambiente distribuido y heterogéneo de la información. Los datos de análisis corresponden a diez bases de datos de diferentes tipos de organismos (peces, aves, reptiles, anfibios, plantas etc.) del museo de la Universidad de los Andes.

Para describir la distribución de los datos de biodiversidad de Colombia, se usaron téc-

nicas de clustering sobre una arquitectura peer to peer.

Manejo de datos en biodiversidad

A través de la correcta administración y análisis de datos de biodiversidad, diferentes estudios han permitido identificar posibles riesgos y comportamientos biológicos en diferentes contextos. En [7] se presenta un sistema que realiza análisis relacionados con el monitoreo y predicción de comportamientos de las especies del parque nacional de las montañas de Rodna. A su vez, presenta estrategias para la predicción de la evolución de ciertas especies usando algoritmos específicos del dominio, junto con varios métodos matemáticos presentes en el estudio. Este artículo analiza diferentes fuentes de datos como clima, geografía, especies, entre otras.

Estas relaciones entre diferentes tipos de datos, puede ser útil para identificar relaciones entre especies, que favorecen o entorpecen el desarrollo de algún proceso industrial o natural. Por ejemplo, en [8] se logró demostrar que en Australia la presencia de insectos en los envíos de flores cortadas, generando un problema importante que afecta el comercio de exportación de flores.

Este es particularmente el caso de los países, como Japón y los Estados, que tienen normas muy estrictas de cuarentena. Un tratamiento de poscosecha para el control de insectos. Es una parte necesaria del proceso de manejo de flores de corte para la exportación.

En la actualidad, la mayoría de los exportadores no están satisfechos con los tratamientos de desinfección que ahora están en uso. Estos tratamientos convencionales no siempre matan a los insectos sin dañar las flores. En consecuencia, los métodos mejorados e innovadores están siendo investigados.

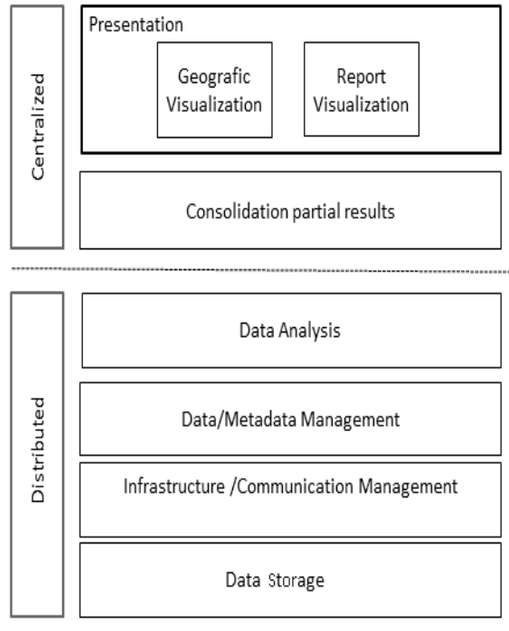
Los trabajos mostrados anteriormente son útiles para desarrollar tareas de análisis en un contexto específico, pero no tienen en cuenta variables asociadas a la recolección y administración de datos de biodiversidad en Colombia. En las siguientes secciones se presenta una arquitectura que tiene en cuenta la forma de recolección de información y necesidades de análisis de una comunidad en particular.

Definición del problema

El museo de la Universidad de los Andes quiere construir una aplicación que permita realizar análisis de distribución de las especies registradas, teniendo en cuenta la información de las 10 colecciones disponibles de los diferentes laboratorios (botánica, ecología molecular de vertebrados acuáticos, biología evolutiva de vertebrados, zoología, ecología acuática, entre otros).

La aplicación debe permitir realizar agrupaciones de poblaciones en determinadas regiones geográficas, de forma que se puedan visualizar regiones, estadísticas de la distribución y comportamiento de la toma de datos del museo.

Figura 1. Arquitectura de solución propuesta



La disponibilidad y privacidad de los datos depende del funcionamiento y políticas de los diferentes laboratorios de la Universidad. La anterior especificación del problema motiva a proponer la arquitectura que será presentada a continuación.

Arquitectura propuesta

Dadas las restricciones de desarrollo de la aplicación, se propone la siguiente arquitectura.

Figura 1. Arquitectura de solución propuesta

Los componentes de la arquitectura son los siguientes:

1. **Componente de almacenamiento:** Almacena la información de biodiversidad de una base de datos SQL Server por cada nodo en ejecución.
2. **Componentes de Infraestructura:** Administra la localización y comunicación entre los nodos del sistema (FreePastry).
3. **Componente de Administración de tareas:** Administra los diferentes tipos de tareas que se tienen entre los nodos (Scribe).
4. **Componente de Análisis:** Transforma los datos y ejecuta el modelo de minería (Weka). En la siguiente sección se detalla las decisiones tomadas con respecto a la técnica y algoritmo de minería a utilizar.
5. **Componente de visualización**: Muestra los resultados del modelo de minería en forma de gráficas y mapas al usuario final.

Para realizar análisis de la información distribuida es necesario definir qué técnicas de minería de datos son útiles para poder resolver la principal pregunta de minería que tienen los expertos.

Por esta razón, usar técnicas de clustering para logar caracterizar y describir los datos. A su vez, la elección de qué algoritmo usar para realizar agrupaciones depende del objetivo propuesto.

La técnica más común para realizar clustering es el algoritmo de K-medias. Esta técnica permite elegir el número centroides iníciales para desarrollar el proceso de segmentación. Además, cada registro pertenece a un solo cluster de información. Por otro lado, el algoritmo de Expectation Maximization (EM) desarrollado por Microsoft permite que los datos pertenezcan a más de un clúster solapando los resultados.

Esta ventaja que ofrece EM, permite que se observen de una mejor forma los clusters de distribución de especies sobre un mapa o una región geográfica. Por esta razón, los análisis presentados en la siguiente sección, realizan una comparación de ambos algoritmos teniendo como uso base el algoritmo de Expectation Maximization.

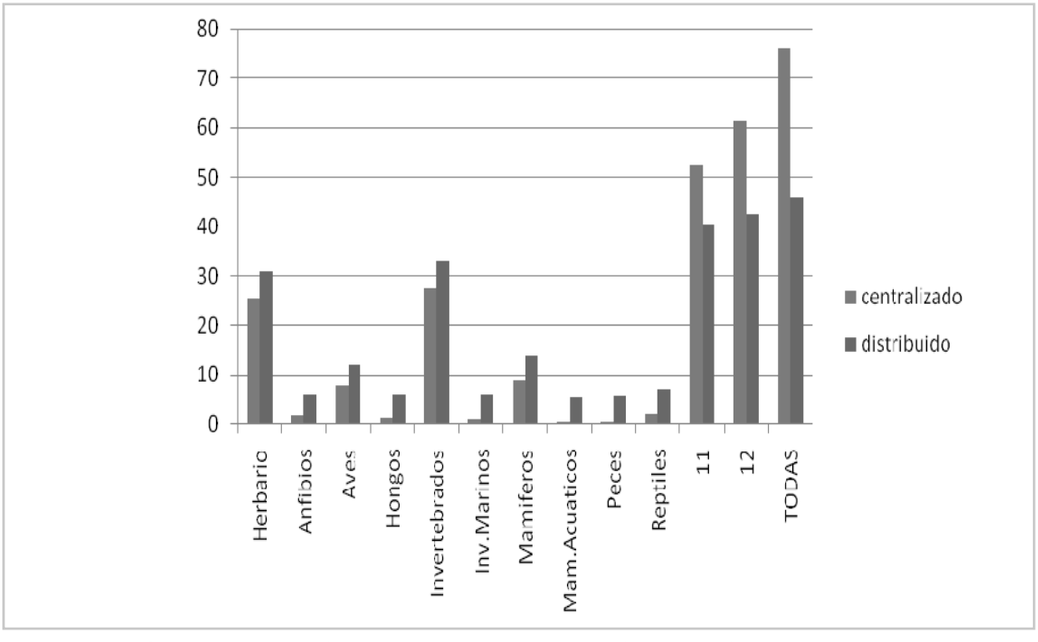
Evaluación

Para la evaluación de la aplicación desarrollada se tuvo en cuenta tanto la arquitectura como el algoritmo de minería utilizado. Para la arquitectura se tiene como objetivo probar el desempeño de la aplicación y la veracidad de los datos. Para el algoritmo se quiere evaluar la calidad de los datos.

Arquitectura

Para evaluar el desempeño de la arquitectura P2P se implementó la versión centralizada del proyecto. Acá hay un único nodo que es el encargado de los trabajos de minería y de visualización, el cual se contrapone con los diez nodos que se utilizan para el sistema distribuido. Cabe mencionar que el nodo de la arquitectura centralizada tiene las mismas especificaciones de hardware que los nodos usados para la prueba descentralizada, con el fin comparar los datos de manera equitativa.

Figura 1. Comparacion en tiempo de la arquitectura distribuida Vs centralizada



La siguiente figura muestra el tiempo en que se demoran ambas implementaciones, es decir, centralizada y distribuida mediante arquitectura P2P ejecutando las mismas tareas.

Como se puede ver en la figura 2, las primeras diez tareas son ejecuciones de una sola colección. Esto muestra que las colecciones más demoradas de procesar son Herbario e Invertebrados. De igual forma deja en evidencia que una arquitectura centralizada toma menos tiempo que la distribuida cuando hay una sola tarea o cuando las tareas son muy pequeñas, porque la arquitectura distribuida gasta tiempo en el intercambio de mensajes. Las tareas 11 (Peces, Herbario, Invertebrados) y 12 (Mamíferos, Herbario, Invertebrados) muestran que para tareas más grandes el sistema distribuido es mucho más eficiente, dado que las tareas se distribuyen y hacen uso de procesos de paralelización. La última tarea, equivalente a todas las colecciones, muestra que el sistema distribuido realiza la tarea casi al 50% del tiempo tomado por la arquitectura centralizada.

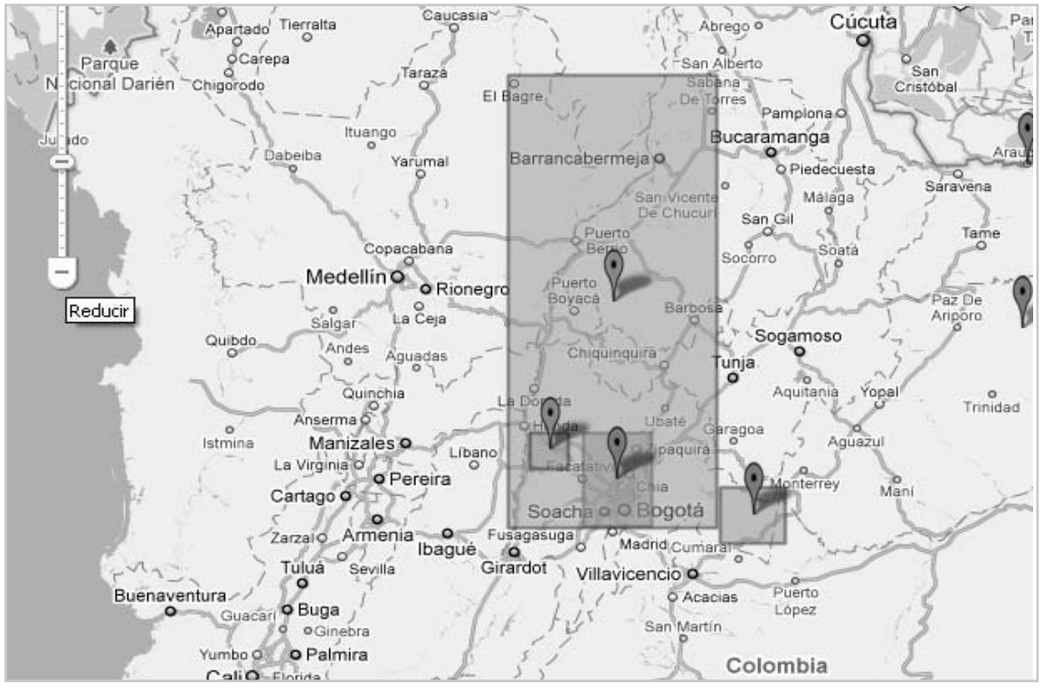
La explicación de este comportamiento es que mientras el centralizado ejecuta secuencialmente todas las colecciones y el tiempo total es la suma que se demora procesando cada una de ellas, el sistema distribuido tiene diez nodos a su disposición que trabajan de manera paralela, por lo que el tiempo no va a ser la suma de los tiempos de cada nodo sino el tiempo del nodo más lento.

Por otro lado, se ha contrastado los datos obtenidos en la arquitectura centralizada con la arquitectura distribuida para evaluar la veracidad de los datos. Al hacer esta prueba se confirma la hipótesis de que el algoritmo distribuido se comporta de manera exacta que el centralizado, es decir, que ambas soluciones producen el mismo resultado.

Algoritmo

Para comparar la calidad de los datos se ha implementado el algoritmo K-means. Sin embargo en la solución se ha escogido el algoritmo Expectation Maximization para la técnica de minería de datos porque este

Figura 2. Visualización de los resultados de minería para la tarea Invertebrados-Reptiles



algoritmo permite solapar los datos dentro de diferentes clústeres.

A manera de ejemplo se muestra la figura 3, la cual es el resultado de minería para la tarea invertebrados (rosado) y reptiles (azul). Acá se puede apreciar que los clústeres de reptiles están solapados. Este mecanismo permite al usuario obtener más información sobre cómo están distribuidas las especies.

Además de la diferencia anterior, el número de clústeres que genera K-means para las colecciones son muy diferentes a las de

Figura 3. Resultados de Aves por k-means.

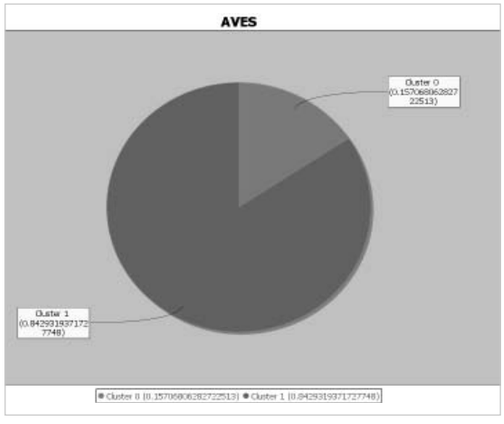


Figura 4. Resultados de Aves por EM.

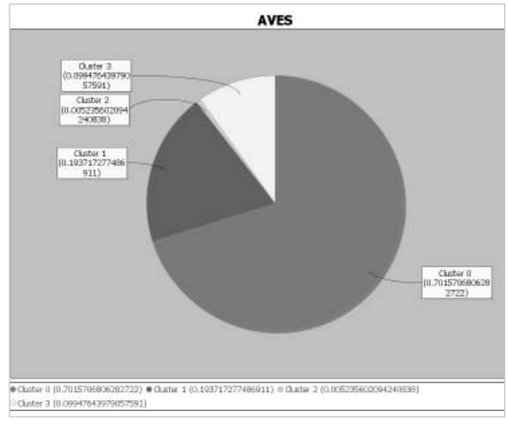


Figura 5. Resultados del sistema con todas las especies.

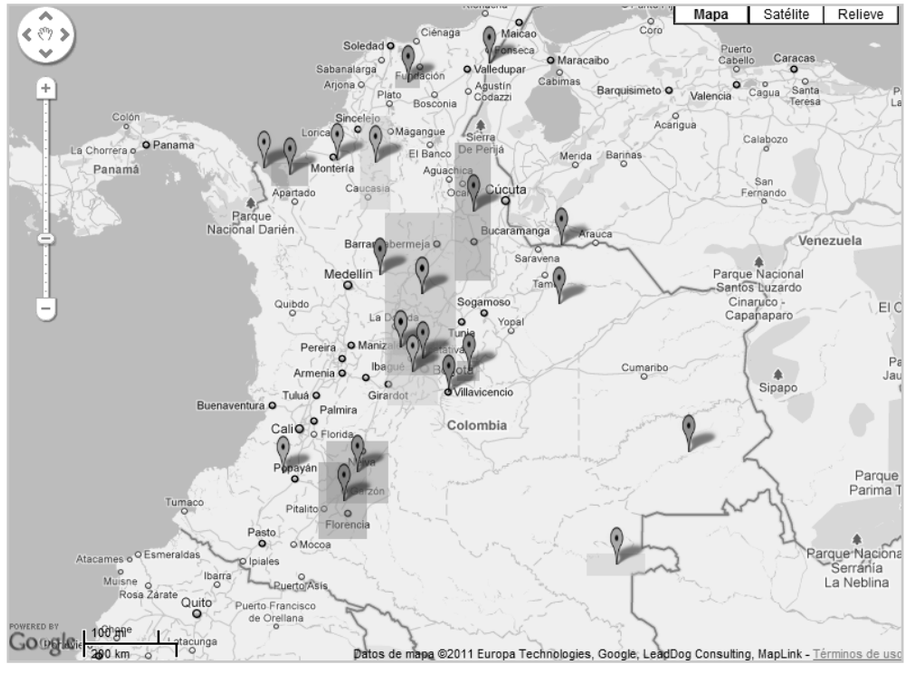
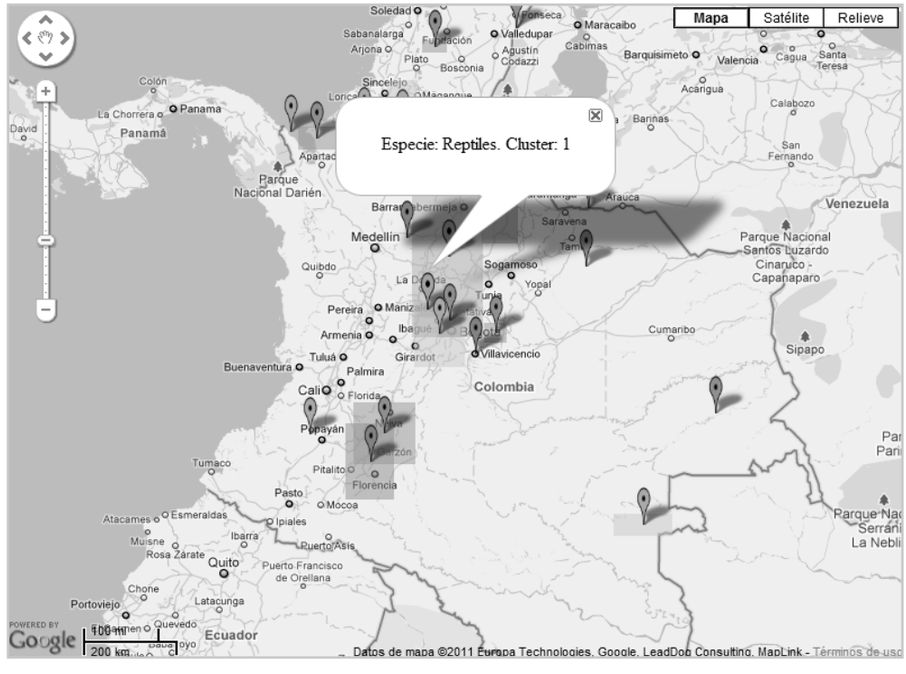


Figura 6. Resultados del sistema con todas las especies.



EM. K-means agrupa a cada una de las 10 colecciones en 2 clústeres mientras EM varía entre 2 y 9. De esta forma los resultados de EM permiten un mejor entendimiento de la distribución de la información.

A manera de ejemplo se muestra el resultado de analizar la colección Aves para ambas técnicas.

La siguiente figura presenta los resultados de todo el proceso ejecutado en el sistema distribuido tomando todas las especies descritas anteriormente, mediante la técnica de Expectation Maximization.

En la figura anterior se puede apreciar que cada especie tiene varios cluster en donde el marcador hace referencia al centroide y este a su vez provee información del cluster como se muestra en la siguiente figura.

Conclusiones y trabajo futuro

Este artículo presenta una arquitectura distribuida para realizar análisis de información sobre datos de biodiversidad. Los resultados obtenidos y los algoritmos de minería de datos utilizados demuestran que es posible realizar análisis sobre información distribuida con buenos tiempos de respuesta, presentando resultados útiles para el usuario.

La decisión de realizar técnicas de minería de datos distribuidas sobre el objetivo propuesto requiere de un acompañamiento del experto, que permita establecer las políticas de intercambio y confidencialidad de los datos.

Como trabajo futuro se piensa realizar análisis de distribución mundiales, integrando diferentes fuentes de datos de clima, geografía, etc., teniendo en cuenta variables que se manejan en el museo de la Universidad de los Andes.

Agradecimientos

Los autores quieren agradecer al museo de la Universidad de los Andes por facilitar la recopilación de la información base para el desarrollo del proyecto.

Referencias

Cluj-Napoca. Biodiversity management system in Rodna Mountains National Park. 2010 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR). ISBN: 978-1-4244-6724-2.

Cluster Analysis. (s.f.). Recuperado el 5 de Mayo de 2011, de StatSoft: http://www.statsoft.com/textbook/cluster-analysis/#k

GBIF, Global Biodiversity information Facility. Recuperado 17 de Marzo de 2011 de http://www.gbif.org/

WorldClim, Global Climate Data. Recuperado 17 de Marzo de 2011 de http://worldclim.org/