Factores socioeconómicos que influyen en el nivel de protección a nivel provincial de los mamíferos de Argentina

PRA1 | Tipología y ciclo de vida de los datos

Antonio Gutiérrez Blanco
Nicolás Caruso

Universitat Oberta de Catalunya

ÍNDICE

1. CONTEXTO

En el año 2018 la Sociedad Argentina de Mastozoología, junto con la Secretaría Nacional de Ambiente y Desarrollo Sustentable pusieron en marcha el proceso de categorización de los mamíferos de Argentina que culminó con la publicación, en el año 2019, de un portal web que contenía la información recopilada durante el proceso. Si bien la idea original del portal era contener toda la información abierta para quien quisiera consultarla, los mapas de distribución de las especies no están disponibles para su descarga.

Este trabajo tiene como objetivo generar un pipeline que permita la descarga de la información geoespacial de la distribución de las especies de mamíferos presentes en Argentina en un formato acorde que permita su procesamiento posterior. Pero además, nos propusimos recolectar información georeferenciada de las áreas protegidas presentes en Argentina que nos permita evaluar el nivel de protección efectiva de dichas áreas sobre las poblaciones de mamíferos. Para esto, recurrimos al portal Protected Planet que es la base de datos más actualizada respecto de las áreas protegidas a nivel global. Esta información, si bien está disponible para su descarga, se actualiza constantemente por lo que decidimos también aplicar técnicas de web scraping que permita tener la información actualizada. Por último, nos pareció interesante complementar el dataset con datos a nivel provincial que sean indicadores socioeconómicos y que puedan utilizarse en modelos subsecuentes tendientes a evaluar el nivel de protección de las especies en cada provincia y que resulten en herramientas útiles para la gestión a nivel nacional.

2. TÍTULO

Factores socioeconómicos que influyen en el nivel de protección a nivel provincial de los mamíferos de Argentina.

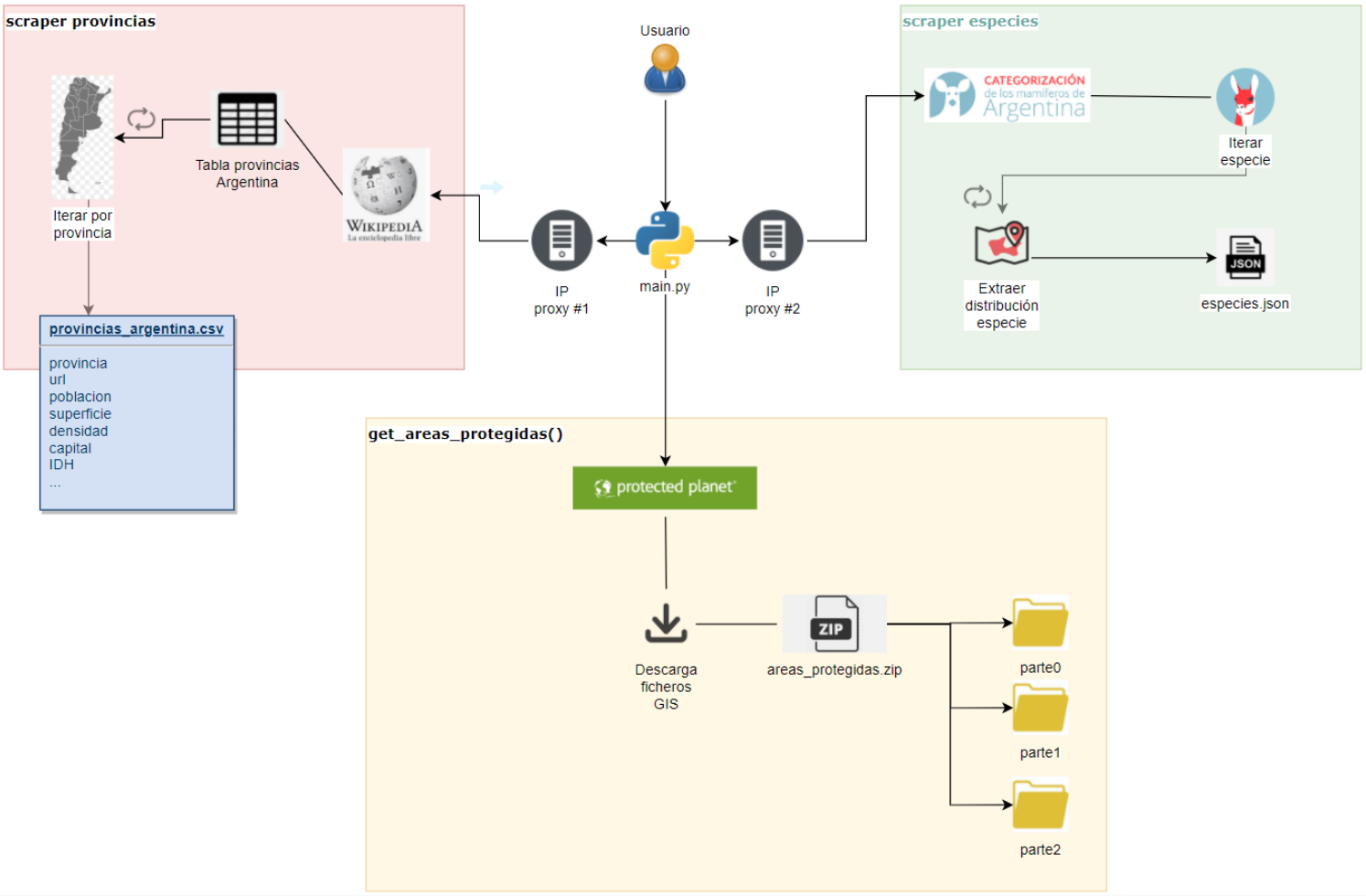
3. DESCRIPCIÓN DEL DATASET

Este trabajo consistió en la recopilación de datos georeferenciados acerca de la distribución de las especies de mamíferos presentes en Argentina, combinando esta información a nivel provincial con una serie de indicadores socio-económicos y de protección ambiental. Como resultado del proyecto, se han generado tres conjuntos de datos diferenciados:

1. ***provincias_argentina.csv***: Información demográfica y socioeconómica de las provincias de Argentina.
2. ***especies.json:*** Información georeferenciada de la distribución de los mamíferos de Argentina, junto con el nombre científico y su categoría de conservación.
3. ***areas_protegidas.zip:*** Información georeferenciada de la extensión de las áreas protegidas de Argentina.

4. REPRESENTACIÓN GRÁFICA

A continuación se muestra un representación alto nivel del diagrama de flujo del proyecto, donde se ilustran las tres scrapers (arañas) que capturan la información necesaria para construir los diferentes conjuntos de datos:

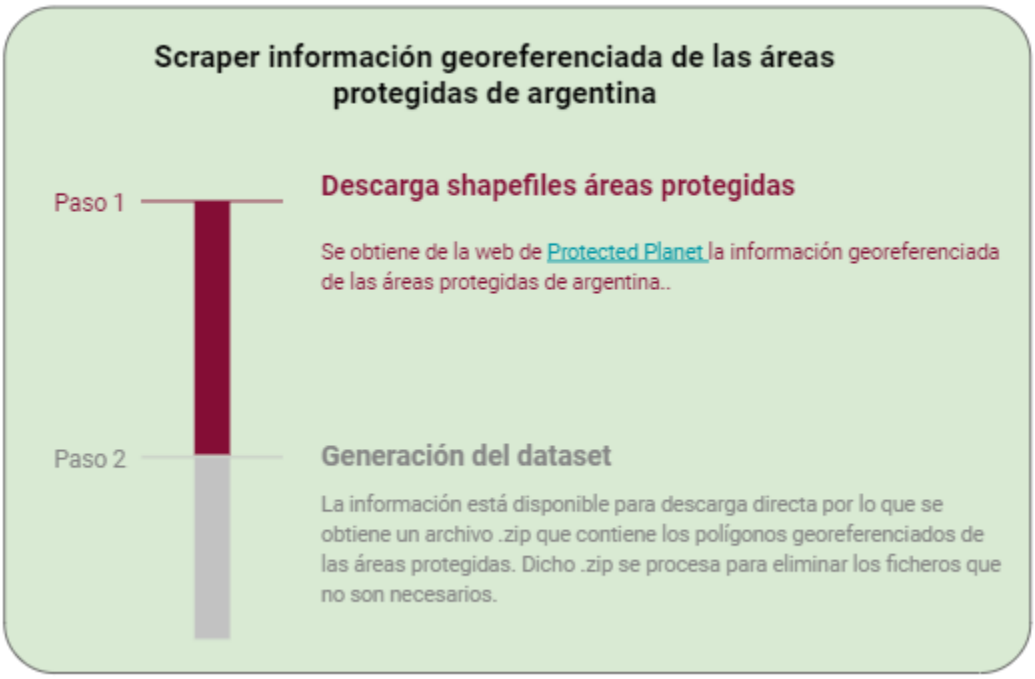


Adicionalmente, los siguientes diagramas muestran en detalle los pasos seguidos para cada uno de los tres procesos de *web scraping* recogidos en la librería Python *scraper_functions* que se puede ver en el repositorio Git:

- Generación del conjunto de datos con la información demográfica y socioeconómica de las provincias de Argentina (*provincias_argentina.csv)*:

- Generación del conjunto de datos con la información georeferenciada de las especies de mamíferos de Argentina (*especies.json)*:

- Obtención de los datos con la información georeferenciada de las áreas protegidas de Argentina (*areas_protegidas.zip*):



5. CONTENIDO

El contenido de cada uno de los juegos de datos generados se detalla a continuación:

provincias_argentinas.csv

El conjunto de datos con la información demográfica y socioeconómica de las provincias de Argentina contiene la siguiente información:

- ***provincia:*** nombre de la provincia de Argentina
- ***url:*** enlace a wikipedia con la información de detalle para la provincia
- ***población:*** número de habitantes
- ***superficie:*** extensión territorial de la provincia
- ***densidad:*** densidad de población
- ***capital:*** capital de la provincia
- ***IDH:*** Índice de Desarrollo Humano (ver detalle ***aquí***) (año 2018)
- ***analfabetismo:*** tasa de analfabetismo (año 2010)
- ***autonomía:*** fecha de declaración de la autonomía

- ***altitud media:*** altitud media sobre el nivel del mar
- ***altitud maxima:*** altitud máxima sobre el nivel del mar
- ***altitud minima:*** altitud mínima sobre el nivel del mar
- ***coordenadas:*** coordenadas geográficas (latitud y longitud)

Para la obtención de la información, se han seguido los pasos descritos en el punto anterior, obteniendo (mediante *web scraping* con la librería *Selenium*) la lista de provincias de la siguiente página de Wikipedia, la cual no está recogida en el fichero *robots.txt* como página cuya información no esté permitida obtener:

https://es.wikipedia.org/wiki/Provincias_de_Argentina

Posteriormente, se ha obtenido información detallada de cada provincia de Argentina disponible en las respectivas páginas de Wikipedia para cada provincia. Cada atributo data de una fecha determinada, siendo por ejemplo el IDH del 2018 y la tasa de analfabetismo del 2010. Existe otra información que, por su naturaleza, resulta invariante en el tiempo (p.e. la declaración de Autonomía, altitudes o capital).

especies.json

El conjunto de datos con la información georeferenciada de la distribución de las especies de mamíferos en Argentina está en formato json, el cual fue elegido ya que la información espacial que contiene puede ser muy extensa para ciertas especies y este formato se ajusta mejor a esos casos. El archivo json contiene la siguiente información para cada una de las especies:

- ***nombre de la especie:*** nombre científico de la especie
- ***categoría:*** iniciales de la categoría de conservación de la especie (por ej., “VU”: vulnerable)
- ***información espacial:*** string que tiene la información de latitud y longitud de cada punto que conforma el o los polígonos de distribución de la especie. Este dato requiere de una limpieza posterior ya que contiene otra información, además de la espacial, no relevante para este trabajo.

Este conjunto de datos se extrajo de la web de Categorización de los Mamíferos de Argentina. En primer lugar se obtuvo la lista de las URL de cada una de las especies de mamíferos. Luego, para cada una de dichas direcciones, se extrajo la información del nombre científico, la categoría y los datos georeferenciados de las áreas de distribución. Toda esta información se guardó en un archivo json.

areas_protegidas.zip

Este archivo contiene información georeferenciada de las áreas protegidas presentes en Argentina, descargando la información de la web https://www.protectedplanet.net

Se trata de un archivo .zip que en su interior contiene tres carpetas con diferentes archivos *.shp* que representan las áreas protegidas (polígonos y puntos). Estos archivos están acompañados por otros cuatro archivos auxiliares (*.cpg, .dbf, .prj y .shx*) necesarios para poder leer la información georeferenciada en cualquier software de procesamiento de datos espaciales. La partición en tres subcarpetas facilita la lectura de los ficheros con cualquier librería o software GIS ya que de esta manera ocupan menos tamaño, una vez cargados las diferentes áreas, se pueden combinar (merge) para crear una única área con toda la información.

6. AGRADECIMIENTOS

La situación de la biodiversidad a nivel mundial ha sido reconocida como crítica en numerosos foros y publicaciones. La pérdida del área de distribución histórica debida a actividades humanas representa un paso previo a la pérdida de especies, y este proceso se ha acrecentado durante las últimas décadas (Barnosky et al., 2011). Las áreas protegidas (APs) son instrumentos que ayudan a frenar este proceso, ya que son sitios que preservan los hábitats de las especies, sus servicios ecosistémicos y recursos culturales asociados. Actualmente el sistema de áreas protegidas cubre aproximadamente un 12% de la superficie de la Tierra y, si bien el número de APs muestra una tendencia creciente, aún es insuficiente para proteger ciertas especies, sobre todo aquellas en peligro de extinción (Brooks et al., 2004). Frente a esta situación, conocer el estado de conservación de las especies de fauna silvestre es fundamental para diagramar planes de gestión y manejo que aseguren la persistencia de sus poblaciones y la coexistencia armónica con los humanos.

Contar con mapas actualizados de distribución de las especies e información georreferenciada de las áreas protegidas permitiría hacer una análisis que muestre el nivel de protección que ofrece el sistema de áreas protegidas de una región a las especies que allí habitan. Caraballo et al. (2020) realizaron un estudio donde evalúan el nivel de conservación de las especies del género *Ctenomys* en Argentina, solapando las áreas de distribución con los polígonos de las áreas protegidas nacionales. Estos autores utilizaron las mismas fuentes

de información que las propuestas en este trabajo. Otro estudio similar, es el de Rodriguez et al. (2004) donde combinaron datos de distribución de especies y de áreas protegidas para realizar el primer análisis respecto de la efectividad del sistema de APs para representar la biodiversidad a nivel global. Este trabajo fue pionero y sentó las bases para la realización de este tipo de estudios a nivel regional/nacional, como el que estamos proponiendo aquí. Adicionalmente, éstos análisis pueden llevarse a cabo a nivel provincial analizando la influencia de variables socioeconómicas (cómo la población, el índice de desarrollo humano, etc.) sobre el nivel de protección de las especies en cada provincia. Baldi et al. (2019) desarrollaron un estudio similar, pero a nivel nacional para los países de América Latina, donde evaluaron la eficacia de las áreas protegidas en relación a aspectos demográficos, económicos y geopolíticos.

Bibliografía:

- Baldi G, Schauman S, Texeira M, Marinaro S, Martin OA, Gandini P, Jobbágy EG (2019). Nature representation in South American protected areas: country contrasts and conservation priorities. *PeerJ*, *7*, e7155.
- Barnosky AD, Matzke N, Tomiya S, Wogan GO, Swartz B, Quental TB, et al. (2011). Has the Earth's sixth mass extinction already arrived?. *Nature*, *471*(7336), 51-57.
- Brooks TM, Bakarr MI, Boucher T, Da Fonseca GA, Hilton-Taylor C, Hoekstra JM, et al. (2004). Coverage provided by the global protected-area system: is it enough?. *BioScience*, *54*(12), 1081-1091.
- Caraballo DA, López SL, Carmarán AA, Rossi MS (2020). Conservation status, protected area coverage of *Ctenomys* (Rodentia, Ctenomyidae) species and molecular identification of a population in a national park. Mammalian Biology, 100(1), 33-47.
- Rodrigues AS, Andelman SJ, Bakarr MI, Boitani L, Brooks, TM, et al. (2004). Effectiveness of the global protected area network in representing species diversity. Nature, 428(6983), 640-643.

7. INSPIRACIÓN

Como se mencionó en el apartado anterior, la información recopilada en este trabajo es fundamental para entender el nivel de protección que tienen las especies de mamíferos en Argentina. Esta información es básica para diagramar nuevas áreas protegidas o planes de manejo tendientes a mejorar la protección de las especies más vulnerables. Pese a que ya

existen estudios similares a nivel nacional, **con la información obtenida en esta práctica se puede analizar la información a nivel provincial y nacional de las especies y las respectivas provincias** donde habitan, lo cual no es posible con los actuales estudios realizados (punto 6), siendo por tanto una mejora respecto al actual estado del arte.

En particular, con la información proporcionada en los diferentes conjuntos de datos se puede dar respuesta a preguntas como las siguientes:

- ¿Cuál es el nivel de protección que el sistema de áreas protegidas aporta a las especies de mamíferos en Argentina?
- ¿Existen diferencias en el nivel de protección entre las especies más vulnerables vs. las menos vulnerables?
- ¿Están las especies mejor protegidas en aquellas provincias con un IDH más alto?
- ¿Existe alguna relación entre las especies en peligro de extinción y la tasa de analfabetimos de las provincias donde habitan?
- ¿Existe correlación entre las especies con menor extensión de hábitat (km^2) y las provincias más pobladas?

Además por ejemplo, mediante la historificación de los datos de la distribución de las especies, también se podría dar respuesta a preguntas de la siguiente índole:

- ¿Cuáles son las especies de mamíferos en Argentina para las que el área de distribución crece y en cuáles decrece con el paso de los años?
- ¿Cuál es la tasa de expansión/retracción interanual de las áreas de distribución de los mamíferos en Argentina?
- ¿Sobreviven mejor las especies con mayor representación en las áreas protegidas?

8. LICENCIA

Los conjuntos de datos generados se encuentran bajo licencia CC0: Public Domain License.

De esta forma, la información es completamente abierta al dominio público, permitiendo la copia, modificación y distribución de la obra sin pedir permiso a los autores. De esta manera se pretende fomentar el análisis de la información, a favor de la protección de las especies en Argentina. De igual forma, los autores no ofrecen ninguna garantía respecto a la obra y renuncian a cualquier responsabilidad por cualquier uso de la obra.

9. CÓDIGO

El código del proyecto se puede encontrar en el repositorio siguiente:

https://github.com/ntaxus/WEB-SCRAPING

Para la extracción de la información web se han utilizado técnicas de scraping de la información. Esto es debido a que esta información que se quiere capturar no es accesible mediante una API o descarga directa (excepto para el caso de las áreas naturales protegidas de Argentina), es por ello que se ha decidido utilizar la técnica del web scraping. Adicionalmente, la información alojada en estos sitios webs se actualizan periódicamente, por lo cual resulta útil contar con un código que automatice la descarga de la información. Esto posibilitará el desarrollo de visualizaciones en *real time* respecto de la evolución de las métricas que se obtengan a partir del análisis de los datos.

La librería elegida para realizar las tareas de scraping es *Selenium*, esto es así ya que en algunas páginas webs utilizan componentes en JavaScript (concretamente con la librería JQuery), siendo Selenium una buena herramienta para estos casos (aunque también se podría resolver con otras librería como *Scrapy* o *Beautiful Soup*). También, para la descarga del fichero de las áreas protegidas se puede emplear la librería Selenium para simular los eventos que lanzan la descarga, aunque se ha considerado que es más sencillo realizar directamente una petición http *GET* con la librería *requests* de Python para que se ejecute la descarga. También, se planteó Selenium inicialmente como una buena solución para descargar imágenes de cada una de las especies, ya que no era posible obtenerlas directamente accediendo a la URL (*Selenium* permite hacer screenshots), pero finalmente se descartó esta opción ya que las imágenes están protegidas por derechos de autor.

Adicionalmente, **se han usado técnicas para la prevención de obstáculos en web scraping** como son:

1. Se da la opción de especificar direcciones IP al ejecutar el programa que actúen como Proxy, de esta forma cada una de las arañas se aparecen detrás de una dirección IP distinta (previene bloqueo de IP).
2. Modificación del User-Agent de las peticiones, tanto en las realizadas con *Selenium* como cuando usamos la librería *request.*

3. Espaciado temporal de dichas peticiones para evitar saturar el servidor al que se realiza la petición.
4. Se han respetado los sitios marcados en el archivo *robots.txt* de cada página, accediendo exclusivamente a información marcada como permitida.

10. DATASET

Los tres datasets se encuentran publicados en el siguiente enlace:

https://zenodo.org/record/5644095#.YYL4apuvFH4

11. VIDEO EXPLICATIVO

El video explicativo del proyecto se encuentra alojado en la siguiente dirección:

https://drive.google.com/file/d/1fia7jf30IjG_f1ecylkLytKFe_8tpE7y/view?usp=sharing

12. CONTRIBUCIONES

Finalmente se lista la contribución de cada uno de los integrantes a la práctica, siendo:

- **N.C.**: Nicolás Caruso
- **A.G.B.**: Antonio Gutiérrez Blanco

Contribuciones	Firma
Investigación previa	N.C. y A.G.B.
Redacción de las respuestas	N.C. y A.G.B.
Desarrollo del código	N.C. y A.G.B.