Islas Canarias
Del 15 al 19 de noviembre de 2021

XXI Jornadas de Estadística de las Comunidades Autónomas

Web scraping para características de empresas

Ander Iparraguirre García
Matias Ledesma Sereno

https://www.linkedin.com/in/ander-iparraguirre/?originalSubdomain=es
https://www.linkedin.com/in/matiasledesmasereno/?originalSubdomain=es

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

Eustat

Erakunde autonomiaduna
Organismo Autónomo del
EUSKO JAURLARITZA
GOBIERNO VASCO

1. Introducción

- Eustat, el **Instituto Vasco de Estadística**, consciente de la potencia de las nuevas tecnologías para el uso de internet para recolectar información y alineado con los objetivos del ESSnet Big Data, siguiendo la estela de otros países.
- El objetivo del mismo es el uso del **webscraping**, **text mining** e **inferencia estadística** para recoger y procesar información de empresas y mejorar o actualizar la información registrada en el directorio de actividades económicas, generando **estadística experimental**.
- Con la colaboración de la **Universidad del País Vasco** (EHU-UPV)

¡Se ha lanzado un proyecto piloto para el scrapeo de webs de empresas!

1. Introducción

La técnica webscraping presenta una oportunidad única para mejorar la calidad de los registros sin incrementar **la carga de respuesta**.

Se han detectado varios **ítems** en la encuesta relacionados con la información objetivo del proyecto .

Tabla 1. Encuesta ESIE del Eustat sobre Medios de Comunicación Sociales

B.3.5. Uso de los medios de comunicación sociales (Indique cuáles utiliza su empresa)
a) Redes Sociales (Facebook, LinkedIn, Tuenti, Google+, …)
b) Blogs de empresas o microblogs (Twitter, Blogger, …)
c) Webs para compartir contenido multimedia (Youtube, Flickr, Instagram, …)

2. Objetivos

Tratar de obtener una nueva fuente de datos utilizando técnicas de **webscraping** para sustituir o validar a los obtenidos vía encuesta.

El objetivo ha sido **comparar los resultados** de la encuesta y lo que se obtiene mediante la técnica de webscraping**.**

3. Metodología

- En primer lugar, hemos usado la muestra de compañías extraída de la población de referencia incluída en la encuesta ESIE de Euskadi del año 2021 con solo las empresas con **sitios web conocidos**.
- A continuación se retiran las compañías **sin URL disponible**.

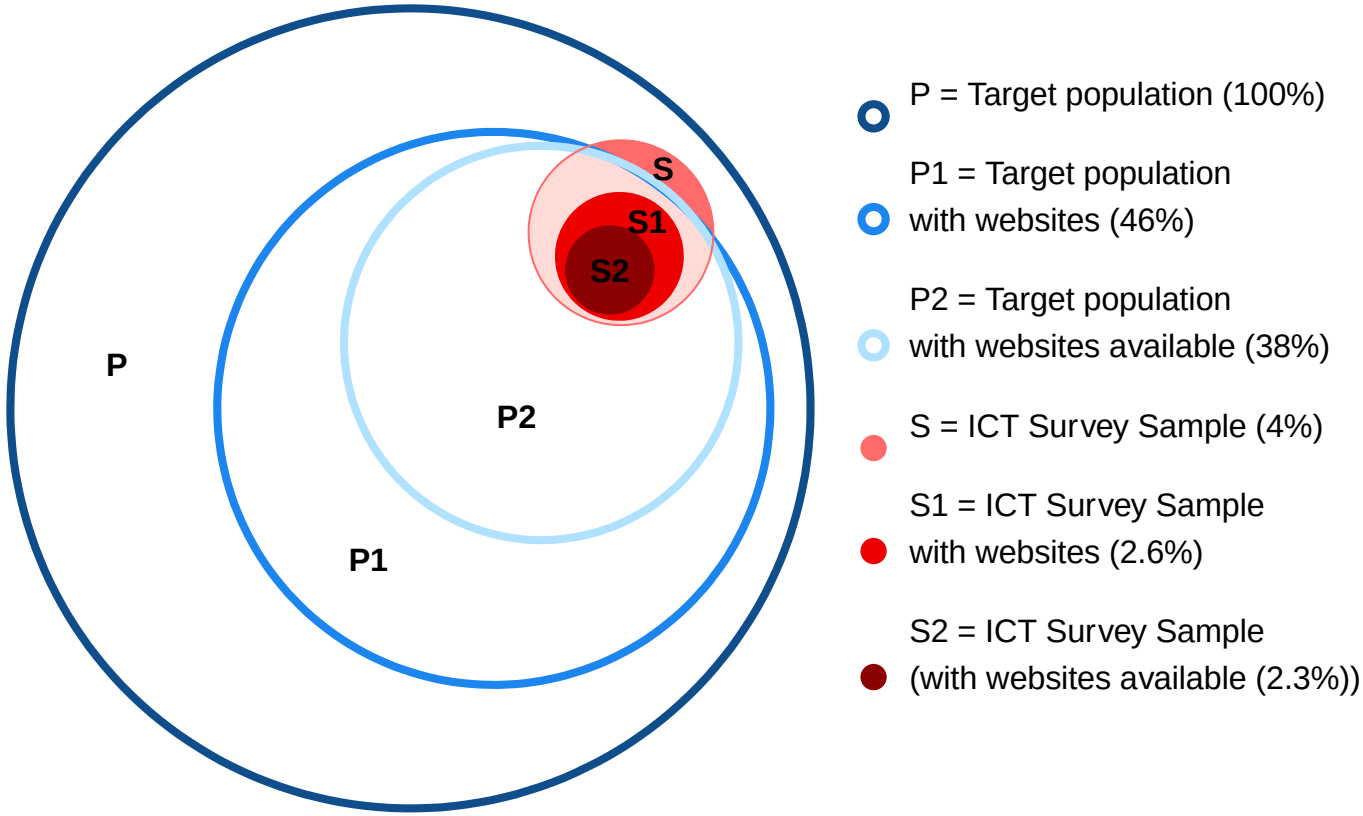
Se obtiene una submuestra de compañías con sitio web y URL accesible

4. Resultados

4.1. Muestra

Del total de empresas de Euskadi la muestra resulta en un subconjunto del 2,31% de la población.

Figure 1. Euler diagram of the process of selecting the training set.



4.2. Comparación de resultados

Scraping

Por el método de scraping se estima que el 69% de las compañías usan medios de comunicación sociales de acuerdo con la muestra **S2**.

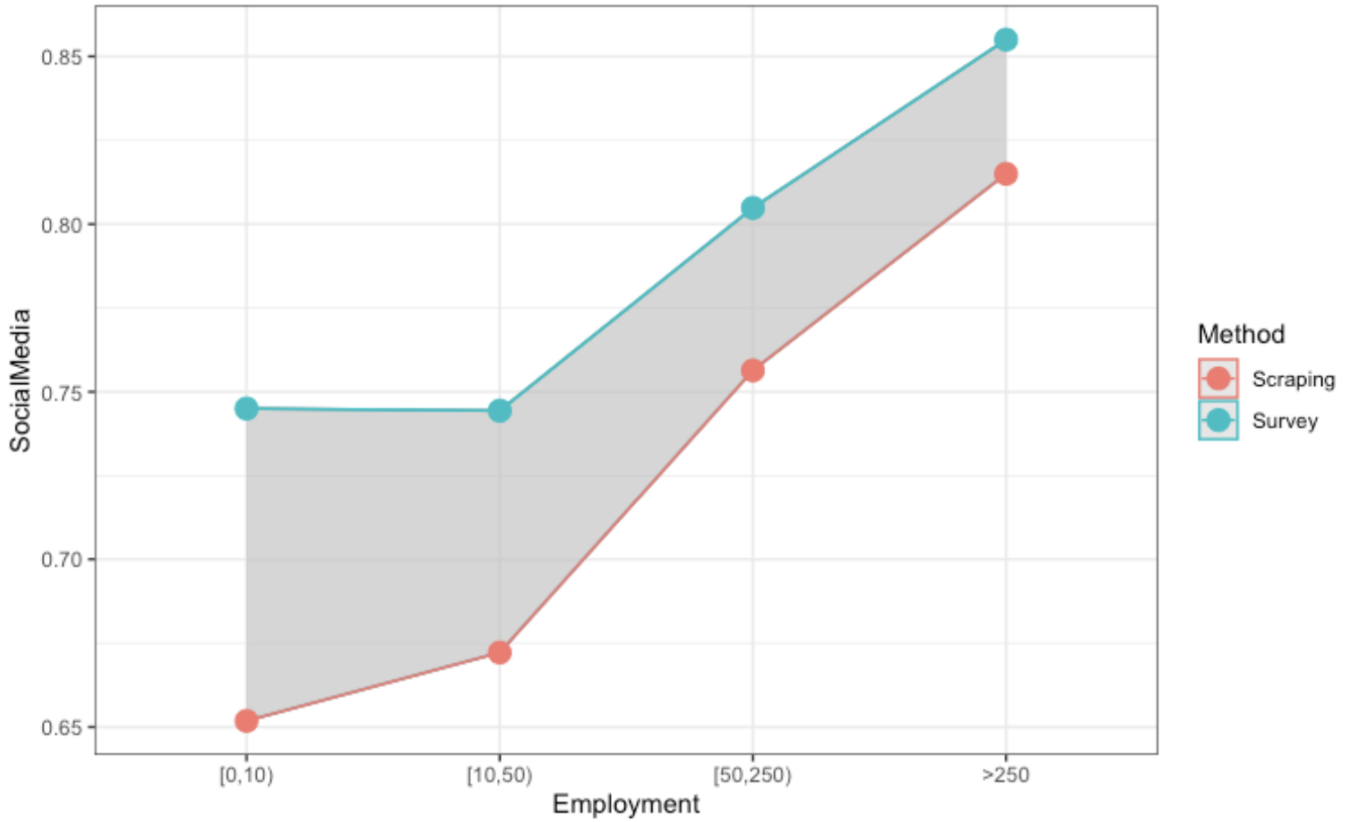
Vs

Encuesta

El 76,33% de las compañías respondieron afirmativamente a la pregunta “Usa su empresa medios de comunicación social?”.

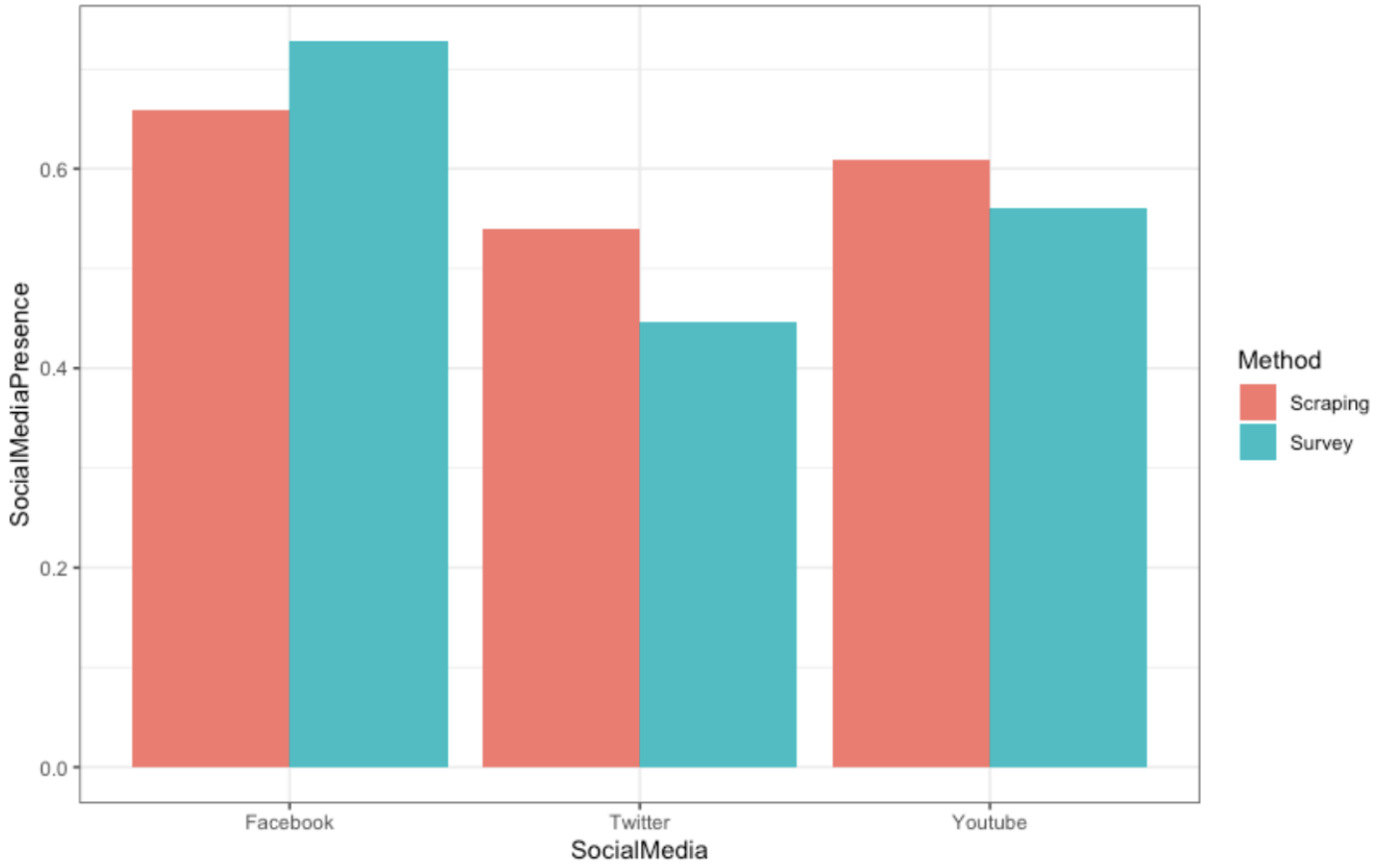
La discrepancia entre la estimación y los resultados de la encuesta puede deberse a la falta de medios de comunicación social de las pequeñas compañías.

Figure 2. Point chart showing how the survey and the forecast gradually converge as the size of the enterprise grows.



Analizando estadísticas más concretas presentadas anteriormente en la Tabla 1, observamos en la Figura 3 que el método de web scraping es razonablemente preciso.

Figure 3. Bar graph of the comparison of specific social.



5. Conclusiones

1. Los resultados arrojan una perspectiva razonable para la obtención por la vía alternativa presentada en este trabajo una vez se ha refinado el software original.

2. La discrepancia entre métodos se puede deber a diversos errores. En primer lugar, se ha detectado que la corrección manual de la muestra utilizada parece no ser suficientemente precisa y **difícilmente escalable** si se aumenta el volumen de la encuesta. Se esperan grandes resultados del proyecto piloto para solucionar dicho problema. Por otro lado, una vez se **discrimina por tamaño de empresas** los resultados son esperanzadores incluso si ponemos el foco en redes sociales concretas, lo que permitiría poder publicar datos más desagregados que hasta ahora no se habían publicado.

5. Conclusiones

3. **Rápido** de procesar, **sin desfase**, permite buscar otra información y modelización entre empleo y presencia en redes sociales.

4. Para **futuros trabajos**, el método y el tamaño de la muestra nos permite observar en mayor detalle las discrepancias de resultados entre métodos, mejorar el algoritmo para accede a más URL y encontrar la solución para elevar resultados a la población objetivo.

5. Finalmente añadir que el método de recogida permite publicar datos más **desagregados** en el campo de las medios de comunicación sociales.

Eskerrik asko!

¡Muchas gracias!