

Análisis del 10% de mayores ingresos en Chile

Marcos González, Agustín Rabie

18 January, 2025

Contents

Este proyecto analiza la distribución y características del 10% superior de los hogares de mayores ingresos en Chile, utilizando datos de las encuestas CASEN 2020 y EBS 2021. El objetivo es desarrollar un análisis descriptivo detallado y posteriormente un modelo de machine learning que permita predecir la pertenencia a este grupo.

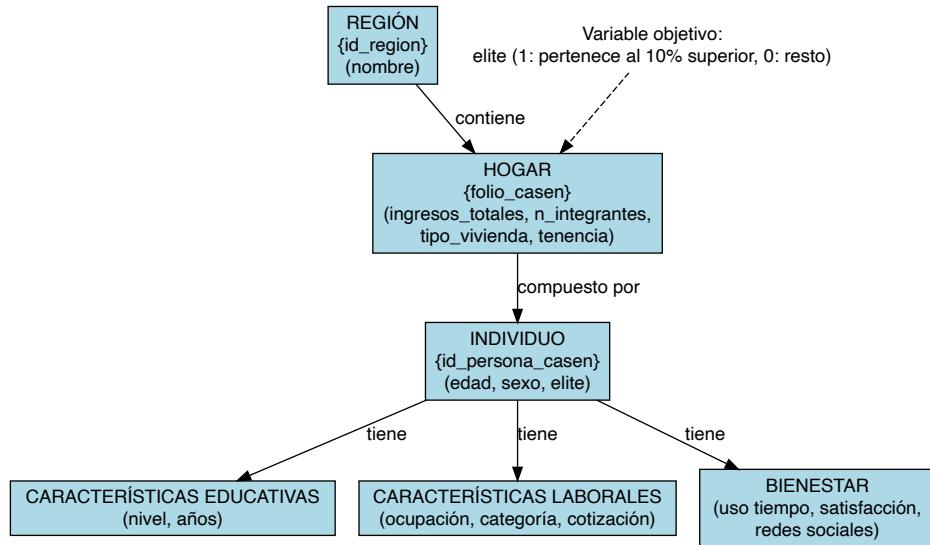
1. Preparación y descripción de los datos Para comenzar el análisis, se cargan las bibliotecas necesarias y se establecen los parámetros de configuración. Se utiliza una combinación de paquetes para manipulación de datos (dplyr), visualización (ggplot2), manejo de datos espaciales (sf, geodata, chilemapas) y presentación de resultados (kableExtra).

2. Carga y preparación inicial de datos El análisis utiliza dos fuentes principales de datos:

CASEN 2020 (versión reducida, en formato rds para poder subirse a GitHub): Proporciona información socioeconómica detallada EBS 2021: Complementa con información adicional y factores de expansión actualizados

En el proceso de preparación, se realizan los siguientes pasos: a. Cálculo de deciles de ingreso para identificar el 10% superior según la muestra total de la encuesta CASEN, no de la submuestra de la EBS b. Creación de variable binaria para el grupo objetivo c. Merge de ambas bases de datos manteniendo la estructura de la CASEN

3. Modelo Entidad-Relación La estructura de datos se organiza en cuatro entidades principales:



4. Análisis descriptivo inicial Para comprender la distribución de ingresos en Chile, comenzamos examinando las estadísticas básicas por decil. Esto nos permite contextualizar el 10% superior dentro de la

Table 1: Estadísticas por decil de ingreso

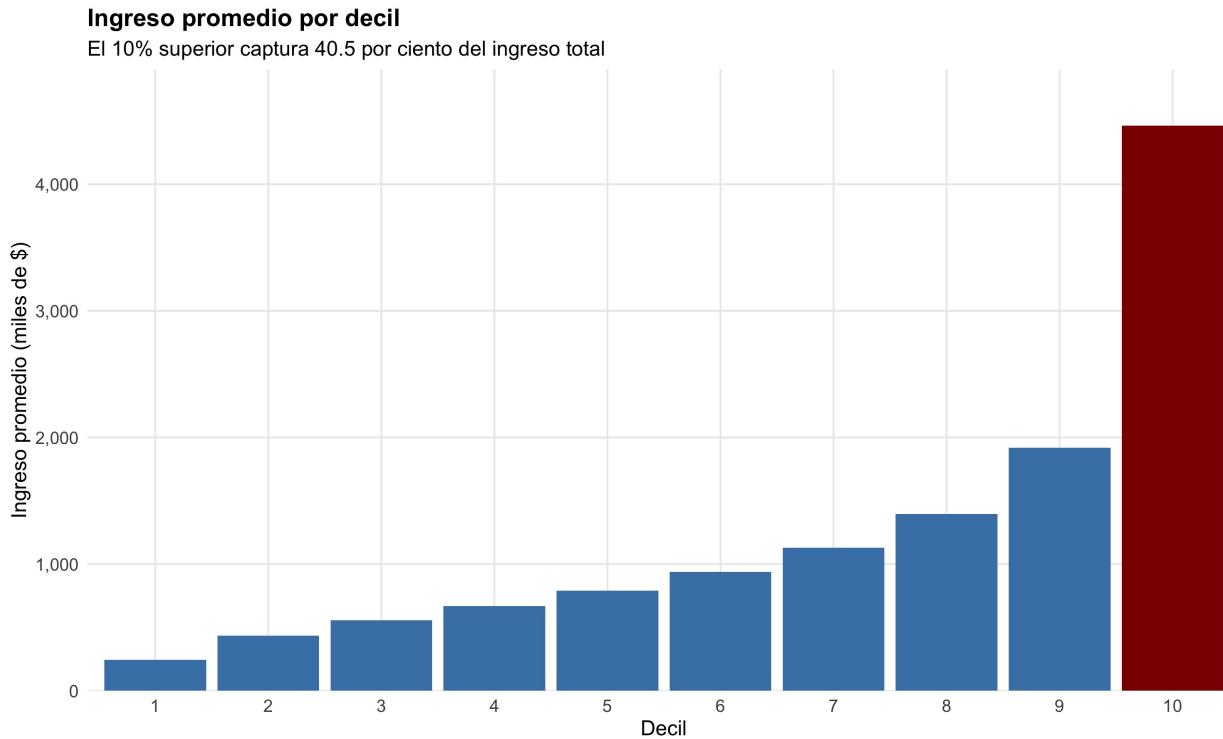
decil	n	n_expandido	ingreso_medio	proporcion_poblacion
1	1117	1325185	243723.2	0.0873880
2	1204	1499566	436750.2	0.0988873
3	1099	1389221	558032.0	0.0916107
4	1093	1443258	667421.0	0.0951742
5	1096	1423513	790322.5	0.0938721
6	1082	1541201	937046.2	0.1016329
7	1100	1597200	1130165.5	0.1053257
8	1104	1533102	1397455.4	0.1010988
9	1052	1552609	1917867.4	0.1023852
10	974	1859535	4460647.5	0.1226251

distribución general de ingresos del país.

Tenemos, para nuestro caso positivo, 974 casos de una muestra de 10921, sin NAs, esto es, alrededor de un 9% de la muestra, lo que se considera suficiente para informar un modelo de Machine Learning.

5. Visualización de la distribución por deciles El gráfico siguiente muestra la marcada desigualdad en la distribución de ingresos en Chile. Se destaca el porcentaje del ingreso total que captura ese 10%.

```
## pdf
## 2
```



6. Caracterización del 10% superior vs resto de la población La tabla siguiente compara características clave entre este grupo y el resto de la población. Se incluyen variables territoriales (región, urbano/rural), demográficas (edad, sexo), socioeconómicas (educación) y de bienestar. Los valores están

ponderados usando los factores de expansión provistos por la EBS: seleccionamos éstos en vez de los de CASEN por tratarse de una encuesta bifásica.

\begin{table} \caption{Características del 10% superior vs resto}

elite_label	n	n_expandido	ingreso_promedio	desv_est	edad_promedio	prop_hombres	prop_urbano
10% superior	974	1,859,535	4,460,648	2,993,049	41.27	0.56	0.95
Resto	9,947	13,304,855	917,774	500,956	44.87	0.48	0.87

\end{table}

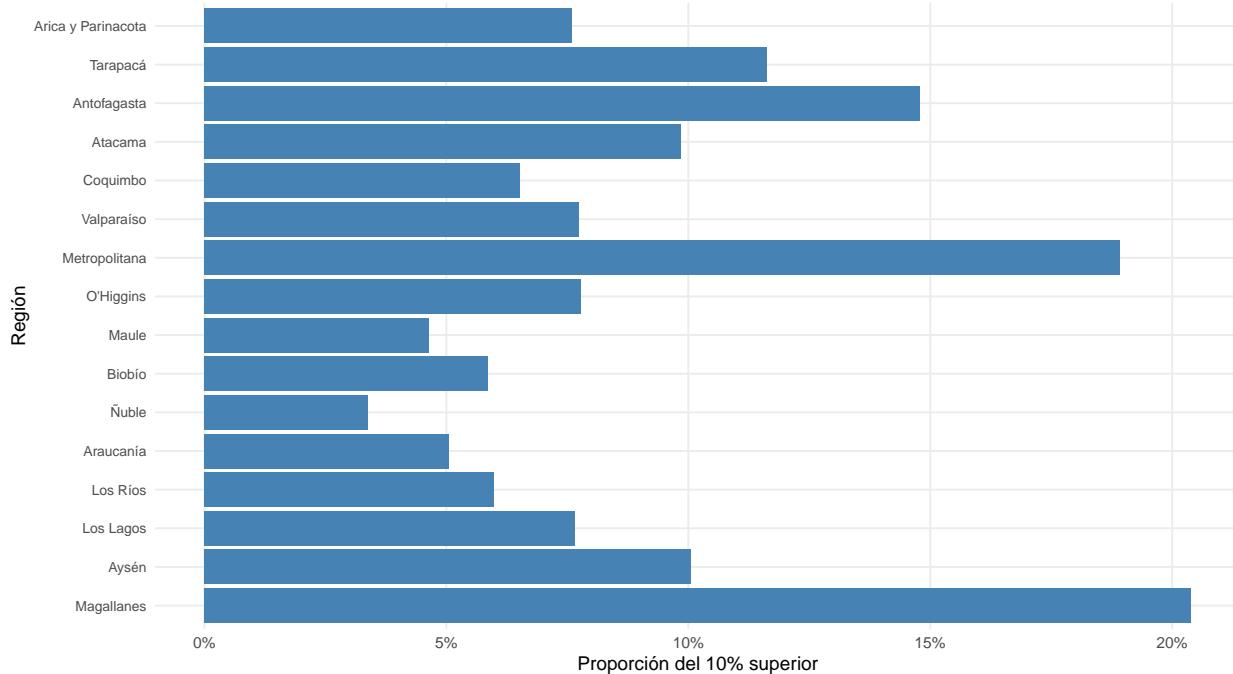
7. Análisis territorial El análisis territorial se desarrolla en dos dimensiones complementarias:

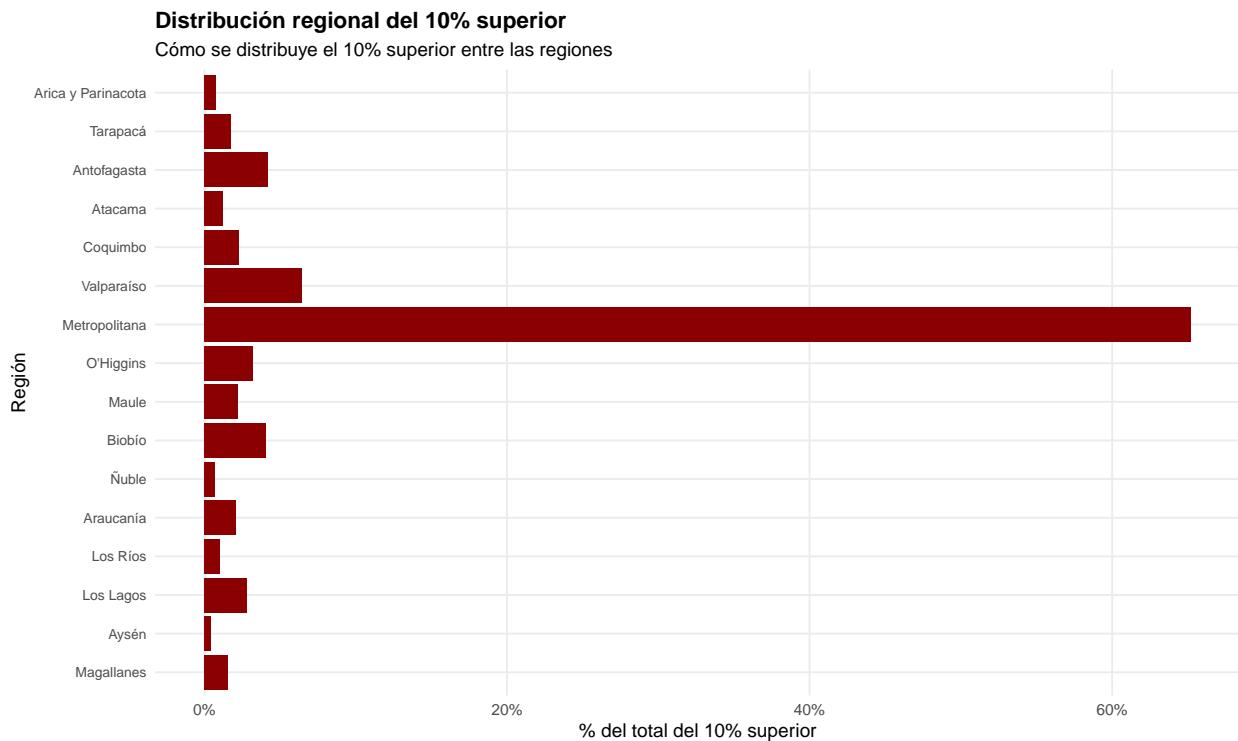
Proporción interna: Qué porcentaje de la población de cada región pertenece al 10% superior Distribución nacional: Cómo se distribuye el total del 10% superior entre las regiones

Para facilitar la interpretación, se ordenan las regiones de norte a sur.

Proporción del 10% superior por región

% de habitantes de cada región que pertenece al 10% superior





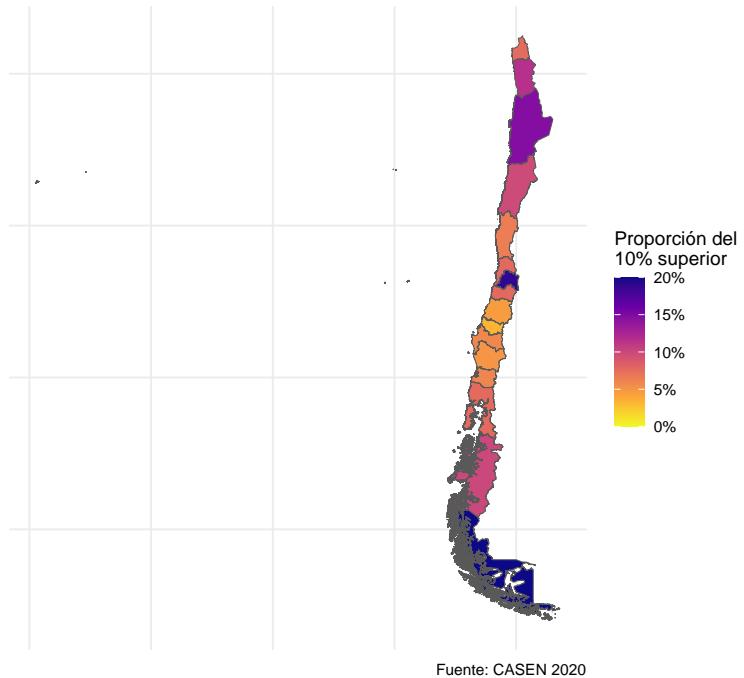
Los gráficos resultantes revelan patrones interesantes. La proporción de elite dentro de cada región (gráfico azul) muestra una concentración en ciertas áreas. La distribución del total de la elite (gráfico rojo) evidencia una fuerte centralización

8. Visualización espacial Aquí generamos mapas que permiten una visualización más intuitiva de los patrones espaciales. Se utilizan dos mapas que corresponden a las mismas dimensiones analizadas en los gráficos de barras:

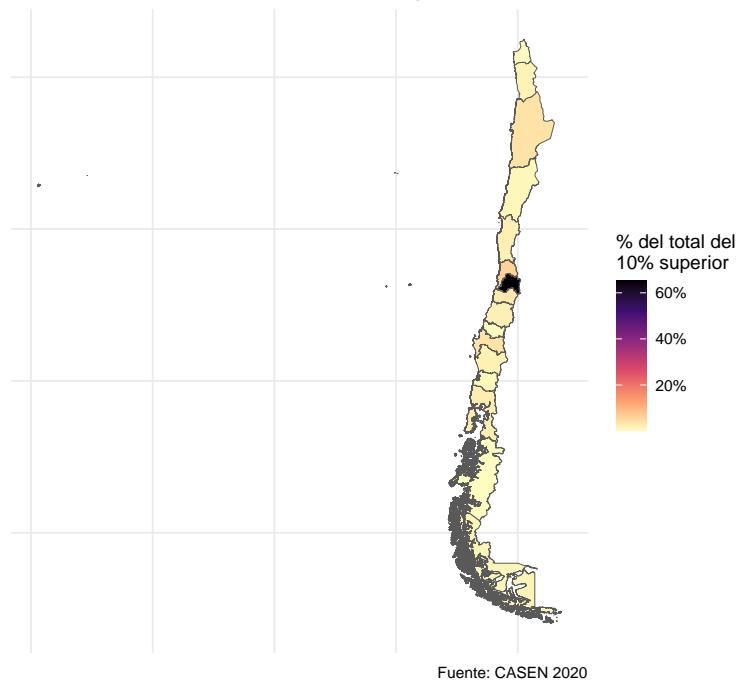
El primer mapa muestra la proporción de habitantes de cada región que pertenece al 10% superior. Este mapa ayuda a identificar dónde es más probable encontrar miembros de la elite económica. El segundo mapa (en tonos magma) visualiza cómo se distribuye el total de ese 10% entre las regiones.

Para la construcción de estos mapas, se enfrentaron varios desafíos técnicos: La necesidad de compatibilizar diferentes codificaciones de regiones: Magallanes tenía problemas de visualización que lo convertía persistentemente en NA.

Proporción del 10% superior por región
% de habitantes de cada región que pertenece al 10% superior



Distribución regional del 10% superior
Cómo se distribuye el 10% superior entre las regiones



9. Sexo jefe/a de hogar

Recodificación para identificar sexo de jefe de hogar. No es posible saberlo para el 100% de los casos, pero con variables de sexo y de relación de quien responde con jefe/a de hogar, se puede reducir para cerca del 70% de los casos.

Table 2: Distribución y proporción de elite por sexo del jefe de hogar

Sexo	N	N expandido	Proporción elite	Proporción muestra	Proporción población
Mujer	2710	3129383	6.2%	0.525	0.454
Hombre	2454	3757075	14.1%	0.475	0.546

Table 3: Distribución y proporción de elite por grupo de edad

Grupo de edad	N	N expandido	Proporción elite	Proporción muestra	Proporción población
18-29	293	435288	2.2%	0.057	0.063
30-34	397	670537	19.5%	0.077	0.097
35-39	433	694170	17.3%	0.084	0.101
40-44	486	770243	13.2%	0.094	0.112
45-49	564	674818	14.3%	0.109	0.098
50-54	584	717096	9.3%	0.113	0.104
55-59	593	782963	8.9%	0.115	0.114
60-64	563	625209	8.7%	0.109	0.091
65-69	466	552611	6.5%	0.090	0.080
70-74	361	436181	5.5%	0.070	0.063
75-79	248	282900	5.4%	0.048	0.041
80 o más	176	244442	0.3%	0.034	0.035

```
## [1] "Distribución de parentesco en la muestra:"  
  
##  
##   1   2   3   4   5   6   7   8   9   10  11  12  13  14  
## 5164 2442  18 1078 1331  65 123  34 119 234 138  20  98  57  
  
## [1] "\nDistribución del sexo del jefe de hogar (incluyendo inferidos):"  
  
##  
##   1   2 <NA>  
## 4357 3267 3297  
  
Sexo del/la jefe de hogar  
  
## [1] "Distribución de parentesco en la muestra:"  
  
##  
##   1   2   3   4   5   6   7   8   9   10  11  12  13  14  
## 5164 2442  18 1078 1331  65 123  34 119 234 138  20  98  57  
  
## [1] "\nNúmero de hogares (debería coincidir con jefes):"  
  
## [1] 5164
```

Edad

10. Nivel educacional

Table 4: Distribución y proporción de elite por nivel educacional

Nivel educacional	N	N expandido	Proporción elite	Proporción muestra	Proporción población
Hasta básica	2453	2879582	1.3%	0.225	0.190
Hasta media	4613	6227468	4.3%	0.422	0.411
Postgrado	188	334451	58.7%	0.017	0.022
Técnica superior	1302	1879090	9.9%	0.119	0.124
Universitaria	2365	3843799	30.5%	0.217	0.253

Table 5: Análisis de variables predictoras potenciales

variable	elite_label	media	desv	na_prop
edad	10% superior	41.267	15.089	0.000
edad	Resto	44.872	17.755	0.000
sexo	10% superior	1.441	0.496	0.000
sexo	Resto	1.521	0.500	0.000
zona	10% superior	1.050	0.218	0.000
zona	Resto	1.125	0.331	0.000
region	10% superior	10.780	3.740	0.000
region	Resto	9.529	3.959	0.000
e6a	10% superior	13.809	2.498	0.000
e6a	Resto	10.317	3.013	0.000
v13	10% superior	1.313	0.962	0.000
v13	Resto	1.807	1.691	0.000
v13_propia	10% superior	1.531	0.533	0.201
v13_propia	Resto	1.180	0.403	0.363

Variables de bienestar:

Resumen variables potencialmente predictoras Para preparar la fase de modelamiento, se realiza un análisis exploratorio de las variables que podrían predecir la pertenencia al 10% superior. El análisis de estas variables se realiza considerando:

Su distribución diferenciada entre elite y no elite La presencia de valores faltantes que podrían afectar el modelamiento La necesidad de transformaciones o recodificaciones para su uso en modelos predictivos

La tabla resultante proporciona una primera aproximación a la capacidad predictiva de cada variable, mostrando diferencias significativas en varias dimensiones entre el grupo elite y el resto de la población.