

# Predicción de la percepción de sismos por la población a partir de datos básicos de terremotos

Trabajo final de la asignatura  
“Taller de tesis”



**Maestría en Explotación de Datos y  
Descubrimiento del Conocimiento**



Facultad de Ciencias Exactas y Naturales  
Facultad de Ingeniería

**Autor:**  
Víctor A. Bettachini

**Fecha:**  
4 de julio de 2024

## Resumen

Una decena de terremotos se producen en el territorio nacional en forma diaria sin que sean detectados más que por instrumentos. Son muy pocos los casos en que las ondas sísmicas que estos producen son percibidos por la población. Este trabajo busca cuantificar la capacidad de predecir esta eventualidad por dos herramientas de la ciencia de datos usando solo datos básicos de terremotos.

# Índice general

|  |           |
|--|-----------|
| <b>1. Introducción</b>   | <b>3</b>  |
| 1.1. Contexto y motivación científica . . . . .                            | 3         |
| 1.2. Objetivos del trabajo / Pregunta . . . . .                            | 4         |
| 1.3. Estructura del documento . . . . .                                    | 4         |
| <b>2. Marco teórico</b>  | <b>6</b>  |
| 2.1. Relevamiento de trabajos previos y relevantes . . . . .               | 6         |
| 2.2. Conceptos y técnicas de ciencia de datos utilizados en el trabajo     | 7         |
| 2.2.1. Regresión logística para la predicción binaria . . . . .            | 7         |
| 2.2.2. XGBoost para una predicción binaria . . . . .                       | 8         |
| <b>3. Metodología</b>  | <b>9</b>  |
| 3.1. Presentación y descripción de los datos utilizados . . . . .          | 9         |
| 3.1.1. Unidad de magnitud en el conjunto de datos . . . . .                | 10        |
| 3.2. Adquisición y formateo de los datos . . . . .                         | 11        |
| 3.2.1. Carga y verificación de faltantes o duplicados . . . . .            | 11        |
| 3.2.2. Inspección y formateo de datos . . . . .                            | 11        |
| 3.3. Delimitación del espacio geográfico . . . . .                         | 11        |
| 3.4. Análisis exploratorio de datos (AED) . . . . .                        | 14        |
| 3.5. Ingeniería de características en función de la física del sistema . . | 17        |
| 3.5.1. Linealización de la magnitud . . . . .                              | 17        |
| 3.5.2. Descarte de terremotos de poca profundidad . . . . .                | 18        |
| 3.6. Variables a correlacionar con la percepción . . . . .                 | 19        |
| 3.7. Preprocesamiento . . . . .  | 20        |
| 3.8. Métricas de evaluación de los modelos . . . . .                       | 21        |
| <b>4. Resultados y discusión</b>   | <b>23</b> |
| 4.1. Presentación de resultados . . . . .                                  | 23        |
| 4.1.1. Predictor por regresión logística . . . . .                         | 23        |
| 4.1.2. Predictor por XGBoost . . . . .                                     | 24        |
| 4.2. Relevancia de los resultados . . . . .                                | 24        |
| 4.3. Limitaciones y posibles mejoras . . . . .                             | 24        |
| <b>5. Conclusión</b>   | <b>25</b> |
| 5.1. Resumen de los hallazgos principales . . . . .                        | 25        |
| 5.2. Conclusiones generales y su relación con los objetivos del trabajo    | 25        |
| 5.3. Aplicaciones y relevancia de los resultados . . . . .                 | 25        |
| <b>Bibliografía</b>  | <b>26</b> |

|   |           |
|---|-----------|
| <b>Anexos (opcionales)</b>                            | <b>29</b> |
| 5.4. Código fuente utilizado en el análisis . . . . . | 29        |
| 5.5. Tablas y gráficos adicionales . . . . .          | 29        |
| 5.6. Otros materiales relevantes . . . . .            | 29        |

# Capítulo 1

## Introducción

### 1.1. Contexto y motivación científica

Un rápido desprendimiento entre dos facetas enfrentadas de sendas placas tectónicas que se traban mutuamente su desplazamiento relativo produce una rápida liberación de energía que se denomina terremoto. Esto sucede a cierta profundidad en la corteza terrestre en el punto denominado hipocentro a partir del cual parte de esta energía se encauza como ondas elásticas. El estudio de estas ondas es el área llamada sismología y de ahí el termino sismo para un evento particular detectado, pero que debiera aclararse si se produjo por un terremoto u otra fuente de ondas [1, sección 4.1.1].

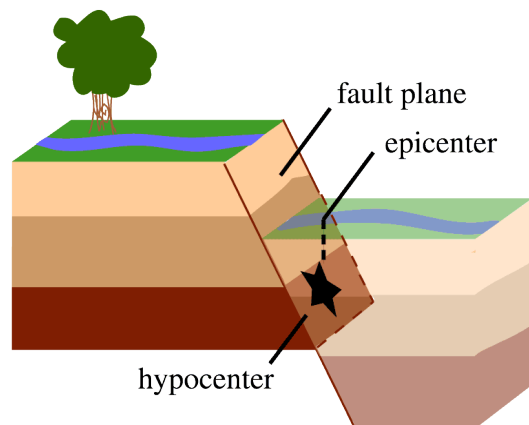


Figura 1.1: El epicentro es el punto de la superficie inmediatamente sobre el hipocentro que es el punto a cierta profundidad al que se adscribe un terremoto. Reproducido de [2].

Sea que las ondas de compresión longitudinal de la corteza, las tipo P, o las del tipo S transversales y más lentas, ambas arriban con mayor intensidad al punto de la superficial terrestre que se encuentra directamente sobre el hipocentro, que se denomina epicentro [1, sección 4.1.2] como ilustra la figura 1.1.

Cuanto más próximo es un punto en la superficie al hipocentro, la amplitud de las ondas sísmicas registradas en la superficie es mayor. Tanto esta amplitud como el período de oscilación son mucho mayores que el de otros desplazamientos de la corteza como los de las mareas solares y lunares de la corteza [1, sección 4.1.4]. Como resultado, estructuras artificiales pueden agitarse poniendo en riesgo su estabilidad estructural y haciendo caer elementos que no estaban fijados a esta o perdieron tal adhesión a causa de la agitación misma. Como consecuencia, los sismos más fuertes pueden generar graves daños, poniendo en riesgo la integridad física y la seguridad de las personas al generar daños en las viviendas y edificios, derrumbes de puentes, rompimiento de vidrios, entre otros [3].

La mayor parte de los sismos presentan ondas de pequeña amplitud y no generan daños materiales. Que sean o no detectados por la población es un factor relevante en su percepción de confianza vis-à-vis de los organismos de monitoreo y prevención de riesgos. Cuando informar a la población de la ocurrencia de un sismo es una decisión de política pública que debiera apuntar a no alertar innecesariamente sobre sismos menores imperceptibles [4]. El caso inverso es también problemático, en que ante una carencia de una comunicación oficial de la poca importancia de un evento sísmico llevó a la autoevacuación por parte de la población que lo percibió [5].

## 1.2. Objetivos del trabajo / Pregunta

Contar con una estimación rápida a partir de los datos sísmicos registrados por instrumental de si un dado evento será percibido por la población o no permitiría a las autoridades tomar decisiones informadas sobre la comunicación a la población. Este trabajo busca determinar el grado de certeza con que ciertos métodos de ciencias de datos pueden predecir si la población percibirá actividad sísmica producto de terremotos a partir de unos pocos datos básicos sobre los mismos publicados por el Instituto Nacional de Prevención Sísmica (INPRES) que es el organismo público de la República Argentina que realiza estudios e investigaciones básicas y aplicadas de sismología [6].

## 1.3. Estructura del documento

Se estructuró en los siguientes capítulos con temáticas diferenciadas

**Introducción** : se presentará el contexto y la motivación científica, los objetivos del trabajo y la estructura del documento.

**Marco teórico** : se revisarán los trabajos previos y relevantes, se presentarán los conceptos y técnicas de ciencia de datos utilizados en el trabajo.

**Metodología** : se describirán los datos utilizados, el preprocesamiento y limpieza de los mismos, el análisis exploratorio de los datos, las técnicas de análisis y modelado utilizadas, la selección de características, las métricas de evaluación de los modelos y los métodos estadísticos utilizados.

**Resultados y discusión** : se presentarán y analizarán los resultados obtenidos, se discutirán los resultados y su relevancia, se identificarán las limitaciones y posibles mejoras.

**Conclusión** : se resumirán los hallazgos principales, se presentarán las conclusiones generales y su relación con los objetivos del trabajo, se discutirán las aplicaciones y relevancia de los resultados.

## Capítulo 2

# Marco teórico

### 2.1. Relevamiento de trabajos previos y relevantes

En la prensa general se busca transmitir la importancia de un fenómeno sísmico informando un valor sea en la escala de Richter o de Mercalli como si fueran intercambiables. La primera se refiere a la magnitud, una característica física del terremoto (ver definición en la sección 3.5), en tanto que la segunda se refiere a una caracterización subjetiva del sismo, la llamada intensidad. Esta última cantidad es función de fenómenos percibidos por el público como si hubo grietas en el suelo o vidrios o si temblaron edificios [1, sección 4.2.3]. Los organismos de monitoreo de sismos cuentan con portales en la world-wide web para que el público ofrezca tal información voluntariamente en formularios con preguntas de elección múltiple. Tal es el caso de los portales “Encuesta de sismos” del INPRES [7] y “Felt Report - Tell Us!” [8] parte del programa de riesgo de terremotos del Geological Survey de los Estados Unidos de América (USGS) [9]. Este último aporta a la sistematización de reportes del público denominado “Did you feel it?” (DYFI) que ha nutrido en la última década y media a varios estudios relacionando la intensidad con diversos fenómenos citados en [10].

La intensidad expresada en la escala modificada de Mercalli (MMI) [11] en la base de datos DYFI mostró estar bien correlacionada con medidas de espectro de oscilación lo que impulsó a buscar un modelo que le relacione con parámetros físicos de terremoto. Una regresión múltiple basada en el método de máxima verosimilitud se encontraron los factores  $c_1...7$  de la [12, ecuación 1]

$$MMI = c_1 + c_2(M_w - 6) + c_3(M_w - 6)^2 + c_4 \log R + c_5 R + c_6 B + c_7 M_w \log R, \quad (2.1)$$

donde  $M_w$  es la magnitud del momento, según la definición discutida en la sección 3.1.1,  $R = \sqrt{D + h}$  es la distancia entre observador e hipocentro en función de la distancia del primer al epicentro  $D$  y profundidad del segundo  $h$ , y  $B$  es nulo para  $R < R_t$  una distancia de transición en la forma de atenuación de las ondas. Una posterior actualización de este modelo por los mismos autores se basó en agrupar los valores promedio de MMI en cada código postal

$$MMI = c_1 + c_2 M_w + c_3 \log R + c_4 R + c_5 B + c_6 M_w \log R, \quad (2.2)$$

con  $R = \sqrt{D + h + 14^2}$  donde el factor adicional,  $14^2$  da cuenta de la saturación a cortas distancias permitiendo simplificar el factor  $B = \max(0, \log(R/50))$  que ya no requiere establecer un valor de transición  $R_t$  [13].

Lo anterior muestra que hay antecedentes de que un ajuste permite predecir la percepción de los mismos por parte de la población a partir datos físicos de los terremotos. Para este trabajo lo que busca determinarse es una decisión binaria entre lo que correspondería a la intensidad 0 y 1 de la escala MMI, respectivamente una percepción solo con instrumentos o una por al menos una fracción de la población [11]. No es el enfoque a seguir discriminar con un umbral tras una predicción del valor MMI sino utilizar las técnicas de aprendizaje automático que generen una decisión binaria directamente desde los datos de parámetros físicos.

## 2.2. Conceptos y técnicas de ciencia de datos utilizados en el trabajo

Se utilizan en este trabajo dos técnicas con aplicación a la predicción pero en sendos extremos de complejidad y transparencia en cuanto al peso relativo que tienen las variables independientes en el resultado, el de regresión logística y el de máquinas de potenciación de gradiente, más conocidos por su nombre en inglés Gradient Boosting Machines (GBM).

### 2.2.1. Regresión logística para la predicción binaria

Lo siguiente es un resumen del material de la novena clase de la asignatura *Enfoque estadístico del aprendizaje* titulada *Regresión Logística* elaborado por Juan Barriola, Azul Villanueva y Franco Mastelli.

Un modelo de regresión lineal con coeficientes  $\beta_j$  para cada variable  $X_j$  que busque la probabilidad de una dependiente binaria  $P(Y)$

$$P(Y) = \beta_0 + \sum_{j=1}^p \beta_j X_j, \quad (2.3)$$

no presentaría un punto de corte claro para clasificar los datos en dos categorías. Para sortear esta dificultad se toma la salida de esta regresión como la variable dependiente de una regresión logística,

$$P(Y|X) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}}, \quad (2.4)$$

lo que asegura un valor entre 0 y 1. De esta expresión puede arribarse a

$$\log\left(\frac{P(x)}{1 - P(x)}\right) = \beta_0 + \sum_{j=1}^p \beta_j X_j, \quad (2.5)$$

cuyo lado izquierdo es la función *logit*, la inversa de lo que sería la logística de la probabilidad  $(1 + e^{-P(x)})^{-1}$ .



Para implementar tal función en código en lenguaje *R*, el utilizado en este trabajo, se hace uso de la función *glm* provista por la biblioteca *glmnet* [14]. Esta tiene por argumentos *formula* y *data*, los mismo que la popular función *lm* incluida en el paquete base de *R* para modelos lineales. Pero como *glmnet* genera modelos lineales generalizados, requiere un argumento adicional, *family*, para indicar la distribución del error de la variable a predecir:

- Binomial: link=logit
- Poisson: link=log
- Gaussiana: link=identidad

La correspondiente función de enlace (link) relaciona el modelo lineal con función de probabilidad. Como se busca predecir un resultado booleano se indica la distribución binomial.

Aunque se parte de pocas variables independientes, se ensayará forzar una mayor simplificación del mismo mediante la técnica de regularización tipo Lasso (L1) que fuerza a que los coeficientes de las variables independientes que menos aportan a predicción vayan anulándose.

### 2.2.2. XGBoost para una predicción binaria

En otro extremo entre las herramientas de ciencia de datos por su complejidad se ensayará utilizar máquinas de potenciación de gradiente, más conocidos por su nombre en inglés *Gradient Boosting Machines* (GBM) para la predicción. Las distintas implementaciones de estos algoritmos, como *XGBoost*, *LightGBM* o *CatBoost* son capaces de producir un único modelo con fuerte poder predictivo a partir de la síntesis de resultados de modelos de predicción débiles, típicamente árboles de decisión. Para este trabajo se utiliza XGBoost implementado en lenguaje *R* por la biblioteca *xgboost* [15].

## Capítulo 3

# Metodología

### 3.1. Presentación y descripción de los datos utilizados

En el marco de los *Proyectos de Asistencia Estadística* del *Instituto de Cálculo* (IC) de la *Facultad de Ciencias Exactas y Naturales* (FCEyN) de la *Universidad de Buenos Aires* (UBA) se publicaron conjuntos de datos en un repositorio curado con el objeto de ser aplicados a la enseñanza de la estadística y la ciencia de datos por Daniela Parada, investigadora del IC [16]. De estos conjuntos el utilizado en este trabajo es el que se publica en el apartado “Visualización” que corresponden a datos de sismos de Argentina de la última década [17]. En este repositorio alojado por la firma GitHub, se provee un front-end html que da un contexto, hace una exploración inicial, un análisis para una provincia en particular, muestra una estimación de probabilidad y provee otra información sobre los datos.

Los datos corresponden a detecciones por parte de estaciones de monitoreo sísmico en la República Argentina recopilados y publicados por el INPRES en su sitio web [18]. En el sitio de publicación de los datos se indica que el conjunto de datos comprende las fechas desde el 7 de enero de 2012 hasta el 18 de mayo de 2022 y fue realizado con datos *scrappeados* del buscador de sismos del INPRES por Gustavo Juantorena [17, sección 4.1].

Allí mismo se describe que el conjunto de datos reducido y curado denominado “sismos”, el que se utilizó en este trabajo, es accesible a través de la importación de la biblioteca `datosIC` en lenguaje R [17, sección 5.1.1]. Este mismo conjunto reducido puede descargarse en formato de valores numéricos separado por comas (CSV) apuntando a su URL en el repositorio alojado en GitHub [19].

Las variables reportadas para cada sismo son:

- *Fecha*: en el formato `verb'aaaa-mm-dd'` de la norma ISO 8601 [20].
- *Hora*: una cadena de caracteres en formato `hh:mm:ss` con una exactitud al segundo.
- *Latitud*, *Longitud*: un número con una exactitud de un decimal con grados como unidad.

- *Provincia*: cadena de caracteres del nombre de la provincia donde se produjo el sismo (no donde se ubicó quién potencialmente lo percibiera) según se afirma en el sitio de publicación [17, pág. 5.1.1].
- *Magnitud*: un número función de un logaritmo de la amplitud de las ondas sísmicas, específicamente la escala de magnitud de momento ( $M_w$ ) (leer discusión en el párrafo siguiente).
- *Profundidad*: un número entero con kilómetros como unidad que indico que tan bajo la superficie se ubicó el epicentro.
- *Percibido*: variable booleana de si hubo reportes de percepción del fenómeno por parte de la población,

Esta última variable es la que se busca predecir en este trabajo en función de las demás.

### 3.1.1. Unidad de magnitud en el conjunto de datos

Hay una multitud de especificaciones derivadas de la escala logarítmica originalmente propuesta por Richter en 1935 [1, sección 4.2.3]. Sin embargo no hay una especificación de en cual de estas se expresa la *magnitud* del conjunto de datos en [17]. Tampoco hay una nota adjunta a los datos que lo indique en el repositorio [19]. Asimismo, en la fuente original de donde se hizo el previamente mencionado *scrapping*, el buscador de sismos del INPRES [18], al solicitar datos solo se indica el término *magnitud* sin más detalle.

A pesar de carecer de una indicación explícita al respecto se opta por asumir que la magnitud en el conjunto de datos corresponde a la escala de magnitud de momento ( $M_w$ ) que es la más comúnmente utilizada en la actualidad como se detalla en el párrafo siguiente. Avala tal suposición que en la página *Cálculo de la Magnitud* de la sección de educación del sitio del INPRES se la presenta como una escala englobadora [21]. Pero se considera un avala más crucial un documento de la *Comisión de trabajo de gestión de riesgo* en que el INPRES figura no solo como el *organismo con responsabilidad operativa* sino también entre los *organismos que generan información de base* en [22, anexo X] detalla “actualmente la más utilizada es la magnitud momento,  $M_w$ ”.

**Unidad de magnitud más corrientemente utilizada** A medida que se fueron instalando más estaciones sismográficas en todo el mundo, se hizo evidente que el método desarrollado por Richter sólo era estrictamente válido para determinados rangos de frecuencia y distancia. Para aprovechar el creciente número de estaciones sismográficas distribuidas por todo el mundo, se desarrollaron nuevas escalas de magnitud que son una extensión de la idea original de Richter en adelante denominada  $M_L$ . Entre ellas se incluyen la magnitud de onda de cuerpo,  $m_b$  y la magnitud de onda de superficie  $M_s$  (ver sección ??). Cada una de ellas es válida para un rango de frecuencias y un tipo de señal sísmica concretos. En su rango de validez, cada una es equivalente a la magnitud Richter. Debido a las limitaciones de las tres escalas de magnitud,  $M_L$ ,  $m_b$  y  $M_s$ , se desarrolló una nueva extensión de la escala de magnitud, conocida como magnitud de momento o  $M_w$ , de aplicación más uniforme. En particular, para los terremotos de gran magnitud,  $M_w$  ofrece la estimación más fiable de la importancia de mismo [23].

## 3.2. Adquisición y formateo de los datos

Puesto que el conjunto de datos es curado por un equipo de investigación de la UBA, se asume que los mismos son confiables y que no se requiere de un proceso de limpieza de los mismos. De todas formas se realizaron las verificaciones usuales cada vez que se utilizan datos tabulares en un estudio de estadístico y/o de ciencia de datos.

### 3.2.1. Carga y verificación de faltantes o duplicados

Tras descargar el archivo de datos en formato CSV se le importó en una estructura de datos *data.table* de en un entorno de trabajo en lenguaje R. Esta estructura de datos permite una consulta de los datos análoga a la del lenguaje SQL de bases de datos relacionales lo que le hace una herramienta versátil para el análisis de datos tabulares [24].

La presencia de valores faltantes indicados con el símbolo NA se descartó cuando la ejecución `sum(is.na(sismos_arg))` arrojó un cero como resultado. Por el contrario ejecutar `sismos_arg[duplicated(sismos_arg, fromLast = TRUE)]` mostró unos 23 registros duplicadas sobre un total de 55817 registros.

En una nueva tabla con nombre más corto, `sismos` se copiaron los registros sin duplicados ejecutando `sismos_arg[!duplicated(sismos_arg)]`.

### 3.2.2. Inspección y formateo de datos

Ejecutar la función `colnames` con la *data.table* denominada `sismos_arg` como argumento permitió verificar que contuviera las columnas con los nombres anunciados en el sitio que publica los datos en su sección [17, Exploración inicial]. Las transformaciones o ingeniería de características que se detallan luego en esta sección se realizaron en función de los tipos de datos de cada columnas constatados con la función `str`.

**Fecha y hora del sismo** A partir de la variable Hora se escribió una función que genera otra continua contando el número entero de segundos transcurridos desde la medianoche del día en que se produjo el terremoto, *Segundos en el día*, apoyandose para esto en la función `strptime` del paquete base de R. Esto habilita posibles análisis segmentando el día en franjas horarias.

La fecha se pasó del formato ISO 8601 a dos variables. Una guarda el número consecutivo de día del año, *Día del año*, lo que permite segmentar por rangos del año. La otra simplemente guarda *Año*, para hacer contrastaciones interanuales.

Resumiendo las nuevas características:

- *Segundos en el día*: número entero de 0 a 86399
- *Día del año*: número entero de 1 a 366
- *Año*: número entero de 2012 a 2022

## 3.3. Delimitación del espacio geográfico

**Distribución geográfica de los datos** Como es esperable la mayor parte de los hipocentros se ubican en regiones con orografía elevada, como la cordillera

de los Andes, producto de la subducción de la placa de Nazca bajo la placa Sudamericana, o las sierras cordobesas producto de procesos mucho más antiguos. Una ubicación de los mismos sobre un mapa físico lo ilustra en la figura 3.1. Esto conlleva a que la mayor parte de los terremotos estén alejados de las mayores urbanizaciones reduciendo la probabilidad de que sean percibidos por la población.

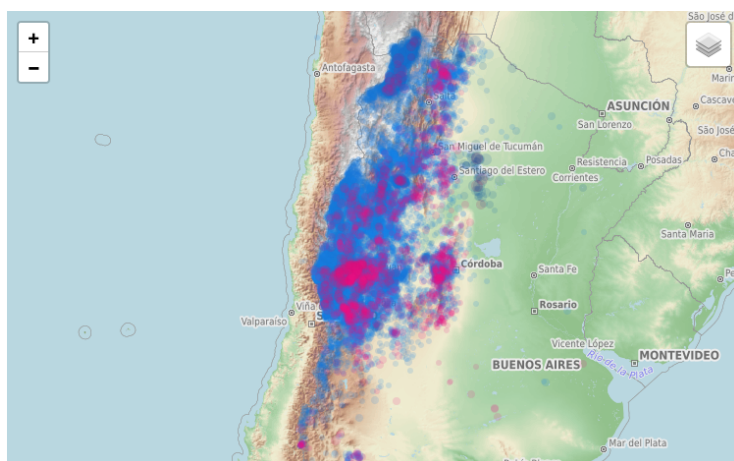


Figura 3.1: Superponer las ubicaciones de los terremotos en un mapa físico muestra que la mayoría de los terremotos se registran en zonas de montaña o cerros en su mayoría alejados de las mayores urbanizaciones lo que afecta negativamente la estadística de los percibidos. La escala de colores muestra los más superficiales en azul y profundos en rojo. Reproducido de [17]

La ubicación de los terremotos sobre un mapa político que muestra la figura 3.2 deja a las claras que no tienen una distribución homogénea aún en provincias con abundante actividad.

**El problema de la distancia** Evidentemente la proximidad de un epicentro a una población es un factor relevante en la percepción de un sismo. El INPRES tiene una metodología operativa para registrar la ubicación de los usuarios que reportan haber percibido un sismo a través de su página web [25]. Su buscador de sismos indica los sismos “sentidos” por usuarios con un color en sus resultados de búsqueda como ilustra la figura 3.3.

Lamentablemente en el conjunto de datos curados no figura la ubicación del sismógrafo o la población que hizo el registro lo que imposibilita incorporar la distancia al epicentro como una variable en el modelo. Puesto que no se tiene control posible sobre la posición de quienes potencialmente percibieron un sismo se decidió para reducir el impacto del parámetro distancia limitar el espacio geográfico de los terremotos. Las posibles fuentes de datos para realizar tal operación son su latitud, longitud y provincia.

**Elección de la provincia de San Juan** Respondiendo a lo discutido en el párrafo anterior se buscó trabajar en un extensión geográfica relativamente limitada, pero cuidando que el número de sismos restantes sea aún elevado.

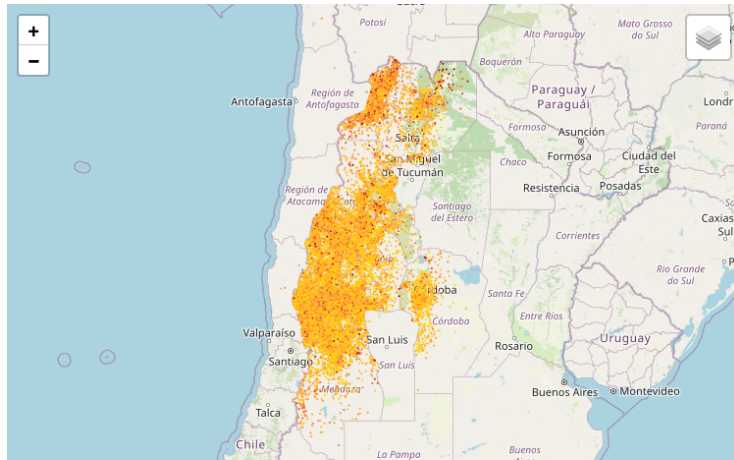


Figura 3.2: Sobre este mapa político se representan las ubicaciones de los terremotos como puntos. Esto permite apreciar que en algunas provincias, como San Luis, se omiten datos como evidencia el corte abrupto de reportes en su frontera. En otras se hacen más claros sesgos geográficos, como en Córdoba, Salta o Jujuy, Reproducido de [17]

Tras una inspección visual de la figura 3.2 se determinó que la provincia de San Juan cumple con tales requisitos, por lo que se procedió a filtrar los datos con el comando `sismos_SJ <- sismos[Provincia == "San Juan"]`. El subconjunto de datos de esta provincia representa un  $\approx 54\%$  de los datos hasta aquí disponibles. Con 29 917 registros es aún un número suficiente para realizar análisis como los métodos elegidos.

**Latitud y longitud** La distribución de terremotos en la provincia de San Juan dista de ser homogénea pues hay una preponderancia de las ubicaciones meridionales como puede apreciarse en la figura 3.4.

Lo anterior podría hacer sospechar de un sesgo con la latitud, pero la inspección visual de la figura muestra que la proporción de puntos rojos y amarillos, los de terremotos percibidos o no es similar con la que hay en el norte. En función de esto se decidió no eliminar las variable de latitud en este estadio y analizar más adelante su correlación con la percepción de los sismos. A fin de cuentas el objetivo del trabajo es determinar si un modelo puede predecir la percepción de un sismo por parte de la población y no determinar los factores que influyen en la percepción de los sismos por parte de la población. En caso de que haya factores geográficos regionales, estos serán capturados por el modelo de clasificación y no necesariamente por la variable de latitud.

Un razonamiento similar se aplica para no coartar la variable de longitud. Como se comentó al presentar la figura 3.1, son esperables más terremotos y de mayor intensidad en las regiones occidentales de la provincia de San Juan. Nuevamente será una virtud de los modelos si pueden explotar tal información para mejorar la clasificación de los sismos percibidos por la población.



|            |   |                               |                             |                            |                          |               |           |
|------------|---|-------------------------------|-----------------------------|----------------------------|--------------------------|---------------|-----------|
| Sismología | Red Nacional de Estaciones Sismológicas | Red Nacional de Acelerógrafos | Ingeniería Sismorresistente | Laboratorio de Estructuras | INPRES-CIRSOC Reglamento | Publicaciones | Servicios |
|------------|---|-------------------------------|-----------------------------|----------------------------|--------------------------|---------------|-----------|

Buscador de sismos

Resultado de la búsqueda

| 12sismos encontrados según las características especificadas. (Los sismos listados en color rojo han sido sentidos.) |            |          |         |          |          |       |   |           |      |
|--|------------|----------|---------|----------|----------|-------|---|-----------|------|
| Id   | Fecha      | Hora     | Latitud | Longitud | Profund. | Magn. | Intensidad  | Provincia |      |
| 1  | 14/06/2024 | 23:32:49 | -31.860 | -69.663  | 119 Km.  | 2.7   |   | SAN JUAN  | mapa |
| 2  | 14/06/2024 | 22:12:32 | -23.274 | -66.653  | 233 Km.  | 3.5   |   | JUJUY     | mapa |
| 4  | 14/06/2024 | 20:15:52 | -31.897 | -69.960  | 126 Km.  | 2.9   |   | SAN JUAN  | mapa |
| 5  | 14/06/2024 | 18:51:09 | -31.547 | -69.410  | 125 Km.  | 2.8   |   | SAN JUAN  | mapa |
| 6  | 14/06/2024 | 18:44:59 | -28.521 | -67.919  | 112 Km.  | 2.5   |   | LA RIOJA  | mapa |
| 7  | 14/06/2024 | 16:12:04 | -31.250 | -68.664  | 108 Km.  | 3.8   | II a III -Ciudad de San Juan, San Juan; II a III -Albardón, San Juan; II a III -Caucete, San Juan; II a III -Villa San Martín, San Juan | SAN JUAN  | mapa |
| 8  | 14/06/2024 | 14:41:57 | -31.538 | -69.343  | 102 Km.  | 2.8   |   | SAN JUAN  | mapa |
| 9  | 14/06/2024 | 14:02:37 | -31.898 | -69.009  | 111 Km.  | 2.8   |   | SAN JUAN  | mapa |
| 10   | 14/06/2024 | 12:38:36 | -31.818 | -69.715  | 100 Km.  | 2.5   |   | SAN JUAN  | mapa |
| 13   | 14/06/2024 | 04:38:59 | -31.297 | -68.555  | 108 Km.  | 2.8   |   | SAN JUAN  | mapa |
| 14   | 14/06/2024 | 04:15:16 | -23.687 | -66.592  | 227 Km.  | 3.2   |   | JUJUY     | mapa |
| 15   | 14/06/2024 | 03:30:23 | -31.210 | -68.433  | 104 Km.  | 3.1   |   | SAN JUAN  | mapa |

1

volver a Búsquedas

Figura 3.3: Resultados de una búsqueda manual de sismos en el sitio del INPRES. El indicado en rojo fue percibido por la población. En la columna *intensidad* se dan datos de ubicación de la población que lo percibió.

### 3.4. Análisis exploratorio de datos (AED)

Un primer vistazo sobre los datos con `summary(sismos_SJ)` permitió obtener un resumen de las variables numéricas y categóricas. Saltan a la vista que hay valores extremos en el extremo superior de la variable **Magnitud** bastante alejados de la mediana y que hay un fuerte desequilibrio en la variable “Percibido”, la de clase de clasificación, en favor de los terremotos no percibidos.

**Desequilibrio en la clase de clasificación** Sobre el total de registros solo un valor cercano al 98 % fueron percibidos por el instrumental y no por la población. Restan tan solo unos 619 registros que efectivamente fueron percibidos por la población. Este desequilibrio llama al uso de técnicas de balanceo de clases en los modelos de clasificación a utilizar.

**Distribución de la magnitud** La magnitud de terremotos forzosamente presenta una distribución asimétrica por el hecho de que cuanto mayor es la energía liberada, más infrecuente es el fenómeno. Esta relación logarítmica de la frecuencia con la magnitud de los sismos que ilustra la figura 3.5 para San Juan en los años 2012 a 2022 es universal y se conoce como la ley de Gutenberg-Richter [1,

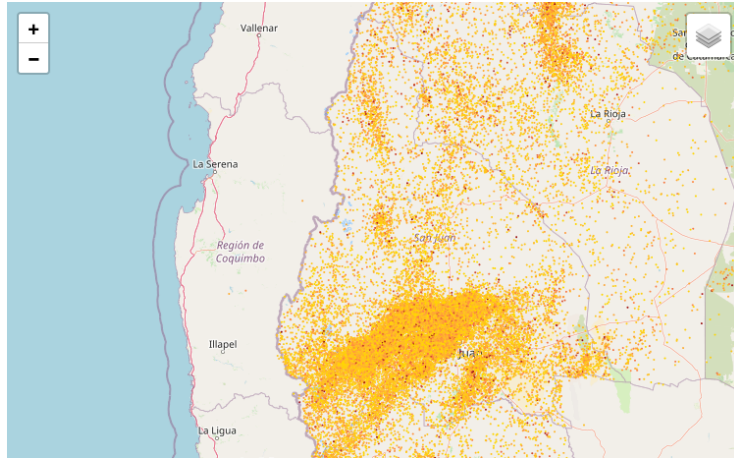


Figura 3.4: Ampliación del área de San Juan en la figura 3.2. Los terremotos que figuran en el conjunto de datos registrados en la provincia de San Juan se presentan mayoritariamente en su región meridional. Los puntos rojos son los percibidos por personal, los amarillos solo por instrumental. Reproducido de [17]

ec. 4.24].

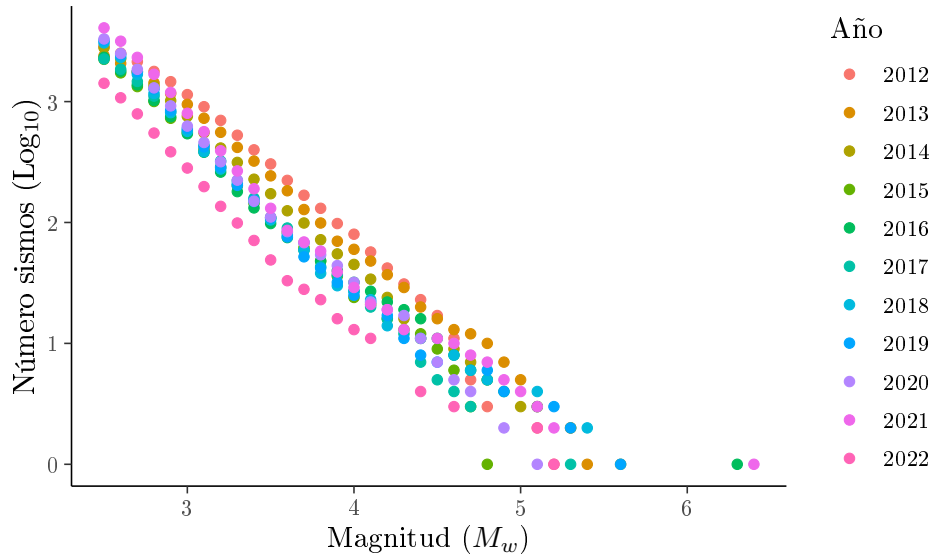


Figura 3.5: Los sismos de mayor magnitud son más infrecuentes. Generado con código provisto junto con los datos [17, sección 4.2.1]

Los sismos percibidos por la población no cumplen esta relación, quedando relegados los de menor magnitud dando cuenta de la dificultad en percibirlos en ese caso. La figura 3.6 parece sugerir que esto es aún más pronunciado para



magnitudes inferiores a  $\approx 4$ .

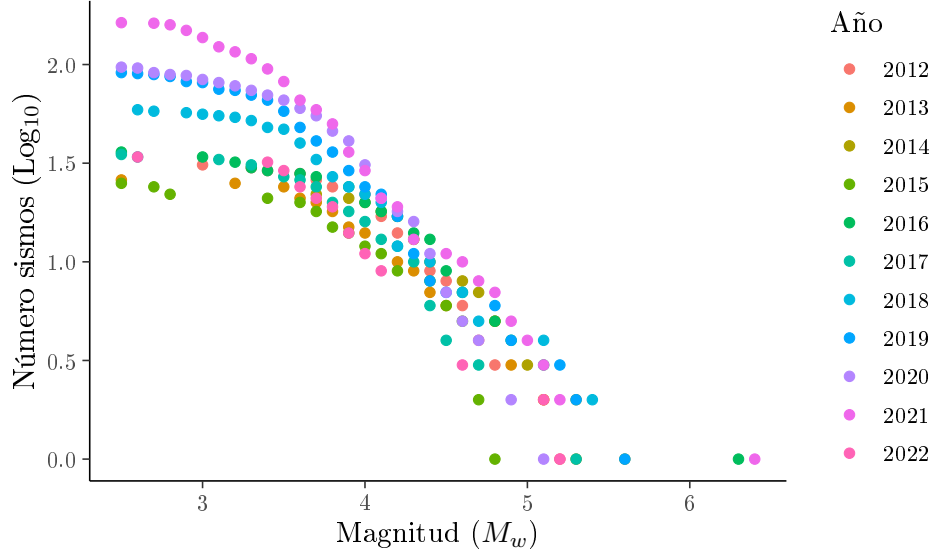


Figura 3.6: El apartamiento de la tendencia de la ley de Gutenberg-Richter en los sismos percibidos por la población muestra que la dificultad para que esto se produzca se incrementa con la baja de la magnitud.

**Horario de los percibidos** Se puede hipotetizar que en los horarios de sueño de la mayoría de la población sea menor la proporción de sismos percibidos. Segmentando la `Hora_decimal` de los registros en intervalos de una hora se calculó la proporción entre casos percibidos o no para graficar las 24 proporciones en la figura 3.7.

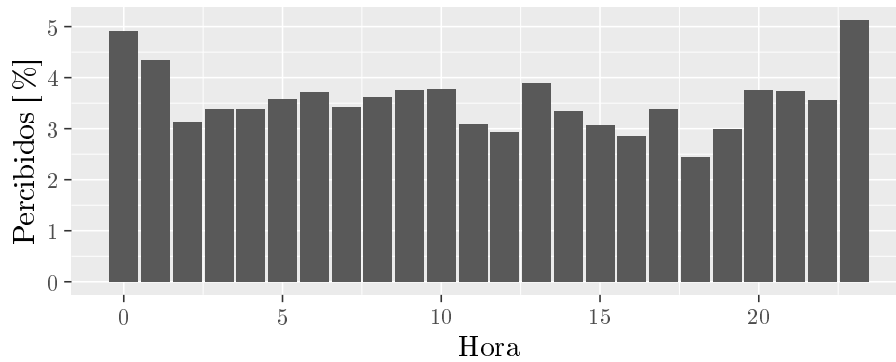


Figura 3.7: Proporción de sismos percibidos por la población en función de la hora del día.

Contrariando la hipótesis, la única merma observada se produce entre las 10 y las 19 horas, y es muy ligera. Donde es claro un cambio llamativo es durante las tres horas que comienzan a las 23 de cada día, en las que hay un incremento de percepciones. Habría que ver si tal variación es estadísticamente significativa, pero por sobre todo, buscar una posible causa sociológica o fisiológica relacionada a que sean las primeras horas de sueño nocturno de la mayor parte de la población.

No surge una acción a realizar en función de esta observación.

### 3.5. Ingeniería de características en función de la física del sistema

#### 3.5.1. Linealización de la magnitud

Todas las escalas de magnitud buscan dar cuenta de la energía liberada en el terremoto. La escala utilizada en los datos del INPRES es la de magnitud de momento (ver sección 3.1) definida como [1, ec. 4.23]

$$M_w = \frac{2}{3} \log_{10} (M_0) - 6,0 = \frac{2}{3} \log_{10} (\mu A u) - 6,0, \quad (3.1)$$

donde  $M_0$  es el llamado *momento sísmico* a su vez función del módulo de cizalladura,  $\mu$ , el área de falla involucrada,  $A$ , y su desplazamiento promedio,  $u$ , todas características del fenómeno en profundidad [1, sección 4.2.4].

Como se discutió en la sección 3.1.1  $M_w$  fue precedida por otras escalas de magnitud ligadas a la amplitud de las ondas sísmicas percibidas en la superficie, como la de onda de cuerpo,  $m_b$ , y la de onda superficial,  $M_s$ , derivadas de la original de Richter de 1935. Puesto que este trabajo busca predecir la percepción por personas en la superficie, interesa relacionar los datos de  $M_w$  con estas escalas que tienen una forma genérica

$$M = \log_{10} \left( \frac{A}{T} \right) + q(\Delta, h) + a, \quad (3.2)$$

donde  $M$  es la magnitud, y aquí  $A$  es la amplitud de las ondas sísmicas detectadas,  $T$  su período de oscilación,  $q$  es función de la profundidad del hipocentro,  $h$ ,  $\Delta$  el ángulo entre éste y el sismógrafo y la vertical, y  $a$  es una constante de ajuste [1, ecuación 4.13].

Dado que se carece del dato del punto de detección del sismo, no puede determinarse  $\Delta$  y aunque se eligiera la escala que corresponde por  $h$ <sup>1</sup> no podría determinarse  $q(\Delta, h)$ . Frente a esto las diferencias entre las relaciones empíricas entre  $M_w$  con  $M_s$  o  $m_b$ [26], son irrelevantes a los fines de este trabajo. Se asume entonces por válida una fuerte aproximación

$$M_w \approx M_s \approx m_b. \quad (3.3)$$

---

<sup>1</sup>Si  $h < 50$  km la mayor parte del aporte a la sismicidad la hacen ondas de propagación superficial por lo que se utiliza la *magnitud de onda superficial*,  $M_s$ , con distintos coeficientes de ajuste en la ecuación 3.2, que la *magnitud de onda de cuerpo*,  $m_b$ , que se usa para mayores profundidades [1, sección 4.2.3].

Para las ondas de tipo S y P, las que contribuyen a los sismos percibidos en superficie (ver sección 1.1) se puede hacer uso de la escala propuesta originalmente por Gutenberg en 1945 con la función de calibración para  $q(\Delta, h)$  propuesta por *Gutenberg-Richter* en 1956 [1, ecuación 4.18] en la que se omite de la expresión 3.2 la constante de ajuste  $a$  y para definir la razón  $\frac{A}{T}$  se toma la mayor registrada por los sismógrafos,

$$m_b = \log_{10} \left( \frac{A}{T} \right)_{\text{máx}} + q(\Delta, h). \quad (3.4)$$

Para  $q(\Delta, h)$  se usan valores tabulados, e.g. para ondas P en  $\Delta = 10$  a  $110^\circ$  corresponden  $q(\Delta, h) \approx 6$  a  $8$  [27].

Para obtener un valor lineal  $A$  a partir de la escala 3.4, que es la utilizada en los datos de magnitud del INPRES, se consideró usar un valor medio del rango comentado en el párrafo anterior,  $\bar{q}(\Delta, h)$ , con lo que podría despejarse

$$\left( \frac{A}{T} \right)_{\text{máx}} = 10^{(m_b - \bar{q}(\Delta, h))}. \quad (3.5)$$

Pero el menor  $m_b$  en el conjunto de datos es 2,5 por lo que si restase tal promedio,  $\bar{q}(\Delta, h) = 7$ , se obtendrían valores negativos para el valor  $A$  a la izquierda de la ecuación y esto sería algo sin validez física (¡amplitudes negativas!). Se decidió utilizar entonces un ficticio  $\tilde{q}(\Delta, h) = 2,5$ .

De querer despejar la amplitud de las ondas,  $A$ , debiera tenerse información sobre el período de las ondas,  $T$ , algo que no figura en el conjunto de datos. No se encontró otra alternativa que asumir que todos los fenómenos registrados tienen el mismo y en consecuencia asumir  $T_{\text{constante}}$ . Así realizando despejes a partir de la ecuación 3.4 y asumiendo tales condicionantes, se puede obtener un valor linealmente relacionado con amplitud registrada por un sismógrafo,  $A_{\text{máx}}$ ,

$$\left( \frac{A_{\text{máx}}}{T_{\text{constante}}} \right) = 10^{(m_b - 2,5)}. \quad (3.6)$$

Con el comando `sismos[, Proxy\_amplitud := 10^(Magnitud- 2.5)]` se generó una columna para este valor denominada `Proxy\_amplitud`.

### 3.5.2. Descarte de terremotos de poca profundidad

Hay sismos cuyo origen no son terremotos sino desplazamientos superficiales de tierra, explosiones para la minera o el fracturado hidráulico para la extracción de hidrocarburos. Se busca omitir tales orígenes en los datos informados.

Siendo que la variable se informa como enteros de kilómetros, estos representarían los hipocentros hasta una profundidad de 500 m, compatibles con estas actividades artificiales. La omisión de estas fuentes a baja profundidad es una práctica usual en el análisis de datos orientados a sismos originados en terremotos [28]. El pequeño número que estos representan el conjunto de datos se aprecia en el histograma según profundidad en la figura 3.8. Son filtrados con la instrucción `sismos <- sismos[Profundidad > 0]`.

Este recorte tiene un efecto cuasi-irrelevante en la distribución de la profundidad de los terremotos en el conjunto de datos para San Juan. La figura 3.8 mostró que los hay con  $h < 50$  km, pero los hay hasta con  $h = 750$  km siendo los de  $h > 50$  km  $\approx 79\%$  del total como evidencia el histograma que reproduce la figura 3.9.

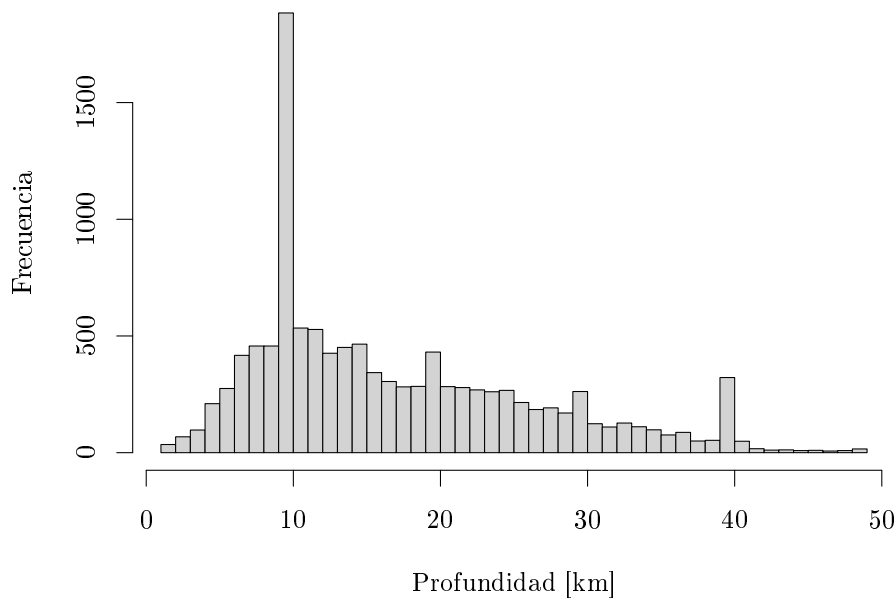


Figura 3.8: Los sismos con origen a profundidad de 0 km son pocos en el conjunto de datos de terremotos con hipocentros poco profundos,  $< 50$  km.

### 3.6. Variables a correlacionar con la percepción

De las originales referidas al fecha y hora en que se produce el terremoto Fecha y Hora se trabajará solo con la generada a partir de la segunda Hora\_decimal obviando la primera pues no es esperable que la percepción de un sismo por parte de la población.

De los datos sobre parámetros físicos del terremoto, Magnitud y Profundidad, se trabajará con la segunda sin modificación, en tanto que se espera un mejor desempeño en modelos que se basen en la variable elaborada a partir de la primera, Proxy\_amplitud por lo expuesto en la sección ???. De todas formas, como la relación entre ambas no es lineal, la covarianza entre ambas es elevada,  $\approx 0,233$ , pero no es perfecta por lo que podría ensayarse el impacto de usar alternativamente Magnitud o Proxy\_amplitud en distintos modelos.

Quedan así potenciales seis variables para trabajar en el modelo de clasificación del estado de Percibido: Hora\_decimal, Latitud, Longitud, Profundidad, Magnitud y Proxy\_amplitud.

Con la función `cor` se verificó que la covarianza de Percibido con Hora\_decimal es casi nula, que cuanto más profundo es el terremoto es menos percibido y que era de esperarse que terremotos que liberan más energía sean más percibidos, Magnitud y Proxy\_amplitud muestran una correlación positiva. La correlación de la percepción con la latitud resultó ser menos significativa que con la longitud. Se resume esto en el cuadro 3.1.

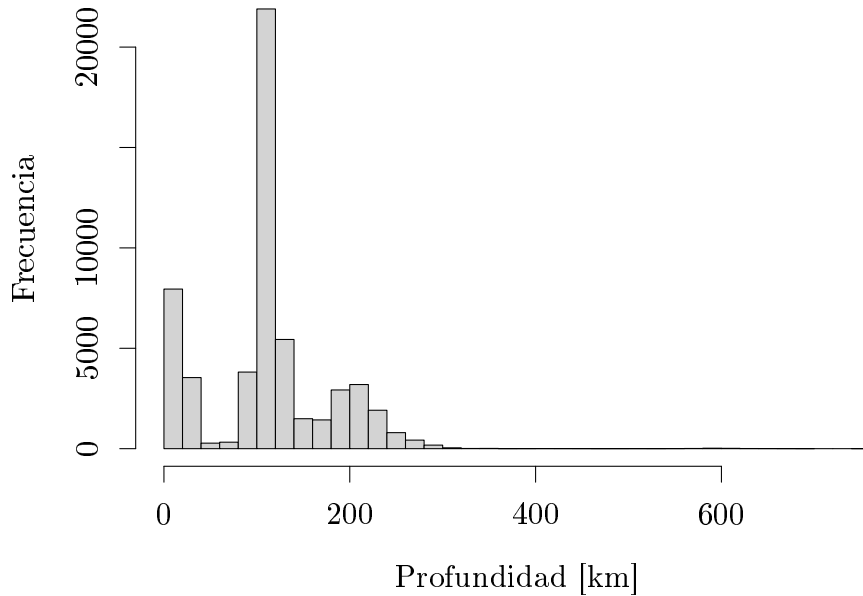


Figura 3.9: Los sismos con origen a mayor profundidad de 50 km son mayoritarios en el conjunto de datos.

| Hora_decimal | Profundidad  | Proxy_amplitud | Latitud      | Longitud    |
|--------------|--------------|----------------|--------------|-------------|
| -0.004116917 | -0.142358740 | 0.165371903    | -0.010351411 | 0.055488097 |

Cuadro 3.1: La fila para la variable de percepción de la matriz de covarianza entre las variables de los sismos a analizarán muestra una débil correlación positiva con la variable que depende de la amplitud de las ondas sísmicas y una negativa con la profundidad del terremoto en concordancia con las expectativas lógicas que pueden tenerse sobre el fenómeno. Un orden de magnitud menor hay una positiva con la longitud, lo que se esperaba por acercarse a la cordillera, y una menor en magnitud con la latitud. Relativamente la hora a la largo de todo el día no prácticamente correlación con la percepción de los sismos. Para analizar esto último hay que segmentar por bandas horarias.

### 3.7. Preprocesamiento

**Escalamiento** Previo a la partición (splitting) se realiza un escalado uniforme sobre todo el conjunto de datos (scaling) de las características numéricas. Con esto variables de entrada a los modelos tendrán una media de cero y una desviación estándar de uno, lo que tiene consecuencias para interpretabilidad del modelo de regresión:

- los coeficientes serán comparables entre sí, no dependiendo de la escala de

los datos,

- el intercepto del modelo será la predicción esperada para el caso en que los factores contemplados sean nulos, es decir, no tengan efecto,
- y finalmente si se hace uso de regularización se evita que los coeficientes de las variables de mayor escala tengan un peso desproporcionado en la función de pérdida [29, sección 3.4.1].

Para esto se hace uso de `scale` función del conjunto base de R para generar `sismos_SJ_escalado`

**Partición con estratificación** Las evaluaciones sobre la calidad de los modelos de clasificación generados se realizarán sobre un subconjunto de ensayo (test) del 20 % de los datos de la provincia de San Juan, el resto se utilizará para el entrenamiento (train).

Dado el fuerte desequilibrio de la clase Percibido comentado en la sección 3.4, ante una división de los datos en subconjuntos entrenamiento y prueba estocástica está el riesgo de que el subconjunto de prueba quede con muy pocos casos positivos y no sea representativo de la distribución de la clase en el conjunto de datos. Para evitar esto se realiza una división estratificada de los datos en subconjuntos de entrenamiento y prueba usando la función `CreateDataPartition` que indicó los índices para generar los conjuntos de datos `entrenamiento_SJ` y `ensayo_SJ` este último con un número aún adecuado para su función de 5982 registros.

**Desequilibrio en entrenamiento** Para contrarrestar el desequilibrio en la clase de clasificación se utilizará la técnica de sobremuestreo de la clase minoritaria que genera nuevos casos sintéticos de la clase minoritaria a partir de los existentes. Para esto se utiliza la función `ovun.sample` de la biblioteca `ROSE` que genera un conjunto de datos de entrenamiento con un número de casos de la clase minoritaria igual al de la clase mayoritaria. Se generó así un nuevo conjunto de datos de entrenamiento `entrenamiento_SJ_balanceado` con 23 934 registros.

### 3.8. Métricas de evaluación de los modelos

Independientemente del modelo de clasificación que se utilice, la evaluación de su desempeño se realiza a partir de la comparación de las predicciones del modelo con los valores reales de la variable objetivo. Se reservará un subconjunto de los datos para evaluar el desempeño del modelo, el conjunto de prueba y generar una matriz de confusión.

**Matriz de confusión** La matriz de confusión es una tabla que muestra el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos del modelo. A partir de esta matriz se pueden calcular como razones entre verdaderos y falso positivos y negativos la exactitud, precisión, sensibilidad, especificidad y el valor F1.

**Exactitud (accuracy)** La exactitud es la proporción de predicciones correctas sobre el total de casos. Se calcula como

$$\text{Exactitud} = \frac{\text{Verdaderos positivos} + \text{Verdaderos negativos}}{\text{Verdaderos pos} + \text{Falsos pos} + \text{Verdaderos neg} + \text{Falsos neg}}. \quad (3.7)$$

**Precisión** La precisión es la proporción de predicciones correctas sobre el total de predicciones realizadas. Se calcula como

$$\text{Precisión} = \frac{\text{Verdaderos positivos} + \text{Verdaderos negativos}}{\text{Verdaderos pos} + \text{Falsos pos} + \text{Verdaderos neg} + \text{Falsos neg}}. \quad (3.8)$$

**Sensibilidad (recall)** La sensibilidad es la proporción de verdaderos positivos sobre el total de casos positivos. Se calcula como

$$\text{Sensibilidad} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos negativos}}. \quad (3.9)$$

**Especificidad** La especificidad es la proporción de verdaderos negativos sobre el total de casos negativos. Se calcula como

$$\text{Especificidad} = \frac{\text{Verdaderos negativos}}{\text{Verdaderos negativos} + \text{Falsos positivos}}. \quad (3.10)$$

**F1-score** El *F1-score* es la media armónica de la precisión y la sensibilidad (recall). Se calcula como

$$\text{F1-score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}. \quad (3.11)$$

Con un gráfico de estas métricas en función del punto de corte entre clases se elige manualmente este que permita clasificar las instancias en una de las dos clases.

Como complemento se pueden trazar las curvas ROC y PR para evaluar el desempeño de los modelos en separar las clases.

**Área bajo la curva ROC** El área bajo la curva ROC (AUC-ROC) es una métrica que evalúa la capacidad de un modelo de clasificación para discriminar entre clases. Se calcula como el área bajo la curva ROC, que es la curva que representa la tasa de verdaderos positivos en función de la tasa de falsos positivos. El valor de AUC-ROC varía entre 0 y 1, donde 0 indica un modelo que clasifica todas las instancias de la clase positiva como negativas y viceversa, y 1 indica un modelo que clasifica perfectamente las instancias de ambas clases.

**Área bajo la curva PR** El área bajo la curva PR (AUC-PR) es una métrica que evalúa la capacidad de un modelo de clasificación para discriminar entre clases. Se calcula como el área bajo la curva PR, que es la curva que representa la precisión en función del *recall*. El valor de AUC-PR varía entre 0 y 1, donde 0 indica un modelo que clasifica todas las instancias de la clase positiva como negativas y viceversa, y 1 indica un modelo que clasifica perfectamente las instancias de ambas clases.

## Capítulo 4

# Resultados y discusión

### 4.1. Presentación de resultados

#### 4.1.1. Predictor por regresión logística

##### Múltiple sin interacción entre variables

El primer modelo ensayado es uno múltiple de regresión logística con las variables `Hora_decimal`, `Profundidad` y `Proxy\_amplitud` como predictores sin interacción entre ellos,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Hora\_decimal} + \beta_2 \text{Profundidad} + \beta_3 \text{Proxy\_amplitud}, \quad (4.1)$$

La función `summary` produce un resumen de los resultados arroja para los coeficientes:

Coefficients:

|                | Estimate  | Std. Error | z value | Pr(> z )   |
|----------------|-----------|------------|---------|------------|
| (Intercept)    | -0.311098 | 0.014614   | -21.287 | <2e-16 *** |
| Hora_decimal   | -0.015398 | 0.013362   | -1.152  | 0.249      |
| Profundidad    | -0.555176 | 0.010705   | -51.863 | <2e-16 *** |
| Proxy_amplitud | 0.167536  | 0.006263   | 26.750  | <2e-16 *** |

Esto indica que excepto el coeficiente  $\beta_1$  para la `Hora_decimal`, todos son significativos ya que sus probabilidades de que no tengan esos valores y se cumpla la hipótesis nula  $Pr(> |z|)$  para el estadístico  $z = \frac{\beta}{\sigma_\beta}$  se indican como muy inferiores a 0,05.

Para definir el punto de corte entre las clases se graficaron las métricas de evaluación en función de este como ilustra la figura 4.1. Si en base a la inspección de esa figura eligiera el punto de corte en 0,5 se obtendrían las métricas de evaluación que se resumen en la tabla 4.1.

Continuando con el análisis de la regresión logística ahora vería si logro mejorar estas métricas con un modelo que incluya interacción entre las variables.



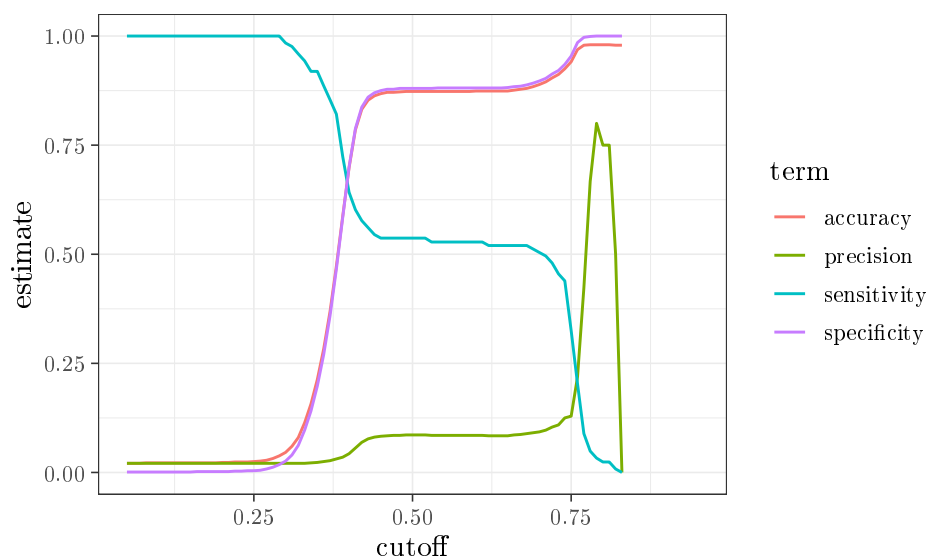


Figura 4.1: Métricas de evaluación para definir manualmente el punto de corte entre clases para el modelo de regresión logística múltiple sin interacción entre variables.

| corte | exactitud   | sensibilidad | especificidad | precisión    |
|-------|-------------|--------------|---------------|--------------|
| 0,5   | 0,872 617 9 | 0,536 585 4  | 0,879 672 3   | 0,085 603 11 |

Cuadro 4.1: Métricas de evaluación para el modelo de regresión logística múltiple sin interacción entre variables con punto de corte en 0,5.

## Múltiple con interacción entre variables

Esto es un placeholder para una sección aún vacía.

### 4.1.2. Predictor por XGBoost

Esto es un placeholder para una sección aún vacía.

## 4.2. Relevancia de los resultados

Esto es un placeholder para una sección aún vacía.

## 4.3. Limitaciones y posibles mejoras

Esto es un placeholder para una sección aún vacía.

## Capítulo 5

# Conclusión

### 5.1. Resumen de los hallazgos principales

Esto es un placeholder para una sección aún vacía.

### 5.2. Conclusiones generales y su relación con los objetivos del trabajo

Esto es un placeholder para una sección aún vacía.

### 5.3. Aplicaciones y relevancia de los resultados

Esto es un placeholder para una sección aún vacía.

# Bibliografía

- [1] C. M. R. Fowler. *The Solid Earth: An Introduction to Global Geophysics*. first. Cambridge University Press, 29 de jun. de 1990. 490 págs. ISBN: 978-0-521-37025-7. DOI: 10.1017/CB09780511819643. URL: <https://archive.org/details/solidearthintrod0000fowl> (visitado 16-06-2024).
- [2] Lisa Wald. *The Science of Earthquakes*. United States Geological Survey. URL: <https://www.usgs.gov/programs/earthquake-hazards/science-earthquakes> (visitado 25-06-2024).
- [3] ¿Qué es un sismo? Sistema Nacional para la Gestión Integral del Riesgo. 12 de nov. de 2018. URL: <https://www.argentina.gob.ar/sinagir/riesgos-frecuentes/sismos> (visitado 16-06-2024).
- [4] Saunders, J. K., Minson, S. E., Cochran, E. S., Bunn, J., Baltay, A. S., Kilb, D. y O'Rourke, C. «A Twist of PLUM: Low-Magnitude Earthquakes and Ground-Motion-Based Early Warning». En: 2021 Southern California Earthquake Center Annual Meeting, SCEC Contribution #11360. URL: <https://www.scec.org/publication/11360> (visitado 17-06-2024).
- [5] Sandra Vaiciulyte, David A. Novelo-Casanova, Allen L. Husker y Ana B. Garduño-González. «Population response to earthquakes and earthquake early warnings in Mexico». En: *International Journal of Disaster Risk Reduction* 72 (1 de abr. de 2022), pág. 102854. ISSN: 2212-4209. DOI: 10.1016/j.ijdr.2022.102854. URL: <https://www.sciencedirect.com/science/article/pii/S2212420922000735> (visitado 17-06-2024).
- [6] *Instituto Nacional de Prevención Sísmica*. Argentina.gob.ar. 28 de oct. de 2022. URL: <https://www.argentina.gob.ar/inpres> (visitado 17-06-2024).
- [7] *Encuesta de sismos*. Instituto Nacional de Prevención Sísmica. URL: [http://contenidos.inpres.gob.ar/encuesta/encuesta\\_sismo.php](http://contenidos.inpres.gob.ar/encuesta/encuesta_sismo.php) (visitado 30-06-2024).
- [8] *Felt Report - Tell Us!* United States Geological Survey. URL: <https://earthquake.usgs.gov/earthquakes/eventpage/tellus> (visitado 30-06-2024).
- [9] David Jay, Vincent Quitoriano, Charles Bruce, Margaret Hopper y James W. «USGS “Did You Feel It?” Internet-based macroseismic intensity maps». En: *Annals of Geophysics* 54.6 (14 de ene. de 2012). ISSN: 2037416X. DOI: 10.4401/ag-5354. URL: <http://www.annalsofgeophysics.eu/index.php/annals/article/view/5354> (visitado 27-06-2024).
- [10] *DYFI Scientific Background*. United States Geological Survey. URL: <https://earthquake.usgs.gov/data/dyfi/background.php>.

- [11] *Intensidad y Magnitud*. Instituto Nacional de Prevención Sísmica. 7 de nov. de 2022. URL: <https://www.argentina.gob.ar/inpres/docentes-y-alumnos/intensidad-y-magnitud> (visitado 01-07-2024).
- [12] G. M. Atkinson y D. J. Wald. «"Did You Feel It?" Intensity Data: A Surprisingly Good Measure of Earthquake Ground Motion». En: *Seismological Research Letters* 78.3 (1 de mayo de 2007), págs. 362-368. ISSN: 0895-0695. DOI: 10.1785/gssrl.78.3.362. URL: <https://pubs.geoscienceworld.org/srl/article/78/3/362-368/143359> (visitado 30-06-2024).
- [13] G. M. Atkinson, C. B. Worden y D. J. Wald. «Intensity Prediction Equations for North America». En: *Bulletin of the Seismological Society of America* 104.6 (1 de dic. de 2014), págs. 3084-3093. ISSN: 0037-1106. DOI: 10.1785/0120140178. URL: <https://pubs.geoscienceworld.org/bssa/article/104/6/3084-3093/332154> (visitado 30-06-2024).
- [14] Jerome Friedman, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay, Noah Simon, Junyang Qian y James Yang. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. Ver. 4.1-8. 22 de ago. de 2023. URL: <https://cran.r-project.org/web/packages/glmnet/index.html> (visitado 22-06-2024).
- [15] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, Jiaming Yuan y XGBoost contributors (base XGBoost implementation). *xgboost: Extreme Gradient Boosting*. Ver. 1.7.7.1. 25 de ene. de 2024. URL: <https://cran.r-project.org/web/packages/xgboost/index.html> (visitado 22-06-2024).
- [16] *IC-datasets-docencia*. URL: <https://daniellaparada.github.io/IC-datasets-docencia/> (visitado 12-06-2024).
- [17] Daniela Parada. *IC-datasets-docencia - 4 Visualización*. URL: [https://daniellaparada.github.io/IC-datasets-docencia/04\\_visualizacion.html](https://daniellaparada.github.io/IC-datasets-docencia/04_visualizacion.html) (visitado 12-06-2024).
- [18] *Buscador de sismos*. Instituto Nacional de Prevención Sísmica. URL: [http://contenidos.inpres.gob.ar/buscar\\_sismo](http://contenidos.inpres.gob.ar/buscar_sismo) (visitado 15-06-2024).
- [19] Daniela Parada. *sismos-arg*. URL: [https://github.com/daniellaparada/IC-datasets-docencia/blob/main/fuente/04\\_visualizacion/sismos-arg.csv](https://github.com/daniellaparada/IC-datasets-docencia/blob/main/fuente/04_visualizacion/sismos-arg.csv) (visitado 30-06-2024).
- [20] *ISO 8601-1:2019(en), Date and time — Representations for information interchange — Part 1: Basic rules*. International Organization for Standardization. 2019. URL: <https://www.iso.org/obp/ui/en/#iso:std:iso:8601:-1:ed-1:v1:en> (visitado 16-06-2024).
- [21] *Cálculo de la Magnitud*. Instituto Nacional de Prevención Sísmica. 7 de nov. de 2022. URL: <https://www.argentina.gob.ar/inpres/docentes-y-alumnos/calculo-de-la-magnitud> (visitado 01-07-2024).
- [22] *Sismos en el territorio argentino*. 5 de mayo de 2015. URL: <https://www.argentina.gob.ar/sites/default/files/sismos.pdf> (visitado 01-07-2024).

- [23] *Moment magnitude, Richter scale - what are the different magnitude scales, and why are there so many?* United States Geological Survey. URL: <https://www.usgs.gov/faqs/moment-magnitude-richter-scale-what-are-different-magnitude-scales-and-why-are-there-so-many> (visitado 30-06-2024).
- [24] *Introduction to data.table*. The Comprehensive R Archive Network. 27 de mar. de 2024. URL: <https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html#1.%20Basics> (visitado 18-06-2024).
- [25] *Acerca de tu ubicación*. Instituto Nacional de Prevención Sísmica. URL: <https://www.inpres.gob.ar/desktop/conoce.html> (visitado 17-06-2024).
- [26] Thomas C. Hanks e Hiroo Kanamori. «A moment magnitude scale». En: *Journal of Geophysical Research: Solid Earth* 84 (B5 1979). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/JB084iB05p02348>, págs. 2348-2350. ISSN: 2156-2202. DOI: 10.1029/JB084iB05p02348. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1029/JB084iB05p02348> (visitado 04-07-2024).
- [27] Willian L. Ellsworth. «Earthquake Magnitude: THE RICHTER SCALE (ML)». En: *The San Andreas Fault System, California*. Ed. por Robert E. Wallace. Vol. Professional Paper 15151. P. United States Geological Survey (USGS), 1991, pág. 177. URL: [https://web.archive.org/web/20160425121745/http://www.johnmartin.com/earthquakes/eqsafs/safs\\_693.htm](https://web.archive.org/web/20160425121745/http://www.johnmartin.com/earthquakes/eqsafs/safs_693.htm) (visitado 14-10-2008).
- [28] Yi Hu, Wentao Wang, Lei Li y Fangjun Wang. «Applying Machine Learning to Earthquake Engineering: A Scientometric Analysis of World Research». En: *Buildings* 14.5 (mayo de 2024). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, pág. 1393. ISSN: 2075-5309. DOI: 10.3390/buildings14051393. URL: <https://www.mdpi.com/2075-5309/14/5/1393> (visitado 17-06-2024).
- [29] Trevor Hastie, Harry Friedman y Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. second. Springer Series in Statistics 0172-7397. New York, NY, USA: Springer, 26 de ago. de 2009. 745 págs. ISBN: 978-0-387-84858-7. URL: <https://doi.org/10.1007/978-0-387-84858-7>.

# Anexos (opcionales)

## 5.4. Código fuente utilizado en el análisis

Enlace al repositorio en GitHub que aloja el código fuente utilizado en el análisis de los datos.

## 5.5. Tablas y gráficos adicionales

Esto es un placeholder para una sección aún vacía.

## 5.6. Otros materiales relevantes

Esto es un placeholder para una sección aún vacía.