0.25
cx

Universidad de
SanAndrés

Big Data

Propuesta de investigación

Pilar Ruiz Orrico
Sebastián Einstoss Mastracchio
Sofía Marinkovic Dal Poggetto

Noviembre 2021

1. Intoducción

La última dictadura militar (1976-1983) puede categorizarse como uno de los eventos más violentos sufridos por la población argentina en el siglo XX. El ya conocido modus-operandi tuvo como una de sus principales características el asesinato y desaparición/detención de una fracción de la población.

Luego del proceso de desaparición/detención, las diferentes fuerzas represivas se distribuían las víctimas y tomaban la decisión sobre el destino de las mismas. En el caso de decidir el asesinato de una persona, se implementaron diversas estrategias. Un ejemplo de esto son los llamados "vuelos de la muerte" o los entierros como "NN" (*Nomen nescio*).

Más de 40 años después, aún continúan exhumándose restos de víctimas del terorrismo de Estado en pos de reconstruir la historia. El proceso en cuestión implica una inversión de recursos por parte de la sociedad tanto en términos económicos como en tiempo.

Si bien el proceso hasta llegar a la exhumación de un cuerpo suele ser complejo, una pieza fundamental es el proceso judicial y la generación de hipótesis sobre la posible localización de los restos de las víctimas. Dentro del proceso judicial se presentan pruebas que sustentan la hipótesis sobre la locación, junto con los argumentos para exhumar un determinado cuerpo. Esta hipótesis se sustenta en una serie de datos recolectados por diferentes organismos con el objetivo de intentar reconstruir el paradero de la persona. La información es tanto cualitativa como cuantitativa y se procesa de modo de intentar reconstruir la historia detrás de la desaparación.

El presente trabajo tiene como objetivo la generación de un modelo predictivo de la posible locación, en cuanto al cementerio de destino, de aquellas personas cuyos cuerpos aún no se ha logrado identificar. De este modo, el modelo tendría input las diferentes características de las personas y en base a esta información daría una predicción del tipo categórica.

En el presente esta tarea es implementada por diferentes entes públicos sin la validación de un modelo estadístico. El presente trabajo potencialmente podría ser de gran utilidad en el proceso de generación de hipótesis por parte de los organismos que intentan recontruir la historia del país. La bibiografía relacionada a esta temática se ha abordado principalmente con modelos para intentar predecir la locación en términos geográficos de las fosas comúnes donde podrían encontrarse víctimas del terrorismo de Estado. Un ejemplo de esto, aplicado para la Guerra Civil española, es el caso de Congram (2010) [1]. En nuestro caso en particular, queremos predecir en qué cementerio (ya existente) es más probable que se encuentre la persona en cuestión y no en qué longitud y latitud deben buscarse posibles fosas.

2. Base de datos

Los diferentes organismos encargados de la reconstrucción de los casos y la conservación de la memoria han recolectado una extensa cantidad de información acerca de las víctimas del terrorismo de Estado. En particular, la información relevante para el proyecto se encuentra en el "Registro Unificado de Víctimas del Terrorismo de Estado" (RUVTE) y más especificamente en el Archivo Nacional de la Memoria.

El RUVTE presenta una ventaja importante en cuanto a la sistematización de la información. Sin embargo, sólo se encuentra disponible públicamente de forma parcial [2]. Para acceder a la base ampliada existen dos posibles alternativas: Una consiste en generar un pedido de información pública (la que podría darnos el derecho a acceder a bases de datos generadas por el sector público). Mediante este proceso, podríamos llegar a obtener la base de datos de forma completa. Otra vía posible es realizar un "Web-Scraping" de la página web del Monumento a las Víctimas del Terrorismo de Estado. Este ente en su sitio web replica la inforación proveniente del RUVTE (Cabe aclarar que mediante una consulta del tipo informal la Secretaría de Derechos Humanos nos confirmaron que el monumento replica online el RUTVE con cierto rezago) [4].

La información disponible para las victimas van desde datos básicos tales como edad, sexo, fecha de nacimiento, nacionalidad, residencia y ocupaciones, hasta información mucho más detallada como posibles víctimas relacionadas y afiliación a algún partido pólitico. Si bien desconocemos ex-ante cuál es la información completa de cada persona presente en el RUVTE, consideramos que los datos públicos ya arrojan un buen punto de partida. En concreto, la base de datos (con denuncia formal) cuenta con un total de 8.857 observaciones.

Por último, esta base de datos contiene información sobre los casos identificados por el Equipo Argentino de Antropología Forense (EAAF). Esta información sobre la locación donde las victimas fueron identificada será lo que nos permitirá generar una muestra de entrenamiento. Un ejemplo de una victima cuyos restos fueron identificados puede encontrarse en la página web del Monumento a las Víctimas del Terrorismo de Estado - Parque de la Memoria [3].

Se puede observar, para este caso, la existencia de variables como estado, fecha de secuestro, lugar de secuestro, fecha de asesinato y datos de la identificación. Luego, en la parte de datos personales, se encuentran datos tales como nombre, edad, sexo, fecha de nacimiento, lugar de nacimiento, nacionalidad, estado civil, domicilios, ocupaciones y víctimas simultáneas. Cabe agregar que las variables disponibles dependen del caso (por ejemplo, otras victimas tienen información sobre su afiliación politica).

Esta información podría complementarse con aquella presente en el Archivo Nacional de la Memoria, la que ha sido usada como insumo para las exhumaciones ya realizadas. Esto resulta un punto fundamental ya que permitiría ampliar la información con detalles valiosos tales como el centro de detención clandestino al que la persona en cuestión fue asignada. En este punto, intuimos que los entes oficiales podrían estar interesados en el desarollo de este proyecto y tal vez podrían brindarnos ayuda en su implementación.

3. Metodología

La naturaleza propia de nuestro problema limita las herramientas a utilizar. Lo que se desea hacer es una clasificación de los datos en una cantidad determinadas de categorías (cementerios).

Un primer punto a aclarar, es que deben encontrarse aquellos casos que efectivamente fueron identificados, ya que serán la base para la elección de los hiperparametros y del método.

Una posible alternativa inicial podría provenir de implementar un método CART en su forma clásica. Éste parte el espacio en diferentes regiones y dentro de cada partición se propone a la media muestral como predicción. Por lo tanto, se debe elegir que variable se usa para patir y en que punto se parte la muestra óptima (mejor ajuste global). Este proceso se repite para todas las variables y "todas" las particiones. Un punto importante es encontrar la poda óptima que minimice la "cost-complexity" (donde se penaliza tanto el mal ajuste del modelo como la complejidad).

CART tendría como ventaja la facíl interpretabilidad del mismo, lo que en estas instancias podría resultar sumamente valioso para comprender la estructura subyacente, pero a su vez presenta una serie de problemas. Como en este caso en particular se iría sumando información acerca de los nuevos casos identificados, la estabilidad de las predicciones resulta algo sumamente importante. Es decir, un modelo que cambie sustancialmente su predicción a medida que Antropoligía forense realiza avances significaría una gran pérdida en términos de credibildiad del modelo. Al mismo tiempo, ex-ante desconocemos si existe una estructura lineal o no en los datos (aunque presuponemos que es plausible que no).

Una alternativa para reducir la varianza de las predicciones es un posible método Bagging. Este generaría una predicción a partir del promedio de las predicciones de diversos modelos bootstrap. Una clara preocupación aquí proviene en que es probable, en este caso, que algunas de las variables sea un predictor muy fuerte y genere que todos los arboles sean muy similares (y, por lo tanto, las predicciones que se promedian sean muy similares entre sí). Una variable que intuimos que puede tener gran capacidad predictiva es el centro clandestino en el que fue visto la persona por última vez.

Finalmente, dos posibles alternativas podrán ser Random Forest o Boosting. Que ambas presentan grandes ventajas con respecto a CART original y reducen los problemas de Bagging.

Random Forest reduce la correlación de los arboles a tráves de bootstrap usando unicamente m¡p predictores elegidos al azar (siendo p la cantidad total de predictores). Tipicamente se utiliza $m = \sqrt{p}$. Boosting genera el promedio ponderado de una sucesión de clasificadores débiles (donde si usaramos arboles muy frondosos podríamos tener un problema de overfit).

La elección del mejor modelo se realizaría mediante cross-validation. A partir de esto, eligiríamos tanto el modelo como los hiperparámetros que generen el menor error cuadratico medio.

Por último, hacen falta una aclaración con respecto al manejo de la información faltante. Luego de convertir las variables categóricas en dummies, en aquellos casos donde no se presente información esta será reemplazada por 0 (es decir, no vamos a tomar ni la media ni la varianza). Esto se debe a que tal vez le estaríamos imputando una caracteristica que puede ser muy sensible a una persona (como por ejemplo una afiliación a un determinado partido pólitico).

4. Conclusiones y limitaciones

Mediante nuestro proyecto de investigación esperamos obtener como resultado una estructura de predicciones que permita hechar luz sobre la posible locación de las víctimas del terrorismo de Estado de la última dictadura militar.

Esperamos que el modelo posea buena capacidad predictiva, siempre y cuando sea capaz de incorporar información sobre una serie de variables que consideramos como centrales para el problema que enfrentamos. Como ya mencionamos, los datos necesarios son obtenibles mediante pedidos de información pública a la Secretaría de Derechos Humanos. Las variables que consideramos centrales para el armado de nuestro modelo predictivo son: centro clandestino asignado, ámbito de militancia, lugar de residencia, fuerza armada encargada de la detención y año de desaparición. En el presente, casi la totalidad de las variables mencionadas se encuentran disponibles públicamente en la página web del "Parque de la Memoria", por lo que su rápida implementación parece plausible.

No obstante, cabe mencionar las principales limitaciones de nuestra propuesta. En este sentido, pensamos que estas provendrán del hecho de que el modelo utiliza como insumo a los casos previamente dcoumentados. Es decir, no podremos identificar, por ejemplo, aquellos casos de las víctimas de los llamados "vuelos de la muerte", porque sus cuerpos aún no fueron identificados. Entonces, habrá información

faltante para entrenar nuestro modelo en un intento de querer predecir la totalidad de las desapariciones y/o asesinatos.

Por último, nos gustaría indicar que consideramos que la implementación de este trabajo podría traer grandes beneficios a las familias de las personas víctimas del terrorismo de Estado, así como a la sociedad en su conjunto (en su búsqueda por la verdad, la memoria y la justicia). De esta manera, este proyecto podría tener un gran impacto social para la población argentina y potencialmente contribuya a marcar un precedente en el camino de la búsqueda de datos de víctimas de regímenes similares al vivido en Argentina en otros países.

5. Referencias Bibliográficas

[1] Derek Reade Congram. "Spatial analysis and predictive modelling of clandestine graves from rear-guard repression of the Spanish Civil War". Tesis doctoral. Arts & Social Sciences: Department of Archaeology, 2010.

[2] Ministerio de Justicia y Derechos Humanos. "Registro Unificado de Víctimas del Terrorismo de Estado -RUVTE". En: *Secretaría de Derechos Humanos y Pluralismo Cultural. Registro Unificado de Víctimas del Terrorismo de Estado* (2021). DOI: http://datos.jus.gob.ar/dataset/registro-unificado-de-victimas-del-terrorismo-de-estado-ruvte.

[3] Monumento a las Víctimas del Terrorismo de Estado Parque de la Memoria. *Registro de Víctimas*. URL: http://basededatos.parquedelamemoria.org.ar/registros/2140/.

[4] Monumento a las Víctimas del Terrorismo de Estado Parque de la Memoria. *Víctimas*. URL: http://basededatos.parquedelamemoria.org.ar/registros/.