Examen Parcial

6. Planear un problema una investigación de análisis multivariado con las técnicas que hemos visto en clase. El objetivo es ver que tengan presentes las técnicas que hemos visto al hacer un análisis real. No necesariamente tienen que programar el uso de la técnica, solo plantear cómo la usarían y que resultados y posibles conclusiones esperarían. La pregunta tiene valor de comprensión. Si hay alguna técnica que hubiera sido lógico usar y no la mencionan afecta a la evaluación de la pregunta. Este proyecto debería servir de base para su proyecto final. Básense en los siguientes lineamientos:

- Escojan un dataset que les interese a ustedes y al que le quieran dar seguimiento con las restantes técnicas que veremos en el curso. Si no saben qué base de datos usar, utilicen adults de UCI.
- Describan el dataset, las variables y su tipo de dato estadístico (no de R).
- ¿Cuáles son los retos desde el punto de vista del análisis multivariado?
- ¿Qué hipótesis quisieran contestarse con sus datos? ¿Qué técnicas del curso podrían usarse para cada problema de investigación?
- Para cada técnica que mencionen deben contestar por qué la técnica es apropiada para resolver su hipótesis y plantear sus posibles resultados y conclusiones.
- ¿Qué técnicas que no hemos estudiado en nuestro temario podrían combinarse con técnicas del curso para solucionar las hipótesis que les interesan?

A raíz de los conflictos sociales al rededor del mundo se ha observado un incremento en la violencia observada en los últimos años. En particular, enfocaremos el análisis en la amenaza de libertad de expresión y libertad de prensa estudiando un conjunto de datos presentado por el Comité para la Protección de Periodistas (*Committee to Protect Journalist CPJ*).

El conjunto de datos antes mencionado es una recolección de asesinatos de periodistas desde 1992 hasta marzo de 2017 en aproximadamente 105 países. Cada evento o asesinato tiene información sobre el trabajo del periodista y sobre el asesinato.

Es importante mencionar que CPJ incluye en su definición de periodistas al staff, freelancers o independientes, bloggers, stringers y ciudadanos periodistas que reportan noticias de dominio público sin importar el medio ya sea impreso, fotográfico, de radio, televisión o electrónico.

A continuación se presenta la lista de variables incluidas en los datos y una breve descripción:

- **Motive:** variable categórica sobre el motivo del asesinato. Tiene tres niveles: *motive confirmed, motive unconfirmed, media workers*. La variable *media workers* registra el asesinato de gente de apoyo clave en el desarrollo de la investigación.
- **Date:** variable numérica de fecha del asesinato. Incluye mes, año y en algunos casos día. Se decide únicamente considerar mes y año.
- **Name:** variable categórica única por periodista excepto cuando no se conoce la identidad de la persona.
- **Sex:** variable binaria del sexo del periodista. Tiene dos niveles: *female, male*.
- **Country Killed:** variable categórica del país donde ocurrió el asesinato. La variable tiene 105 niveles. Los 30 con mayor número de asesinatos en orden decreciente son: *Iraq, Philippines, Syria, Mexico, Pakistan, Colombia, Russia, India, Somalia, Algeria, Brazil, Afghanistan, Bangladesh, Tajikistan, Turkey, Bosnia, Sri Lanka, Honduras, Guatemala, Rwanda, Israel, Yemen, Nigeria, Ukraine, Nepal, Sierra Leone, Peru, Thailand, Democratic Republic Of The Congo, Egypt*
- **Organization:** variable categórica. Organización para la que el periodista laboraba. La variable tiene 1234 niveles.
- **Nationality:** variable categórica. Nacionalidad de la persona. Se tienen 170 niveles. Los treinta niveles con mayor frecuencia en orden descendente son: *Iraq, Syrian, Philippine, Pakistan, Mexican,*

Algeria, Russia, Colombia, Iraqi, Brazil, India, Somalia, Turkey, Bangladesh, France, Sri Lanka, United States, Honduras, Rwanda, Tajikistan, Nigeria, Afghanistan, United Kingdom, Sierra Leone, Germany, Italy, Peru, Yemen, Egypt, Nepal. Esta variable requiere mayor limpieza, por lo que los niveles pueden disminuir y el ordenamiento cambiar.

- **Medium:** variable categórica. Medios para los que trabajaba la persona. La variable tiene 15 niveles que son las combinaciones de los medios: *radio, television, print, internet.*
- **Job:** variable categórica. Combinación de empleos que desempeñaba de las diez posibilidades siguientes: *internet reporter, print reporter, broadcast, camera operator, editor, photographer, producer, publisher, technician, columnist.*
- **Coverage:** variable categórica. Combinación de coberturas de noticias del periodista de los siguientes temas: *crime, politics, war, corruption, culture, human rights, business, sports.*
- **Freelance:** variable binaria. El periodista laboraba para prensa independiente. Dos niveles: *yes, no.*
- **Local/Foreign:** variable binaria. Se refiere a la procedencia del periodista. Dos niveles: *local, foreign.*
- **Source of Fire:** variable categórica. Personas o entidades probablemente responsables del asesinato. Es una combinación de las siguientes opciones: *criminal group, government officials, local residents, military officials, mob violence, paramilitary group, political group, unknown fire*
- **Type of Death:** variable categórica. Clasificación de asesinato de CPJ únicamente para los casos en los que el motivo esté confirmado. Las opciones son *crossfire/combat-related, dangerous assignment, murder, unknown*
- **Impunity (for Murder):** variable categórica. Monitoreo realizado por CPJ sobre el cumplimiento de la ley y el proceso legal para los casos de asesinato confirmado. Tiene tres niveles: *yes, no, partial.* El nivel *partial* se refiere a cuando algunos responsables son condenados pero no la totalidad.
- **Taken Captive:** variable binaria. Se refiere al secuestro del periodista en un periodo inmediato previo a la muerte. Dos niveles: *yes, no.*
- **Threatened:** variable binaria. Se refiere a cualquier amenaza dirigida al periodista cualquier momento antes de ser asesinado. Dos niveles: *yes, no.*
- **Tortured:** variable binaria. Esto significa que el periodista fue demostrablemente físicamente torturado antes de ser asesinado. Dos niveles: *yes, no.*

El reto del análisis multivariado para este conjunto de datos es el gran número de variables categóricas y valores faltantes de las variables. Para disminuir el problema de valores faltantes se decide condicionar el análisis a los asesinatos con motivo confirmado (*motive confirmed*). Finalmente, el conjunto de datos considerando el filtro antes mencionado se compone de un total de 1235 asesinatos al rededor del mundo.

Respecto a líneas de investigación, proponemos las siguientes hipótesis:

a. México ha sido comparado frecuentemente con Siria y otros países en relación a la amenaza contra la libertad de prensa. ¿Es esto real? ¿En qué temas es esto cierto? Existen otros países con los que México se puede relacionar bajo características y tópicos específicos, por ejemplo impunidad o violencia.

Técnicas posibles:

Multiple Correspondence Analysis: esta técnica es apropiada para variables categóricas, que se tienen en el conjunto de datos. Además, estudia las asociaciones entre variables. Ésto nos ayuda a observar la asociación entre países dependiendo de otras variables.

Multidimensional Scaling: esta técnica ayuda a visualizar la similitud entre individuos considerando distancias. Para encontrar la similitud entre los países nos ayuda si recodificamos algunas variables por valores numéricos. En el caso de variables binarias convertir a 0 y 1. O bien, darles un valor numérico a los diferentes niveles asignando un jerarquía específica. Por ejemplo, en el caso de impunidad, se considera de mayor gravedad la impunidad total (*yes*), seguido por impunidad parcial (*partial*) y por último (*no*).

Una vez implementadas las técnicas esperamos obtener gráficamente asociaciones de los países respecto a las variables restantes. Esperamos obtener países con comportamientos similares que permitan aproximarnos a perfiles generales de países.

b. Uno de los principales intereses de CPJ es la creación de un índice de impunidad. Este índice es importante para presionar a los países a perseguir a los responsables en su totalidad, sobre todo en países que se consideran en pro de la democracia. Sin embargo, creemos que adicional a impunidad existen otros temas que caracterizan a los países respecto a la amenaza a la libertad de prensa y se deben de crear otros índices. Por ejemplo, relacionado al nivel de violencia, o amenaza a prensa independiente y cobertura con mayor riesgo, etc.

Técnicas posibles:

Canonical Correlation Analysis: ésta técnica analiza la relación de un conjunto de variables. Esto nos ayuda a encontrar la relación de los conjuntos de variables por los diferentes temas antes mencionados con los que se puede construir uno o varios índices.

Tetrachoric Correlation y Polychoric Correlation: ésta técnica permite estimar las correlaciones entre variables binarias en el caso (*tetrachoric*) y ordinales (*polychoric*). Esta última, puede implementarse asignando un valor numérico a diferentes niveles asignando un jerarquía específica, como se mencionó antes. De esta forma, aplicar factor análisis para resumir la información.

Factor Analysis y Mixed Factor Analysis: estas técnica determina variables que representan a los datos pero con menor número de variables. Esto se implementa en la construcción de índices. En particular si se incluyen únicamente las variables del tema de impunidad (*Impunity*, *Threatened*, *Source of Fire*, etc.) podremos resumir la estructura de estas variables en una dimensión. Mixed Factor Analysis permite realizar lo siguiente pero incluyendo variables categóricas.

Una vez implementadas las técnicas, esperamos obtener índices de variables asociadas por temas generales. Por ejemplo, el índice de impunidad se conforma de las variables (*Impunity*, *Threatened*, *Source of Fire*) que resultaron estar altamente asociadas. De esta forma, existe un nivel de impunidad grave en el caso de existencia de amenazas previas a los periodistas en general, impunidad de los responsables, los cuales están relacionados al gobierno. De forma similar lo veremos en las demás variables.

c. En 2016 México fue considerado el 6º país con mayor impunidad y el 10º con mayor número de asesinatos. Esta información se puede resumir en un índice que permita crear un ranking de países considerando varios temas como impunidad, violencia, cobertura, prensa independiente, entre otros.

Técnicas posibles:

Structural Equations Models: esta técnica nos permite relacionar diferentes componentes o variables mediante variables latentes bajo una estructura predeterminada. Considerando temas específicos, podremos determinar si existe una relación para generar un índice global que permita rankear a los países.

Una vez realizadas las técnicas esperamos confirmar la asociación de los índices antes obtenidos y obtener una variable latente que resuma la amenaza a los periodistas por país. Esto permitirá realizar una evaluación de México en distintos temas para encontrar áreas de mayor necesidad.

Referencias

Committe to Protect Journalists. https://cpj.org/.

Journalists Killed Worldwide Since 1992. https://www.kaggle.com/cpjournalists/journalists-killed-worldwide-since-1992/kernels