UNIVERSITAT OBERTA DE CATALUNYA

Máster Universitario en Ciencia de Datos (*Data Science*)

M2.859 – Visualización de Datos

Proyecto Final de Asignatura

Peace Agreements: Social Groups Perspectives

Documento conjunto de las pruebas PEC2, PEC3 y PEC4

Alberto García Galindo

Enero de 2021

Prueba de Evaluación Continua 2

Peace Agreements: Social groups perspectives

https://public.tableau.com/profile/albergcg#!/vizhome/PeaceAgreements_SocialGroups_v1/SocialGroups

I · Presentación y contexto

Esta segunda práctica se enmarca en la primera fase de desarrollo de un proyecto formal de visualización de datos. Para llevar a cabo dicho proyecto, se ha propuesto trabajar en el análisis de la ***PA-X Peace Agreements Database***[1], una base de datos creada por la *Law School* de la *University of Edinburgh* que contiene información sobre 1868 tratados de paz firmados en todo el mundo desde el año 1990 hasta junio del año 2020. Tal y como se menciona en su página web, se trata de una base de datos creada con el objetivo de proporcionar un punto de partida a usuarios e investigadores que busquen analizar y comprender diferentes patrones y tendencias sobre los acuerdos de paz documentados desde diferentes puntos de vista y atendiendo a objetivos específicos.

El objetivo de esta primera práctica será la familiarización con los datos y la comprensión tanto de su contenido como de su estructura. Analizaremos algunos aspectos de los tratados y propondremos algunas visualizaciones que nos permitan obtener una idea de sus características. Además, prestaremos especial atención a la caracterización de aquellos tratados en los que se haga algún tipo de referencia a grupos sociales, y llevaremos a cabo análisis exploratorios desde diferentes perspectivas.

II · Datos

Para cada uno de los tratados de paz documentados se recogieron, junto con la declaración textual, un total de 266 metadatos que recogían tanto información básica como específica del contenido del acuerdo. Con el objetivo de comprender mejor cada uno de estos metadatos, se hizo uso del material adicional accesible por parte de los autores en la web oficial de la base de datos. En concreto, el documento *codebook*[2] fue la principal fuente analizada, ya que contenía una definición concisa para cada campo asociado a cada metadato. De esta manera, pudimos clasificar la información referente a los tratados de paz en dos grandes subgrupos:

- **Tipología y contexto**. Este primer conjunto de metadatos, compuesto por los primeros 26 campos, especifica las características básicas del tratado en cuestión, pudiendo establecer una taxonomía. De esta manera, podemos clasificar al tratado en función de, por ejemplo, el país o países a los que se refiere, la región asociada, el proceso de paz en el que se contextualiza o la fase en la que se encuentra dicho proceso, así como de la fecha en el que se firmó o acordó. La mayoría de estos campos corresponden con variables categóricas en formato de cadena que utilizaremos para incluir etiquetas en nuestros gráficos, segmentar la información dispuesta en las visualizaciones o filtrar la información en base a un determinado criterio.
- **Referencias**. Por otra parte, el conjunto de metadatos restante describe el contenido del documento, especificando si hay referencias, y en algunos casos de qué tipo, a diferentes

[1] https://www.peaceagreements.org/

[2] https://www.peaceagreements.org/files/PA-X%20codebook%20Version4.pdf

temáticas relevantes enmarcadas en el contexto de la paz. Disponer de todos estos datos nos permitirá, por tanto, analizar los procesos de paz en función de si tratan o no, por ejemplo, aspectos relacionados con los derechos humanos, grupos sociales o el sector socioeconómico. Nos parece interesante recalcar que, en el caso concreto de las referencias a grupos sociales, aparece una variable categórica ordinal que especifica el grado de importancia de la referencia y, además, la propia variable bajo la codificación *one-hot* (en término estadísticos, variable *dummy*), con valores de ausencia/presencia. Por este motivo, y para simplificar, para las referencias en las que se indicaba el grado de importancia, utilizamos únicamente la variable que recoge toda la información en categorías ordinales.

Una vez revisados todos los metadatos, podemos destacar la ausencia prácticamente total de variables continuas. Esto condicionará en gran medida nuestro proyecto, e implicará que la gran parte de nuestros análisis y visualizaciones sean a través del estudio de frecuencias de los tratados de paz desde diferentes puntos de vista.

III · Evidencias, patrones y anomalías

Tras comprender los datos, se llevaron a cabo diferentes procedimientos para extraer la mayor cantidad posible de información sobre los tratados de paz documentados. Toda esta metodología se encuentra detallada en el siguiente apartado. A continuación, enumeramos las evidencias y patrones extraídos durante el procedimiento.

- Más de un 30% de los tratados de paz tuvieron lugar en la zona de África (excluyendo África del Norte), siendo esta por tanto la región en la que más acuerdos se firmaron seguida de la región de Europa y Asia (22%) y de la región de Asia y el Pacífico (20%).
- Gran parte de los acuerdos sucedieron en la década de los 90, siendo 1994 el año en que más acuerdos de paz se firmaron, debido en parte a que únicamente en los procesos de paz de Bosnia y Herzegovina en este año se firmaron hasta 24 tratados.
- Atendiendo al tipo de conflicto en el que se contextualizaban los acuerdos, la gran mayoría (80%) atendían a algún tipo de aspecto gubernamental. Además, analizando la distribución temporal para cada tipo de conflicto, pudimos comprobar que no era uniforme, sino que se podían vislumbrar algunas tendencias:
    - Para los conflictos enmarcados tanto en procesos exclusivamente gubernamentales y como territoriales, se observó una estabilidad en el tiempo en cuanto a número de conflictos.
    - No obstante, para aquellos procesos que se enmarcaban tanto en un contexto gubernamental como territorial, podemos observar una mayor concentración de acuerdos en la década de los 90, en la que se firmaron prácticamente la mitad.
    - Por otra parte, en el caso de los procesos entre grupos, encontramos una mayor concentración en la década de 2010, donde se firmaron mas del 80% de acuerdos de este tipo.

- Curiosamente, tanto el proceso de paz de Bosnia como el proceso de paz de Filipinas – Mindanao tuvieron exactamente el mismo número de acuerdos involucrados: 124. No obstante, a pesar de que ambos se sitúan a la cabeza en el ranking de procesos por número de acuerdos firmados, su duración fue muy diferente. En el caso del proceso de Bosnia, encontramos tratados únicamente entre los años 1992 y 1998, mientras que en el caso del proceso de Filipinas – Mindanao, el primer tratado documentado se sitúa en 1993 y el último en 2018.

- Podríamos destacar como anomalía los casos de España y Reino Unido, puesto que son los únicos países de la Europa Occidental que cuentan con algún conflicto de paz documentado. En el caso de España, los 3 acuerdos recogidos se enmarcan en los sucesos relacionados con el nacionalismo vasco y la organización terrorista Euskadi Ta Askatasuna (ETA).

- Hemos encontrado potencialmente interesante analizar las referencias sobre algunos grupos sociales y caracterizar aquellos procesos y tratados que hacen algún tipo de mención a sus aspectos. De esta manera, hemos podido comprobar que las referencias, por ejemplo, tanto a refugiados como a grupos raciales/étnicos no es uniforme a lo largo del tiempo ni tampoco entre las regiones del mundo. A pesar de haber descubierto algunos hechos interesantes, el proyecto de visualización estará enfocado en profundizar en los aspectos sociales de la base de datos.

IV · Metodología y herramientas

Para llevar a cabo la exploración de datos y la construcción de la visualización, hemos seguido los siguientes pasos.

En primer lugar, descargamos el archivo que contenía los 1868 registros de la base de datos, y utilizamos Microsoft Excel para echar un primer vistazo a la tabla de datos. Como hemos dicho, también hicimos uso del *codebook* proporcionado por los autores para comprender correctamente el significado de cada uno de los metadatos almacenados.

A continuación, utilizamos el lenguaje de programación R para realizar una primera exploración a nivel general tanto gráfica como numérica de los datos. Visualizamos la distribución de algunas de las variables más importantes a nuestro criterio y comprendimos mejor como se clasificaban los acuerdos en función, por ejemplo, de su tipología y su región. Aunque consideramos utilizar este software como principal herramienta de exploración a la hora de realizar gráficos más complejos que nos permitieran conocer más a fondo nuestros datos, finalmente nos decantamos por Tableau gracias a su facilidad de uso y experiencia de usuario. Utilizando todas las posibilidades que ofrece (filtrado, segmentación, selección…) pudimos extraer toda la información descrita en el apartado anterior.

En este sentido, y puesto que no contamos con un dominio específico de la herramienta, su correcto manejo y conseguir una soltura a la hora de realizar gráficos complejos se plantean como objetivos personales del proyecto de visualización. Además, buscaremos profundizar desde la perspectiva estadística en técnicas de análisis de datos multivariante específicas (por ejemplo, Análisis Factorial de Correspondencias) para trabajar sobre datos categóricos, añadiendo un extra de complejidad al proyecto, y profundizaremos en algunos paquetes de R que nos permitan llevar a cabo dichos análisis.

Prueba de Evaluación Continua 3

Máster Universitario en Ciencia de Datos (*Data Science*)

M2.859 – Visualización de datos

Peace Agreements: Social Groups Perspectives

Alberto García Galindo

Diciembre de 2020

Abstract

El presente informe introduce un proyecto de visualización de datos real a partir de un conjunto de datos de sobre tratados de paz a lo largo de todo el mundo en los últimos 30 años. El objetivo del proyecto es llevar a cabo la construcción un cuadro de mandos en Tableau que permita la exploración interactiva de las diferentes características del conjunto de datos mencionado. Para ello, se llevará a cabo un ejercicio de preprocesamiento y análisis inicial de los datos para conocer en profundidad su estructura y el contenido que almacena. Además, se harán uso de técnicas estadísticas sofisticadas que permitirán un mayor nivel de comprensión de las variables analizadas desde un punto de vista multivariante. En concreto, este proyecto hace especial énfasis en los aspectos relacionados con los grupos sociales a los cuales hacen referencia (o no) los tratados de paz analizados.

Índice

Introducción **1**

Datos **2**

PA-X Agreements Database 2

Metodología del proyecto y proceso 3

Evidencias y conocimientos extraídos en la PEC2 4

Tipo de relaciones estudiadas y gráfico utilizado para ilustrarlas 5

Sistema de filtros 6

Diseño **7**

Estructura 7

Importancia de gráficos 7

Formato y aspectos gráficos 7

Bibliografía consultada **8**

I Introducción

Tras todos los conflictos armados acontecidos a lo largo del siglo XX, la sociedad se ha planteado como uno de sus grandes objetivos alcanzar a nivel global un estado de paz ideal entre todas las naciones. En este contexto, el concepto de tratado de paz ha emergido como una de las mejores herramientas para finalizar conflictos bélicos a lo largo de los últimos 30 años.

Con el objetivo de prestar una fuente fiable con la información de todos estos conflictos, desde la *Law School* de la *University of Edimburgh* se crea la *PA-X Peace Agreements Database*, una base de datos accesible para la comunidad científica con las declaraciones de paz firmadas y metadatos asociados de los 1868 acuerdos de paz sucedidos desde el año 1990. Muchos son los aspectos recogidos para cada uno de los acuerdos, desde las zonas geográficas en las que suceden los conflictos a los que pretenden poner fin como la presencia de referencias a todo tipo de aspectos, tanto a nivel social, cultural o político.

En vistas al contenido de la base de datos mencionada, el presente trabajo pretende llevar a cabo un ejercicio de análisis estadístico y, principalmente, de visualización interactiva de datos con el objetivo de descubrir algunas tendencias y patrones que permitan alcanzar un alto grado de comprensión de los datos recogidos sobre los tratados de paz.

Concretamente, en el marco de este trabajo, prestaremos especial relación al análisis de los tratados de paz desde un punto de vista social. Buscaremos estudiar si los acuerdos, que en principio únicamente tienen como objetivo la finalización de una guerra o conflicto entre países o grupos, tienen en consideración la perspectiva social entre sus principales líneas de actuación. De esta manera, la construcción de una visualización de calidad nos permitirá comprender como han variado las menciones a grupos específicos a lo largo de los años y su distribución entre zonas geográficas o tipología de conflictos, algo que nos permitirá caracterizar los acuerdos en función de todas estas características.

Algunas de las hipótesis iniciales que pretendemos abordar y contrastar mediante la consecución del presente proyecto podrían ser las siguientes.

- Los acuerdos que hacen algún tipo de referencia a un grupo específico, por ejemplo, los refugiados, también la hacen a otros grupos, como los que están definido por su raza étnica.
- La proporción de acuerdos que referencian a los grupos sociales se mantiene constante a lo largo del tiempo y es independiente de la región en la que sucede el conflicto o su tipología.
- Existe una agrupación de acuerdos en base a las referencias a los diferentes grupos sociales que hemos considerado en este proyecto.

Para conseguir estos objetivos, realizaremos una selección de los datos más interesantes para la consecución de nuestro proyecto y todas las técnicas de preprocesamiento que sean necesarias. Posteriormente, llevaremos a cabo un análisis estadístico multivariante que nos permitirá comprender las relaciones intrínsecas de los acuerdos como se asocian las características recogidas. Finalmente, llevaremos a cabo la construcción de un cuadro de mandos interactivo que permitirá al lector interactuar con las visualizaciones integradas en el mismo.

II Datos

PA-X Peace Agreements Database

Para llevar a cabo este proyecto de visualización de datos hemos utilizado como principal fuente de datos la ya mencionada *PA-X Peace Agreements Database*. Se trata de un conjunto de datos que recoge 266 variables y metadatos para cada uno de los 1868 acuerdos de paz considerados. Todos los metadatos presentados se pudieron clasificar en dos tipos de bloques de datos.

- **Información taxonómica**. Hablamos del conjunto de metadatos que determina las características principales del tratado en cuestión. Compuesto por los primeros 26 campos del conjunto de datos, este bloque de información permite establecer una clasificación de los tratados en función de numerosos aspectos como la zona geográfica en la que se enmarca el conflicto, la tipología de este, la fase en la que se encuentra o la fecha concreta en que se firmó.

- **Información referencial**. Por otra parte, encontramos el conjunto de metadatos compuesto por los campos restantes, que indican, para diferentes aspectos y temáticas, la presencia de referencias y, en algunos casos, el grado de referencia en el acuerdo en cuestión. Este bloque de información, a diferencia del anterior, nos permite conocer el contenido del tratado en función de si hace referencia, por ejemplo, a aspectos económicos, políticos o sociales.

Como vemos, nos encontramos ante un conjunto de datos que presenta un número muy elevado de variables susceptibles de ser analizadas, algunas sin embargo más relevantes que otras en función del tipo de análisis que se quiera llevar a cabo. En el contexto específico de nuestro trabajo, no todas las variables resultan de vital importancia para llevar a cabo la construcción de nuestra visualización interactiva, y es importante acotar el dominio de metadatos utilizados para establecer unos análisis de calidad. Por este motivo, previamente a la etapa de preprocesamiento y análisis, se ha llevado a cabo un paso previo de selección de las variables más importantes en el marco de nuestro proyecto.

El subconjunto de metadatos utilizado se encuentra resumido en la Tabla 1, mostrando para cada uno de los atributos su información básica, el tipo de variable, su nombre original y su nombre recodificado para una mayor representatividad del concepto que se pretende almacenar.

Nombre original	Nombre definido	Información almacenada	Tipo de variable
`Con`	`Country`	País en el que originó el conflicto	Cualitativa nominal
`Contp`	`ConflictType`	Tipología del conflicto	Cualitativa nominal
`PPName`	`Process`	Proceso de paz	Cualitativa nominal
`Reg`	`Region`	Región en la que tuvo lugar el conflicto	Cualitativa nominal

Dat	Date	Fecha en que se firmó el acuerdo	Fecha
Stage	Stage	Fase en la que se encuentra el conflicto	Cualitativa nominal
Loc1ISO	Loc1ISO	Codificación ISO del país en el que se originó el conflicto	Cualitativa nominal
GRef	Gref	Grado de importancia de la referencia a refugiados	Cualitativa ordinal
Gra	Gra	Grado de importancia de la referencia a grupos étnicos	Cualitativa ordinal
GCh	GCh	Grado de importancia de la referencia a niños	Cualitativa ordinal

Tabla 1. Variables seleccionadas para el proyecto de visualización

Como ya hemos percibido en la primera toma de contacto con los datos, para las variables relacionadas con el tipo de referencia hacia grupos sociales, echamos en falta campos de ausencia/presencia para dichas referencias. Por esta razón, hemos llevado a cabo una recategorización de dichas variables, dando lugar a la creación de nuevos atributos que indicaban la presencia de referencia hacia los grupos sociales considerados, independientemente de su grado de importancia.

Por tanto, a las 10 variables inicialmente consideradas, añadimos 3 variables dicotómicas que indicaban la presencia/ausencia de referencias a los grupos de refugiados, en función de la etnia y niños.

Metodología del proyecto y proceso

Para llevar a cabo el desarrollo del proyecto, se ha elaborado una planificación en X fases que se describen a continuación.

- **Carga de datos en Excel**. Para la primera fase del proyecto, realizaremos una primera exploración y edición de los datos en Excel. En la propia hoja de cálculo realizaremos el cambio de nombre de las variables, tal y como se ha descrito en la Tabla 1. Además, las variables prescindibles para el proyecto serán eliminadas.

- **Carga de datos en Tableau**. Una vez realizada la selección de variables y su renombramiento, llevaremos a cabo la carga de datos en Tableau para comenzar los primeros análisis visuales y construcción del cuadro de mandos. En este punto, también llevaremos a cabo la creación de las nuevas variables definidas en el apartado anterior utilizando las opciones de recodificación de Tableau.

- **Análisis estadístico en R**. En paralelo a la fase anterior, con el objetivo de comprender cómo se relacionan algunas de las categorías de los metadatos recogidos en los tratados, llevaremos a cabo un análisis estadístico de los datos mediante técnicas multivariantes utilizando el lenguaje de programación R. Antes de comenzar con el análisis, llevaremos a cabo todas las transformaciones y reconversiones necesarias para que todas las variables analizadas se

interpreten con el tipo de datos que corresponda en el entorno de programación. En concreto, y dada la naturaleza y tipología de nuestros datos, utilizaremos las conocidas técnicas factoriales, enmarcadas en el contexto del aprendizaje no supervisado. Una de las técnicas de este tipo más recurridas ante la presencia de un número elevado de variables cualitativas es el análisis factorial de correspondencias múltiple. Este método permite, mediante la extracción de nuevos factores a partir de las variables originales, obtener una representación gráfica de las categorías en cuestión, de manera que cuanto más próximas en el espacio se encuentren dichas categorías, mayor será su asociación.

El objetivo de utilizar esta técnica será, por tanto, la extracción de las coordenadas de cada una de las categorías en el espacio de dimensión reducida y su posterior exportación a Tableau para proceder a su representación en un gráfico similar a un diagrama de dispersión.

Adicionalmente, también se llevará a cabo un agrupamiento de los acuerdos en función del grado de referencia a los grupos sociales considerando. Para ilustrar dicho agrupamiento, haremos uso del algoritmo de *clustering* jerárquico específico para variables cualitativas (nominales y ordinales), y mostraremos el dendograma resultante.

Utilizando ambas técnicas podremos identificar tanto la asociación de las variables cualitativas recogidas en los acuerdos como los grupos de acuerdos en los cuales las referencias a grupos sociales son más similares.

- Construcción del cuadro de mandos. Finalmente, tras llevar a cabo todo el proceso de preprocesamiento y análisis mencionado, procederemos a la construcción del cuadro de mandos final que contendrá la visualización de la información de los tratados paz haciendo hincapié en los grupos sociales. El producto resultante será alojado en la cuenta personal del autor y podrá ser explorado de manera interactiva por cualquier usuario o investigador que lo desee.

Evidencias y conocimientos extraídos en la PEC2

Para la realización de la visualización final y el análisis estadístico, hemos utilizado algunas de las evidencias y patrones extraídos en la realización de la actividad anterior. Por ejemplo, vimos que la variable `Status`, que define el estado del conflicto al que hace referencia el acuerdo, no mostraba mucha variabilidad en los valores que tomaba, donde la mayoría de los conflictos se encontraban en un estado en el que ambas partes estaban de acuerdo. No decidimos, por tanto, analizar la información de los acuerdos desde dicha perspectiva.

Sin embargo, vimos que otras variables, como por ejemplo `Reg` y `Contp`, cada una almacenando información sobre la región en la que se desarrolló el conflicto en cuestión y su tipología, respectivamente, sí que presentaban una variabilidad que merecía ser estudiada.

Por otra parte, a la hora de visualizar la presencia y grado de referencias a los diferentes grupos sociales considerados (en el caso de la actividad anterior, únicamente grupos definidos por la raza y en

función de si eran refugiados o no) a lo largo de los años, pudimos apreciar distribuciones diferentes. En consecuencia, comprobamos que analizar los tratados de paz desde las perspectivas de los grupos sociales podía ser potencialmente interesante, lo que nos llevó a elegir dicho enfoque para llevar a cabo nuestro proyecto.

Tipo de relaciones estudiadas y gráfico utilizado para ilustrarlas

A lo largo del cuadro de mandos que se diseñará, el objetivo será estudiar las relaciones entre las variables analizadas en el estudio. A continuación, enumeramos las relaciones que buscamos identificar y el gráfico utilizado para ilustrarlas en el cuadro de mandos.

- Para comprender la distribución en el tiempo del porcentaje de acuerdos en los que se hace referencia a los grupos sociales considerados (y qué tipo de referencia, es decir, el grado de importancia), utilizaremos un diagrama de barras apilado, con una barra para cada uno de los 30 años considerados. De esta manera, en cada uno de los años podremos apreciar el porcentaje de acuerdos en los que se hacía una referencia, por ejemplo, sustancial a los grupos étnicos.

- Para conocer a nivel global la distribución del grado de importancia de las referencias a grupos sociales se hará uso de diagramas de burbuja, donde cada una de las burbujas corresponderá con un proceso de paz en concreto. De esta manera, además de conocer la distribución desde un punto de vista general, podremos conocer qué procesos contribuyen en mayor medida a cada uno de los grados de importancia de las referencias.

- Para conocer la distribución geográfica de los acuerdos en los que se hacía algún tipo de referencia a cada uno de los grupos sociales considerados, si hizo uso de mapas cartográficos, de manera que cada país presentará un tono de color más intenso cuanto mayor número de acuerdos presente con alguna referencia al grupo en cuestión.

- Para estudiar el tipo de relación (positiva o negativa) multivariante entre las categorías de las variables analizadas, haremos uso de la representación factorial resultado del análisis de correspondencias múltiple, ilustrada en forma de gráfico de dispersión, en la que cada punto se identifica con una categoría específica de una variable cualitativa en concreto. De esta manera, los puntos (categorías) más próximos en el espacio factorial calculado indicarán una presencia y comportamientos similares en los tratados.

- Finalmente, para conocer cómo se agrupan los tratados de paz, ilustraremos mediante un dendograma el resultado del algoritmo de c*lustering* jerárquico, que nos permitirá identificar aquellos grupos en los cuales las referencias a la perspectiva social son similares.

Sistema de filtros

Para habilitar al usuario que pretenda estudiar la visualización construida, incluiremos algunos filtros que permitan la transición entre visualizaciones de una manera sencilla. Definiremos, por tanto, filtros que, al utilizarse, afectaran a todas las partes del cuadro de mandos que sean interactivas.

Para que el usuario pueda interactuar con los datos, realizar selecciones y segmentaciones, habilitaremos dos maneras de filtrar la información de los tratados de paz. Por un lado, en la parte superior derecha del cuadro de mandos encontraremos la sección de filtros manuales explícitos, en las que el usuario podrá editar libremente qué subconjunto de los datos visualizará. En esta parte encontramos tres filtros principales.

- **Filtro de fechas**. Este filtro nos permitirá, mediante la selección de un intervalo de tiempo específico, filtrar la visualización y todos los gráficos que integra en un rango de tiempo determinado. De esta manera, el usuario fácilmente podrá seleccionar en el propio cuadro de mandos que período desea visualizar. Esto permitirá filtrar la información, por ejemplo, por una década específica y focalizar la exploración de los datos en dicho intervalo de tiempo.

- **Filtro de región.** Este filtro nos permitirá realizar una selección múltiple de las posibles regiones en las que han tenido lugar los conflictos. Al igual que el filtro anterior, la selección de regiones afectará directamente a todo el contenido interactivo del cuadro de mandos, pudiendo visualizar la información de los contenidos por cada una de las regiones.

- **Filtro de tipología de conflicto**. Por último, este filtro permitirá al usuario realizar una selección múltiple del tipo de conflicto a los que hacen referencia los acuerdos de paz. Este filtro también afectará a los gráficos restantes del cuadro de mando.

En segundo lugar, y para una mayor versatilidad en las posibilidades de la visualización, permitiremos el filtrado de información en base a la selección de datos en los propios gráficos. Esto permitirá al usuario filtrar la información disponible en base a alguna de las propiedades de los diagramas y mapas creados. De esta manera, el usuario podrá filtrar la información, por ejemplo, teniendo únicamente en cuenta un determinado país como Colombia o los tratados en los que se hizo una referencia retórica a refugiados (o ambas).

Este sistema de filtros permitirá un control total sobre las características a visualizar de los propios tratados.

El cuadro de mandos elaborado únicamente presentará una sola página donde se dispondrán todos los elementos visuales. La primera parte del cuadro dispondrá la información para los grupos sociales considerados de manera independiente. La segunda parte del cuadro de mandos contendrá la información sobre las relaciones entre las categorías de las variables (diagrama de dispersión/mapa factorial) y el resultado de la agrupación obtenida mediante la técnica del *clustering* jerárquico.

III Diseño

Para una mayor comprensión del diseño del cuadro de mandos final, recomendamos al lector prestar especial atención al boceto adjunto al proyecto donde se detallan la estructura y disposición del producto final.

Estructura

Para la creación del cuadro de mandos final, lo primero que encontraremos será el título del mismo y la sección de filtros, que nos permitirá filtrar la información en base a los diferentes criterios ya mencionados.

A continuación, dispondremos de manera secuencial los gráficos de cada uno de los grupos sociales considerados. El primer bloque contiene la información asociada al grupo de refugiados. El segundo bloque contiene la información asociada a los grupos étnicos. Por último, el tercer bloque hace referencia al grupo de niños.

Finalmente, encontramos un cuarto bloque de información que contiene la información relativa a los análisis de estadística multivariante.

Importancia de gráficos

A pesar de que los tres primeros bloques son los que permiten analizar la información de manera interactiva, consideramos a los gráficos y la información del cuarto bloque como la más importante puesto que es la que de verdad arrojará algunas conclusiones en base a la realización del proyecto desde un punto de vista multivariante.

Formato y aspectos gráficos

El cuadro de mandos elaborado se ha construido con la intención de que únicamente se encuentre disponible en formato digital, puesto que entendemos que es la única manera de aprovechar las capacidades interactivas de la herramienta con la que se ha construido.

Como puede apreciarse en el boceto adjunto, la visualización a nivel general presentará un tono grisáceo, con el objetivo de utilizar colores más vivos para resaltar algunos aspectos importantes de los gráficos.

Concretamente, utilizamos tonos verdosos para hacer referencia a la información acerca los refugiados, tonos morados para hacer referencia a la información sobre grupos étnicos y tonos azulados para hacer referencia a los niños. Asociar cada uno de estos aspectos a una tonalidad distinta permitirá al lector una rápida asociación de conceptos durante la exploración interactiva de los gráficos.

Con respecto a la tipografía utilizada para el proyecto, utilizaremos por su claridad y su fácil lectura y comprensión la familia de fuentes nativa de Tableau, variando los tamaños de letra en función del elemento que pretende informar.

Destacaremos la información del lector especialmente en el bloque de filtros, donde indicaremos con algún símbolo la presencia del mismo para que no pase desapercibido y pueda aprovecharse el potencial de la visualización construida.

IV Bibliografía consultada

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons.

Koshy J. Effective Data Visualization with 8 Design Principles [en línea]. *PromptCloud Blogs*. (24 de septiembre de 2020). [Consulta: 6 de diciembre de 2020]. Disponible en: https://www.promptcloud.com/blog/design-principles-for-effective-data-visualisation/

Sleeper, R. (2018). Practical Tableau: 100 Tips, Tutorials, and Strategies from a Tableau Zen Master. O'Reilly Media, Inc.

Prueba de Evaluación Continua 4

Peace Agreements: Social Groups Perspectives

Enlace a la visualización: *https://public.tableau.com/profile/albergcg#!/vizhome/pax_visualization/FinalDashboard*

Enlace al repositorio: *https://github.com/albergcg/pax_visualization*

I · Síntesis

Tras llevar a cabo la Prueba de Evaluación Continua (PEC) 2 y la PEC 3, esta PEC 4 pretende finalizar el proyecto de visualización de datos presentado y llevado a cabo en la asignatura *Visualización de Datos* del Máster Universitario en Ciencia de Datos (*Data Science*) de la Universitat Oberta de Catalunya.

Una vez presentada la primera visualización exploratoria de los datos y elaborar la documentación del proyecto, el objetivo de esta última práctica es presentar una visualización de datos final y publicarla para hacerla accesible a la comunidad, además de proporcionar una descripción lo más técnica posible del proyecto, de las herramientas utilizadas para su consecución final y realizar una evaluación general.

Como ya se explicó en las pruebas anteriores, el presente proyecto de visualización de datos se enmarca en la explotación estadística y visual de la perspectiva social de la *PA-X Peace Agreements Database*, la base de datos pública elaborada por la *Law School* de la *University of Edimburgh*, que almacena la información de los diferentes acuerdos de paz firmados a lo largo de todo el mundo desde 1990.

En concreto, en la visualización final del proyecto se ha construido un cuadro de mandos o *dashboard* que:

- Permite analizar desde diferentes dimensiones y mediante un sistema de filtros cómo ha variado la frecuencia de acuerdos que atendían a diferentes aspectos sociales de manera individualizada, como por ejemplo la mención a grupos generalmente infrarepresentados en el contexto del proyecto como los refugiados o los niños.

- Permite establecer una serie de filtros sobre diferentes aspectos (fecha, región y estado del conflicto) de los acuerdos que permiten explorar la información anterior de manera interactiva, así como la selección de características sobre los mismos gráficos.

- Permite explorar mediante un análisis multivariante de las variables asociadas al grado de referencia de los grupos sociales cómo era su asociación. De esta manera, se pudo observar que de manera general que si un acuerdo referenciaba algunos de los grupos sociales considerados, lo hacía adicionalmente para los demás.

- Por último, se llevó a cabo una agrupación de los acuerdos en función del tipo de referencia que hacían a los grupos sociales considerados mediante la aplicación de una técnica de *clustering jerárquico*.

II · Proceso de aprendizaje y valoración de esfuerzo/resultado

A la hora de llevar a cabo proyectos relacionados con la ciencia de datos, uno de los primeros aspectos a tener en cuenta y que a menudo no se le presta la atención que merece es el contexto en el que se enmarcan los datos presentados. En este sentido, el contexto en el que se ha enmarcado el presente proyecto es, precisamente, lo que ha supuesto el primer proceso de aprendizaje sustancial. En la primera toma de contacto con los datos, aprendimos a nivel técnico qué era un acuerdo de paz y cuáles eran las diferentes características que se utilizaban para clasificarlos. Además, pudimos descubrir algunas tendencias y patrones curiosos que nos permitieron conocer un poco más el conjunto de datos, el tipo de información que se proporcionó y motivaron el propósito de nuestra visualización de datos final.

A nivel técnico, el desarrollo del proyecto nos ha permitido conocer y aprender el funcionamiento de una nueva herramienta para la construcción de visualizaciones de datos interactivas desconocida previamente para el autor: Tableau. A pesar de que inicialmente supuso un reto comprender su funcionamiento y las posibilidades que podía ofrecer, tras la consecución del proyecto hemos podido adquirir un nivel de destreza suficiente como para desarrollar la visualización presentada. No obstante, y si bien es cierto que Tableau es una herramienta obligatoria para cualquier científico de datos que se precie, no está exenta de algunas limitaciones a destacar. Como hemos podido comprobar, explotar al máximo su potencial únicamente será posible para conjuntos de datos perfectamente estructurados en tablas y con la información perfectamente organizada, puesto que no dispone de funciones específicas de limpieza de datos. Además, no permite un control de versiones que ayude a llevar a cabo una trazabilidad en las visualizaciones desarrolladas y las opciones de análisis de datos y estadísticas son bastante limitadas, lo que obliga a utilizar herramientas y lenguajes de programación externos para según que propósito se busque.

Como consecuencia de esto, para llevar a cabo las visualizaciones y análisis estadísticos avanzados que se precisaron, se hizo uso del lenguaje de programación R, ampliamente conocido por la variedad de paquetes y funciones disponibles para llevar a cabo dichos propósitos. En este sentido, esta fase del proyecto también favoreció el aprendizaje de nuevas herramientas. Se profundizó en la aplicación de técnicas de análisis multivariante sobre el conjunto de datos, proceso en el cual se concluyó que, dado el alto número de variables cualitativas que se presentaban, se necesitaba hacer especial énfasis en los métodos específicos para este tipo de datos. De manera similar, se profundizó en la aplicación de técnicas de agrupamiento o *clustering* con el objetivo establecer grupos en los acuerdos en función del tipo de referencia que hacían a los grupos sociales considerados.

Una vez finalizado el proyecto, la valoración global del mismo ha sido muy positiva. A pesar de los aprendizajes descritos en las líneas anteriores, el proyecto en sí, por su planteamiento haciendo énfasis en la presentación y visualización de resultados, ha supuesto todo un viaje. Una de las partes en las que pensaba que iba a demorar más en el proceso de aprendizaje es el dominio Tableau. Sin embargo, dada su facilidad de uso y de su interfaz amigable, finalmente la curva de aprendizaje no fue demasiado alta. Un claro ejemplo de ello es que en el inicio de la PEC2, la idea inicial era realizar buena parte del análisis exploratorio inicial de los datos mediante R y el paquete ggplot2. Finalmente, tras comenzar a utilizar Tableau y adquirir destreza rápidamente, se decidió utilizar como principal herramienta para la primera parte del proyecto.

III · Evolución con respecto al proyecto inicial

Si valoramos la evolución del proyecto desde la primera prueba de concepto en la PEC2 hasta la consecución de la visualización final presente en la PEC4, podemos destacar varias mejoras sustanciales.

- Una de las partes más importantes que se ha mejorado es el aprovechamiento del espacio. Se ha rediseñado el cuadro de mandos propuesto inicialmente y se han dispuesto los elementos en un único panel, favoreciendo así la facilidad de exploración de los aspectos de los acuerdos y la comparación entre estos en función del grado de referencia a los grupos sociales.
- Por otra parte, se ha añadido un sistema de filtros más completo, añadiendo nuevas variables que por sus características y naturaleza pueden ser de interés para el lector y que pueden proporcionar nuevas perspectivas para la visualización de los resultados.
- Con respecto a la primera visualización, se ha diversificado el tipo de gráfico utilizado y, para cada uno de los grupos sociales considerados en el proyecto, no se ha repetido ningún tipo de gráfico, enriqueciendo el cuadro de mandos y favoreciéndolo estéticamente.
- El gran cambio con respecto a la primera visualización es la parte inferior del cuadro de mandos. Se ha añadido un gráfico de dispersión útil para explorar la asociación de las variables que definían las referencias a los grupos sociales. Desgraciadamente, y dado el tipo de análisis llevado a cabo, este gráfico a diferencia de los demás se ha presentado de manera estática, sin proporcionar al usuario, aunque carezca de sentido, la posibilidad de editarlo. En este sentido, y de manera análoga a la justificación anterior, se ha visualizado el dendograma resultante del agrupamiento de acuerdos de manera estática.

IV · Descripción técnica del proyecto

Herramientas, lenguajes de programación y librerías

Para el desarrollo del proyecto de visualización de datos llevado a cabo, se han utilizado un número elevado de herramientas y técnicas que enumeramos a continuación de la manera más completa posible.

- Inicialmente, se utilizó la herramienta de ofimática Microsoft Excel 2019 para la primera exploración de los datos y la selección de las variables deseables para nuestro proyecto en concreto.
- Como herramienta principal del proyecto y específica para la creación del producto final se utilizó Tableau 2020.3.
- De manera adicional, para una parte de la exploración inicial del conjunto de datos y para la realización de los análisis estadísticos empleados, se utilizó el lenguaje de programación R (R Core Team, 2020), en su versión 4.0.3.

- Además de R propiamente, se han utilizado varios paquetes y librerías que han permitido llevar a cabo el desarrollo técnico tanto de las técnicas como de los gráficos. Citamos a continuación lo que más relevancia han tenido.
    - Para la carga, tratamiento y recodificación de datos se ha utilizado el conjunto de paquetes y funciones contenidos en el conocido tidyverse (Wickham et al., 2019).
    - Para los gráficos empleados, se ha utilizado la librería ggplot2 (Wickham, 2016).
    - Para la aplicación del método de *clustering*, se han utilizado los paquetes cluster (Meachler et al., 2019) y dendextend (Galili, 2015).
    - Para el cálculo del Análisis de Correspondencias Múltiple, se ha utilizado el paquete factominer (Lê y Husson, 2008).

Licencia

Para el presente proyecto, se ha definido como licencia tanto para la visualización, los contenidos y el código presentados la *MIT License*, al ser una de las más permisivas y utilizadas en la comunidad, con el objetivo de que cualquier usuario que quiera continuar el proyecto pueda hacerlo libremente.

Esta licencia especifica que, en caso de ser utilizado, deberán ser debidamente referenciados e indicar nuevamente el tipo de licencia presentada. Cabe mencionar además que esta licencia no ofrece ningún tipo de garantía por parte de los autores.

Desarrollo del proyecto

Con respecto a las fases de las que ha constado el proyecto, explicaremos brevemente en qué consistió cada una.

- **Fase I (PEC 2)**. En esta primera fase del proyecto, se llevó a cabo una exploración inicial del conjunto de datos. En este punto fue de vital importancia comprender el contexto en el que estábamos trabajando y conocer a qué tipo de datos nos estábamos enfrentando. El mayor reto de esta fase fue la selección del dominio concreto a trabajar, donde nos decantamos por prestar especial atención a la perspectiva social de los acuerdos. Se llevó a cabo dicha exploración tanto con Tableau como con R y se desarrolló un primer cuadro de mandos inicial que sirvió como prueba piloto de nuestro producto final.
- **Fase II (PEC 3)**. En la segunda fase del proyecto, el reto más importante fue la decisión de la construcción del cuadro de mandos final, partiendo de la solución intermedia definida en la PEC2. Se llevó a cabo un proceso de documentación formal del proyecto, en el que se introdujo la problemática de los acuerdos de paz, se describieron de manera muy concisa las variables involucradas en el proyecto y se decidió un primer aspecto mediante un *mockup* del cuadro de mandos final.

- **Fase III (PEC 4)**. Finalmente, en esta última fase del proyecto se ha llevado a cabo la construcción de la visualización final y se ha publicado de manera tanto el código empleado como el propio cuadro de mandos final. Además, todas las herramientas utilizadas se han citado de manera concisa.

Análisis estadístico

Por último, nos parece interesante abordar desde una perspectiva técnica los métodos estadísticos utilizados y empleados en la visualización final tanto por su importancia en el cuadro de mandos como por algunos aspectos en el desarrollo que implicaron un reto en la consecución del proyecto.

Uno de los primeros objetivos que se plantearon una vez se definió la temática del proyecto fue buscar si existía algún tipo de relación entre el tipo de referencia que se hacía de manera conjunta a los grupos sociales considerados. En Estadística, cuando buscar comprender relaciones de manera conjunta, entran el juego las técnicas de análisis multivariante. Dado que lo que buscábamos era comprender la relación entre tres variables categóricas, pudimos comprobar que la primera opción de análisis contemplada, el Análisis de Componentes Principales, únicamente puede ser utilizado con variables continuas.

Por esta razón, decidimos hacer uso del Análisis de Correspondencias Múltiple (Kassambra, 2017), una extensión del Análisis de Componentes Principales para variables categóricas. En concreto, este método extrae un espacio de dimensión reducida en el que se representan las categorías de las variables cualitativas involucradas en el análisis. El procedimiento habitual consiste en extraer el primer plano factorial, que es el que se muestra en el *dashboard* final, e interpretar la proximidad de las categorías en el espacio. Aquellas categorías más próximas en el plano se interpretarán con una asociación positiva, mientras que las que más se distancien se interpretarán con una asociación negativa.

Por otra parte, se buscó la obtención de grupos de acuerdos en función de los grados de referencia a los grupos sociales considerados mediante un algoritmo de *clustering* jerárquico. A la hora de establecer agrupaciones, uno de los conceptos clave que deben de establecerse es el concepto de distancia estadística. En el caso de variables numéricas, la distancia más conocida es la euclídea, pero en nuestro caso particular nuevamente debíamos de buscar una alternativa para variables cualitativas.

Para la realización del agrupamiento, se tomó como distancia la distancia de Gower, especialmente útil para trabajar con todo tipo de datos, incluyendo variables cualitativas ordinales como era nuestro caso. Una vez calculada la matriz de distancias entre cada par de acuerdos, se realizó la agrupación.

Referencias

Galili T. (2015). *dendextend*: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics

Lê, S., Josse, J. y Husson, F. (2008). *FactoMineR*: An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. y Hornik, K. (2019). *cluster*: Cluster Analysis Basics and Extensions. R package version 2.1.0.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Wickham H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.