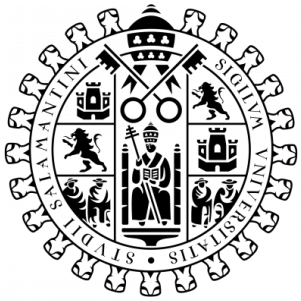


VNiVERSiDAD
D SALAMANCA

Grado en Estadística
Trabajo de Fin de Grado

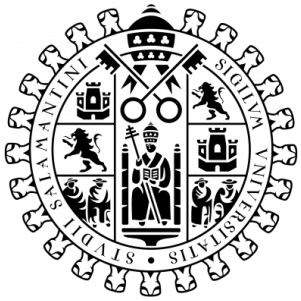
TÉCNICAS ESTADÍSTICAS APLICADAS A LA DIAGNOSIS DE CRIMINALIDAD

APPLICATION OF STATISTICAL TECHNIQUES FOR CRIMINALITY DIAGNOSIS

Autor:
David Briz Benito

Dirigido por:
María Teresa Cabero Morán

7 de julio de 2022



VNiVERSiDAD D SALAMANCA

Grado en Estadística
Trabajo de Fin de Grado

TÉCNICAS ESTADÍSTICAS APLICADAS A LA DIAGNOSIS DE CRIMINALIDAD

APPLICATION OF STATISTICAL TECHNIQUES FOR CRIMINALITY DIAGNOSIS

Autor: David Briz Benito	**Tutora:** Mª Teresa Cabero Morán
Firmado por BRIZ BENITO DAVID - ***1997** el día 07/07/2022 con un certificado emitido por AC FNMT Usuarios **Firma:**_____	Firmado por CABERO MORAN MARIA TERESA - ***6717** el día 07/07/2022 con un certificado emitido por AC FNMT Usuarios **Firma:**_____

Índice general

Capítulo 1

Introducción

Según la Real Academia Española de la lengua, la criminalidad se define como:

1. "Cualidad o circunstacia que hacen que una acción sea criminal"
2. "Hecho de cometerse crímenes"
3. "Número proporcional de crímenes en un tiempo y en un lugar concretos"

Estas tres acepciones se circunscriben alrededor de un mismo campo de estudio en los distintos cuerpos de seguridad, entre otros, en España, los dos exponentes más significativos e interesantes para esta investigación son, el Cuerpo Nacional de Policía, de carácter civil, perteneciente al Ministerio del Interior y la Guardia Civil, de carácter militar, perteneciente al Ministerio de Defensa. De llevar a cabo estas tareas, se encarga la unidad de Policía Científica en la Policía Nacional y el Servicio de Criminalística (SECRIM) en la Guardia Civil, teniendo la función de prestar los servicios de criminalística, identificación, analítica e investigación técnica, así como la elaboración de los informes periciales y documentales.

La investigación científica aplicada al ámbito criminalístico engloba un conjunto muy amplio de técnicas multidisciplinares que abarcan campos científicos muy diversos (como puede ser la química, la medicina, la informática, la matemática, etc.) orientadas a fines diferentes dentro de un mismo ámbito. Así, en este caso se hará uso de la estadística para hacer un balance del valor de las técnicas estadísticas para el estudio de la criminalidad. esta ciencia va a servir como herramienta que va a permitir extraer y comprender información acerca de los datos con los que se trabaje con la finalidad de proporcionar a los miembros pertinentes de las Fuerzas y Cuerpos de Seguridad de la información que necesiten para aplicar las medidas necesarias que sean consideradas por ellos como profesionales en cada caso, quedando en manos de los investigadores estadísticos la decisión de aplicar las técnicas más adecuadas según los datos que se dispongan para optimizar la calidad de los resultados y facilitar la precisión en las decisiones posteriores que queden a cargo de los profesionales de estos cuerpos.

Para determinar la metodología aplicada a la investigación criminal, en primer lugar, el investigador parte de unos conocimientos previos sobre el problema que se quiere estudiar ya sean derivados de otros problemas similares o factores relacionados con el objeto de estudio, así como de todos los demás saberes no solamente técnicos que tiene a su disposición. Antes de realizar una investigación, siempre se parte de una información técnica,

general y específica, amplia, recopilada desde otras investigaciones que permitirán desarrollar la investigación que se pretende. Este problema se manifiesta cuando hace falta información que permita resolver el objetivo que se plantea y para determinar lo que se desconoce es necesario partir de lo conocido aprovechando los medios de los que se dispone a través de los cuáles se obtendrá la información que conducirá a la posterior solución del mismo. En toda investigación criminal, el problema, se debe de considerar como el inicio de la investigación que a través de un procedimiento de pensamiento reflexivo se pretende llegar a la meta que es su solución. Este proceso, se podría decir que es el más importante, sin la identificación del problema no se puede desarrollar el trabajo o si no se consigue identificar correctamente el problema, el resto del proceso, aunque se desarrolle correctamente no va a tener ningún sentido porque los resultados que se extraerán no serán los que se buscan y eso significa que todo el tiempo que se ha invertido en este procedimiento ha resultado inservible y por lo tanto, se ha malgastado. Es importante señalar que la investigación científica se desarrolla en un marco de un conocimiento previo que se debe tener en cuenta por parte del investigador. Un paso importante después de definir el problema es delimitar este mismo, es decir, se debe especificar el sentido de la pregunta, evitando cualquier tipo de ambigüedad.

A continuación, se procede a recoger la información sobre el hecho que se quiere investigar con el único objetivo de tener un concepto previo sobre el fenómeno criminalístico para poder abordar la investigación con mayor facilidad y de esta manera resulte más favorable el desarrollo de los métodos posteriores.

Una vez se dispone de el o los conceptos sobre el fenómeno criminalístico a investigar, se elaboran una serie de hipótesis a las que se tratará de responder posteriormente. Las hipótesis son conjeturas que se plantean con el objetivo de posibilitar una solución al problema que se quiere investigar siempre utilizando el conocimiento implícito del investigador que dispone sobre el problema. Las hipótesis tratan de relacionar variables referentes al problema y el objetivo será confirmar o desmentir estas mismas. Las variables se pueden entender como un tipo de propiedad o dimensión a estudiar que puede tomar diversos valores o clasificarse en distintas categorías o "características de un hecho o fenómeno susceptibles de adoptar valores numéricos" (ALBAJAR y MARTÍN, 2006), así se pueden percibir dos tipos de variables; las variables cualitativas que, como dice su nombre expresan propiedades clasificatorias o características del objeto en estudio y las variables cuantitativas son aquellas que pueden medirse a través de una escala numérica y, a su vez, estas últimas, pueden ser discretas si los valores pertenecen al conjunto de los números naturales y continuas si los valores pertenecen al conjunto de los números reales. Si el proceso de investigación concluye con que dicha relación es existente, el problema habrá sido explicado. También, con toda la información que se conoce hasta el momento, se realiza un análisis causal, teniendo en cuenta todas las variables que intervienen dentro del fenómeno que se está investigando.

Muchas veces, también se debe de llevar a cabo un estudio preoperacional desde el que se pretende determinar si se está en condiciones de investigar. Para ello, se hace un ensayo y se cuestiona sobre los problemas que pueden surgir, los medios de los que se dispone, así como el tiempo para realizarla.

Como paso previo al análisis, se lleva a cabo el proceso de observación y recolección de datos, que, junto a los que se dispone conforman el conjunto de información con la que se va a trabajar. Para que se propicie una buena recolección de datos, se debe de llevar a

cabo el proceso de medición que se compone de varias fases. En primer lugar, el contraste de estos mismos, es decir, de cerciorarse que son correctos y que se han recopilado bien. A continuación, se lleva a cabo el proceso de validación que se basa en estudiar la validez y la fiabilidad de los datos obtenidos, para que los resultados de la investigación sean los más cercanos a la realidad que sea posible.

Por último, el proceso de análisis de los datos que se compone de la combinación de un análisis estadístico y un análisis cualitativo de los mismos. Todos los resultados de estos análisis se tienen que interpretar aplicando distintas técnicas que permiten clarificar los resultados obtenidos, con el objetivo de elaborar finalmente el informe de investigación criminal.

Aunque el objetivo fundamental de este trabajo no es el realizar una investigación sobre unos datos para sacar las conclusiones sobre los mismos, sino que más bien se trata de valorar y demostrar el grado de utilidad de distintas técnicas estadísticas a la hora de hacer una correcta diagnosis de criminalidad, adecuándolas a la tipología de datos que se disponga para poder analizarse, ya que no siempre se van a poder aplicar las mismas técnicas sobre un tipo de datos que sobre otros, además del diseño e implementación en código de técnicas para su análisis; ha sido necesario seguir la gran mayoría de los puntos mencionados anteriormente a la hora de realizar el trabajo. Las conclusiones sobre lo que aportan los resultados de los análisis estadísticos pasarían a segundo plano. En todo el trabajo, se ha seguido en mayor o menor medida el proceso que se describía anteriormente ya que para cada conjunto de datos o mejor dicho, para cada fase del trabajo, se ha establecido un problema a resolver, partiendo de unos conocimientos previos, se han seleccionado una serie de herramientas como pueden ser las técnicas estadísticas en sí y los *software* que se han considerado apropiados para cada una de las dos fases además de otros conocimientos que se han utilizado, una vez se ha contado con esto se ha tratado de dar luz a esos objetivos a través, de un pensamiento reflexivo se ha decidido cuáles son los métodos que mejor convenian a cada caso. También se ha realizado el estudio preoperacional antes de aplicar las técnicas definitivamente para comprobar si eran las más adecuadas y si se estaba en condición de investigar según la "hoja de ruta" que se había planteado y, por último, se realizaban las operaciones estadísticas pertinentes que han dado lugar a extraer las conclusiones de cada uno de los estudios.

La aplicación de las técnicas estadísticas en el trabajo se ha realizado en dos fases como se señalaba anteriormente. En la primera, que corresponde a la aplicación de técnicas para la diagnosis de criminalidad en los datos de origen público, se disponían de los conocimientos previos sobre las técnicas estadísticas que se pensaban aplicar, del lenguaje de programación R sobre el que se han diseñado las herramientas que permiten analizar los datos en esta fase además del aprendizaje y el lenguaje aplicado a la criminalidad con el que se ha tenido que familiarizar para poder comprender cuáles serían las técnicas estadísticas más adecuadas para aplicarse y para entender que se podría hacer con ellos. Son destacables los conceptos de *hechos conocidos* que se definen como el "conjunto de infracciones penales y administrativas, que han sido conocidas por las distintas Fuerzas y Cuerpos de Seguridad, bien por medio de denuncia interpuesta o por actuación policial realizada motu propio (labor preventiva o de investigación)" (Ministerio del Interior, s.f.), *hechos esclarecidos* que "se clasifican como tales cuando en el hecho se dan las circunstancias de detención del autor «in fraganti». Identificación plena del autor, o alguno de los autores, sin necesidad de que esté detenido, aunque se encuentre en situación de libertad provisional, huido o muerto. Cuando exista una confesión verificada, pruebas sólidas o

cuando haya una combinación de ambos elementos. Cuando la investigación revele que, en realidad, no hubo infracción." (Ministerio del Interior, s.f.), *victimización* "viene referido al número de hechos denunciados por personas en los cuáles manifiestan ser víctimas o perjudicados por alguna infracción penal. Se diferencia del concepto de «víctima», ya que éste se refiere a personas individuales." (Ministerio del Interior, s.f.), *detención* "alcanza la lectura de derechos de la persona física, privándole de libertad y poniéndolo a disposición judicial, por atribuirle la comisión de una infracción penal." e *investigado* "será una persona física o jurídica a la que se atribuya la participación en un hecho penal. No se adoptan medidas restrictivas de libertad para esa persona imputada." (Ministerio del Interior, s.f.). Entonces, a partir de este punto de lo que se tratará es de resolver los objetivos planteados empleando un pensamiento reflexivo para saber cuáles son las técnicas más apropiadas para aplicar sobre los datos una vez que se ha desarrollado el estudio preoperacional, en este caso, gráficos de control y números índices, están listos para ser aplicados sobre los datos.

Por otro lado, en la segunda fase, referida al estudio de los datos del Cuerpo Nacional de Policía, los conocimientos previos de los que se dispone se engloban en todas las herramientas estadísticas, los *software* con los que se va a trabajar, en este caso, con SPSS y Weka y los conceptos relativos a la criminalidad que se deben tener en cuenta para entender los datos de los que se dispone. A partir de este punto, se tratará de resolver los objetivos planteados valorando a partir de un pensamiento reflexivo cuáles serán las técnicas más adecuadas para aplicar sobre los datos, en este caso, técnicas de estadística descriptiva, contrastes de hipótesis y un árbol de clasificación, después de haberse desarrollado el estudio preoperacional y tener el convencimiento que serán adecuadasl, y que se van a aplicar sobre los datos.

En el caso de los datos que están referidos a una variable temporal, se procede a pensar en técnicas estadísticas que involucren la variable tiempo, que además puedan considerarse útiles en el ámbito de la criminalidad como pueden ser los gráficos de control, a través de los que se podrá estudiar el comportamiento de las variables de criminalidad estudiadas y a partir de los cuáles, observando la evolución de los puntos correspondientes, los números índices que permitirán hacer comparaciones temporales; o las series temporales, que en este caso no se han utilizado debido a la reducida cantidad de años de los que se dispone para realizar un estudio con ellas significativo. Ahora, en el caso de la segunda fase, relativa a los datos del Cuerpo Nacional de Policía, se opta en primer lugar, por la disposición y la forma de los datos a hacer un estudio descriptivo para tener en primer lugar un conocimiento amplio de los mismos y poder orientar las posteriores técnicas estadísticas que se pueden aplicar. A continuación, se considera hacer una serie de análisis para contrastes de hipótesis para responder una serie de preguntas que se plantean y por último se considera muy valiosa la información que se puede obtener desde bases de datos como esta a través de procesos de minería de datos como pueden ser los árboles de decisión.

Capítulo 2

Métodos estadísticos utilizados

La estructura de este trabajo y los distintos almacenes de datos con los que se trabaja son los que han propiciado la utilización de los diversos métodos estadísticos que se desarrollarán a continuación.

Antes de profundizar en la teoría de estos, se hablará sobre las características generales de los almacenes de datos con los que se han trabajado. Su preprocesamiento se tratará en el siguiente capítulo.

2.1. Estadística Descriptiva

La estadística descriptiva se refiere al conjunto de técnicas encargadas de resumir, sintetizar y explicar de forma breve y general la información que se quiere estudiar.

Tendrán una importancia esencial como fase previa al análisis inferencial de datos ya que serán de gran utilidad puesto que proporcionan gran y valiosa información que permite desarrollar un conocimiento amplio y general de los datos con los que se va a trabajar.

Algunas de las técnicas que se utilizarán en este trabajo serán:

Medidas de tendencia central

Las medidas de tendencia central agrupan a todos los estadísticos encargados de proporcionar toda la información acerca de los valores centrales de la muestra que se dispone, y son representativos del conjunto de esta.

- media: La media aritmética de un conjunto de datos se define como la suma de todos ellos dividida entre el número de datos.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- moda: Se define como moda el valor de la muestra que se repite más veces dentro del mismo. Puede ser unimodal o plurimodal cuando un número de valores distintos se repiten el mismo número de veces.

Medidas de dispersión

Las medidas de dispersión se refieren a una serie de estadísticos de los que se extrae información acerca de la variabilidad de los datos de los que se dispone.

- Rango: El rango de una muestra se define como la distancia que existe entre el valor máximo y el mínimo de esta. Será útil para hacer una primera valoración de la dispersión de la muestra, a mayor rango, mayor dispersión. Aunque siempre se debe complementar con otras medidas puesto que presenta limitaciones como que no tiene en cuenta a todos los elementos de la muestra.

$$R = x_{max} - x_{min}$$

- Cuasivarianza: Se define como la suma de las distancias de cada elemento de la muestra a la media elevadas al cuadrado, dividida entre el número de elementos de la muestra menos uno. Se trata de un estimador insesgado muy apropiado para obtener la dispersión de los datos de una muestra.

$$s_c^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

- Cuasidesviación típica: Se define como la raíz cuadrada positiva de la cuasivarianza. Se define para evitar hablar de unidades elevadas al cuadrado.

Medidas de forma

Las medidas de forma son estadísticos que se refieren a la descripción de la distribución que siguen los datos de los que se dispone sin que sea necesaria su representación gráfica.

Por un lado, está el concepto de asimetría que permite identificar hacia qué lado está desplazada la distribución de los datos y por lo tanto, ver en qué zona hay más concentración de estos mismos. Por otro lado, está el concepto de curtosis o apuntamiento de la distribución que permite hacerse una idea al investigador del grado de concentración de los datos alrededor de su media.

La asimetría se calcula a través de su coeficiente dando como resultado una asimetría positiva, neutral o negativa. La asimetría positiva resultará del sesgo de la función hacia la izquierda siendo así la moda menor que la mediana y esta a su vez, menor que la media; si es neutral, significa que la distribución es simétrica coincidiendo los valores de la moda, la mediana y la media y por último, en el caso de ser negativa, el sesgo de la función estará desplazada a la derecha siendo la moda mayor que la mediana y esta a su vez mayor que la media. Los coeficientes de asimetría no tienen unidades.

Hay diferentes formas de calcularlos el coeficiente de asimetría, el más empleado es el coeficiente de asimetría de Pearson que se define como la media menos la moda dividida entre la desviación típica.

$$AP_1 = \frac{\overline{X} - Mo}{s}$$

.

Por otro lado, la curtosis se calcula a partir de un coeficiente, dando como resultado una distribución leptocúrtica, mesocúrtica o platicúrtica siendo por orden de mención la

primera con un apuntamiento elevado, la segunda, presentando un apuntamiento menos pronunciado y en el tercer caso, con un apuntamiento bajamente pronunciado.

Al igual que pasa con el coeficiente de asimetría, presenta varias expresiones diferentes para su cálculo. Si se utiliza el coeficiente de exceso de Fisher, será el resultado de la división del momento central de orden cuatro dividido entre la desviación típica y todo esto menos tres.

$$g_2 = \frac{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^4}{n}}{s^4} - 3 = \frac{m_4}{s^4} - 3$$

Diagrama de rectángulos

"Es uno de los gráficos más utilizados y puede aplicarse a cualquier tipo de variable" (ALBAJAR y MARTÍN, 2006). Se trata de una gráfica formada por el número de rectángulos correspondientes a cada categoría o valor de la variable de estudio y altura correspondiente a la frecuencia de cada una de las categorías sobre un eje de coordenadas cartesianas.

2.2. Contrastes de hipótesis

Un contraste de hipótesis es un método estadístico perteneciente a la estadística inferencial en la que se tienen dos hipótesis, la primera, denotada por "H_0" se conoce como hipótesis nula, el punto de partida del contraste, que se aceptará siempre y cuando no haya evidencias suficientes para no hacerlo y "H_1" es la hipótesis alternativa a la nula la cual se aceptará en el caso de que haya evidencias significativas que permitan rechazar la otra hipótesis.

Este proceso se lleva a cabo a través de un estadístico de contraste que se calcula a partir de los datos sobre los cuáles se quiere llevar a cabo la prueba, que permitirá aceptar o rechazar la hipótesis nula según el valor que tome, dependiendo de si cae en la región de aceptación formada por todos los valores que permiten aceptar la hipótesis nula o la región de rechazo que refiere el conjunto de valores que hacen que se rechace la "H_0".

Pueden ser unilaterales, cuando solo existe una región de rechazo en la distribución o bilaterales cuando la región de rechazo está dividida en dos partes.

Durante este proceso se pueden cometer errores: el error de Tipo I, que se produce cuando se rechaza la hipótesis nula en el caso de ser cierta, tiene una probabilidad de α denominada nivel de significación. El error Tipo II, cometido cuando se acepta la hipótesis nula cuando esta es falsa con una probabilidad de β, cuya complementaria $1 - \beta$ es la potencia del contraste.

2.2.1. Análisis de la Varianza (ANOVA) de un factor

El ANOVA es un método estadístico para el contraste de hipótesis que se usa cuando se quieren comparar las medias entre más de dos grupos, es decir, cuando se quiere discutir si un número determinado de grupos se pueden considerar iguales o en media.

El Análisis de Varianza está basado en la descomposición de la variabilidad de los datos en dos sumandos, referidos a la variabilidad dentro de los grupos y entre grupos. Se cuenta con dos variables, una cuantitativa también llamada dependiente y otra cualitativa, o factor, que es la que define tantos grupos como categorías tenga (r), siempre que $(r > 2)$

Las hipótesis que se plantean en el ANOVA van a ser:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_n$$

$$H_1 : \mu_i \neq \mu_j, \forall i \neq j$$

La hipótesis nula supone la igualdad en la media de todos los grupos, por otro lado, la hipótesis alternativa, se aceptará cuando haya diferencia en la media de entre al menos dos de los grupos.

El modelo a priori quedaría planteado como

$$y_{ij} = \mu + \alpha_i + u_{ij}$$

Donde α_i corresponde al factor y u_{ij} corresponde al factor aleatorio que interviene en el modelo.

El modelo debe verificar las hipótesis:

- Independencia entre los grupos.
- Varianzas iguales en los grupos.
- Grupos normalmente distribuidos.

La variabilidad se descompone en la suma de la variabilidad explicada, existente entre los tratamientos, más la Variabilidad No Explicada, presente dentro de los tratamientos, lo que da como resultado la variabilidad total.

Se deben calcular, en primer lugar, las sumas de cuadrados para cada una de las fuentes de variabilidad que vienen dadas por las siguientes expresiones para cada una de las fuentes de variabilidad, respectivamente. $SCT = SCE + SCR$, $SCE = \sum_{i=1}^{r} n_i \cdot \hat{\alpha}_i$, $SCR = \sum_{i=1}^{r} \sum_{j=1}^{n_i} e_{ij}^2$, $SCT = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})$ Por otro lado, se dispone de los grados de libertad que se calculan para las fuentes de variabilidad respectivamente como, número de grupos menos uno, $r - 1$, número de elementos menos número de grupos, $N - r$ y número de grupos menos uno $N - 1$ respectivamente.

A continuación, se calculan las medias cuadradas dividiendo las sumas de cuadrados entre los grados de libertad complementarios $MCE = \dfrac{SCE}{r-1}$ y $MCR = \dfrac{SCR}{N-r}$

Por último, se calcula el estadístico de contraste F_{obs} Dividiendo las medias cuadradas

$$F_{obs} = \frac{MCE}{MCR}$$

. Ahora, este resultado se compara con el valor de $F_{1-\alpha}$ que se obtienen al buscar dicho valor a partir de los grados de libertad de la variabilidad explicada y la Variabilidad no Explicada en la tabla de la distribución F de Snedecor.

Si el valor de F_{obs} es mayor que $F_{1-\alpha}$ o si el $p - valor$ es menor de 0.05, se concluye con que el contraste es significativo y, por lo tanto se afirma que para los datos de los que se dispone y un nivel de significación α, existen diferencias entre los grupos.

El siguiente paso sería identificar entre cuáles de los grupos existen diferencias, para esto, se llevarán a cabo los contrastes a posteriori, que pueden ser el contraste de LSD, Tukey, Bonferroni o Dunett, según las necesidades.

2.3. Gráficos de control

Los gráficos de control son herramientas estadísticas que se utilizan para el análisis de datos cualitativos o cuantitativos y tienen en cuenta el orden temporal de los acontecimientos. Se utilizan para definir una meta de cara a una operación que se desee realizar, para ayudar a conseguirla y fundamentalmente para determinar si dicha meta se ha alcanzado o no.

Ayudan a determinar si el proceso de estudio, a lo largo del tiempo se mantiene estable dentro de unos límites de control. En el caso de la evolución de los fenómenos criminalísticos, no se puede parar el proceso, pero sí sirven para tomar medidas adecuadas para corregir los valores fuera de control aplicando una serie de medidas cuando estos se salen de los límites.

Los gráficos de control, en sí, no controlan. Ayudan a supervisar un proceso permitiendo y orientando a la toma de decisiones del investigador cuando estas resulten pertinentes para mantenerlo en estado de control o estable, haciendo modificaciones en la variabilidad asignable del mismo.

Los gráficos de control que se utilizarán en el presente trabajo son los gráficos de Shewart que a su vez son los más comunes y conocidos. Estos gráficos se caracterizan por representar en el eje de abscisas una unidad temporal o alguna magnitud relacionada con el tiempo y en el eje de ordenadas el fenómeno que se está estudiando. Los gráficos de control se componen de tres líneas. La línea central o "LC" corresponde al valor medio de todas las observaciones de las que se dispone y los límites de control, Superior e Inferior que se situarán respectivamente $\pm 3\sigma$ de la línea central, establecen los límites a partir de los cuáles no pueden sobrepasar los valores de las observaciones. Si esto ocurre, esos valores estarán fuera de control. En ciertos casos de estudio de fenómenos criminalísticos resulta útil establecer, los límites de alerta que se encuentran a $\pm 2\sigma$ de la línea central. Estos, servirán para que cuando un valor los sobrepase empezar a tomar las medidas que se consideren para impedir que los siguientes sobrepasen los de control o para estar preparado en caso de que se descontrolen y acabe ocurriendo lo anterior.

Para un buen proceso de construcción del gráfico de control, en primer lugar, se debe identificar el fenómeno criminalístico que se quiere estudiar; en segundo lugar, se debe de seleccionar el número de datos con los que se quiere trabajar, a continuación se procede a establecer los límites de control. Una vez se han identificado, se procede a la representación de los valores sobre el diagrama y, por último, si es necesario, se toman las medidas que se consideren pertinentes en vistas a los resultados que arroje el Gráfico.

En este trabajo, se hará uso de dos clases principales. El gráfico de control por número de Casos y el gráfico de control de proporciones.

2.3.1. Gráfico p

Es un gráfico utilizado para los casos defectuosos, se tiene que p se estima a través del cociente de los casos favorables entre el número total de casos, es lo mismo que decir:

$$\hat{p} = \frac{\text{nº de casos favorables total}}{\text{nº de muestras} \cdot \text{tamaño muestral}} \tag{2.1}$$

Esta proporción, estimada a través de la proporción muestral tiene una esperanza con valor p $E(\hat{p}) = p$ y una desviación típica con valor $\hat{\sigma} = \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$

Por lo que la línea central se define como $LC = \hat{p}$ y los límites de control Superior e Inferior $LSC = p+3\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$, $LIC = p-3\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$ y los límites de alerta $LSA = p+2\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$, $LIA = p - 2\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$.

El cálculo de estos parámetros depende directamente del tamaño de la muestra que se disponga, por lo tanto, si el tamaño de muestra es el mismo para todas las muestras, se mantendrán fijos.

En el caso de que las muestras tengan un tamaño variable, se va a trabajar en este caso con dos casuísticas. La primera de ellas se tiene cuando los tamaños de las muestras tienen una variabilidad pequeña, por lo que se optará por establecer un valor de n promedio de todos los tamaños de muestra de los que se dispone. Por lo tanto, para este caso, habrá que calcular el promedio de las n_i y quedarían los parámetros definidos del siguiente modo: $\bar{n} = \frac{\sum_{i=1}^{m} n_i}{m} \quad \forall n_i, i = 1, 2, \cdots, m$, $LSA = p + 2\sqrt{\frac{\hat{p} \cdot \hat{q}}{\bar{n}}}$, $LIA = p - 2\sqrt{\frac{\hat{p} \cdot \hat{q}}{\bar{n}}}$, $LSC = p + 3\sqrt{\frac{\hat{p} \cdot \hat{q}}{\bar{n}}}$, $LIC = p - 3\sqrt{\frac{\hat{p} \cdot \hat{q}}{\bar{n}}}$, $LC = p = \frac{\sum_{i=1}^{m} d_i}{\sum_{i=1}^{m} n_i} \quad \forall n_i, i = 1, 2, \cdots, m$

Por el contrario, en el caso de que las muestras tengan un tamaño variable, pero la variabilidad entre las mismas sea considerable, entonces el proceso anterior, deja de ser eficaz y por ello, se utiliza el gráfico estandarizado. En este caso, cada uno de los valores se debe de estandarizar y tiene la peculiaridad de que la línea central se encuentra de forma fija en 0 los límites de control se situarán de forma fija a ± 3 y los límites de alerta a ± 2. Los parámetros se calculan del siguiente modo: $Z_i = \frac{\hat{p_i} - p}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n_i}}}$, $LSA = 2$, $LIA = -2$, $LSC = 3$, $LIC = -3$, $LC = 0$.

2.3.2. Gráfico de control para número de defectos

Cuando lo que se quiere tener en cuenta es el número de casos sobre el fenómeno que se quiera estudiar, se hace uso de este tipo de gráfico. Este gráfico puede utilizarse siempre y cuando se tengan los datos en la misma unidad de medida y contabiliza el número total de casos sobre el fenómeno que se quiere estudiar en la muestra que se dispone. Para aclarar esto último, se expone el siguiente ejemplo: Se quiere estudiar el número de abusos sexuales a menores al año durante los últimos diez años.

Se supone que la distribución que siguen el número de casos sigue una distribución de Poisson con parámetro c, pues la probabilidad de que haya un defecto es $P(c)$. Esta distribución tiene como característica principal que el valor de la esperanza es el mismo que el de la varianza y es c. En este caso, se supone que la probabilidad de que ocurra el fenómeno criminalístico en este caso es muy pequeña. Por ejemplo, la probabilidad de que una persona cometa un asesinato es muy pequeña.

Así pues, para la construcción del Gráfico c, se tendrán los siguientes límites:

La línea central, corresponderá a la media, que es $\hat{c}$, los límites de control, Superior e Inferior serán la media más menos tres veces la desviación típica que al ser la varianza igual a la media, será la raíz de $\hat{c}$ y los límites de alerta que se construirán igual que los anteriores, la única diferencia, es que están multiplicados dos veces por la desviación típica.
$LSA = \hat{c} + 2\sqrt{\hat{c}}$, $LIA = \hat{c} - 2\sqrt{\hat{c}}$, $LSC = \hat{c} + 3\sqrt{c}$, $LIC = \hat{c} - 3\sqrt{\hat{c}}$, $LC = \hat{c}$.

2.4. Números Índices

Cuando se quiere estudiar la evolución y el comportamiento de un fenómeno, normalmente, social, o económico a lo largo del tiempo y hacer comparaciones evolutivas entre intersecciones temporales determinadas, se utilizan una serie de herramientas estadísticas muy útiles que permiten analizar el comportamiento del fenómeno de estudio deseado., En este caso criminalístico, permitirá determinar el comportamiento de la variable en el tiempo y, asimismo, evaluar, si ciertos factores que se han considerado de interés han propiciado un aumento o una disminución en el resultado del fenómeno que se propone para estudio. Estos fenómenos, se caracterizan por ser de alta complejidad lo que se consigue a través del empleo de esta herramienta estadística.

La particularidad de los números índices es que las observaciones que se comparan siempre se llevarán a cabo tomando como referencia una de estas mismas como punto fijo de partida, es la primera observación de la que se disponga, siendo así, lo más común considerar la más antigua en el tiempo. También se puede utilizar un punto a lo largo del conjunto temporal de datos que se disponga y se considere como acontecimiento de interés, o también se puede establecer la referencia de forma artificial cuando se considere oportuno, por ejemplo a través de métodos de promediación. Además se puede hacer referencia al espacio en vez de al tiempo

Una definición de números índices se podría establecer como "medida estadística abstracta (sin unidad), diseñada para mostrar los cambios de una variable o grupo de variables con respecto al tiempo, situación geográfica u otra característica" (LÓPEZ MARTÍN, 1982). Al período inicial se le denomina período base o de referencia representando siempre el valor del 1 o 100 % si se contabiliza en tanto por ciento y al período que se desea comparar se le denomina como período actual.

Se trabajará con dos clases de números índices: los números índice simples y los números índice compuestos:

2.4.1. Números Índices Simples

Se utilizan cuando se quiere estudiar la evolución de una única variable a lo largo del tiempo o espacio tomándose como referencia un valor de la misma. Cada uno de los

índices refleja la variación que ha sufrido la variable con respecto a don intersecciones temporales. Se denota por: $I_i = \frac{x_i}{x_0}$, es decir, el cociente del valor de la variable x en el momento "actual" de medición (x_i) con respecto al valor en el momento de referencia (x_0). El resultado es un valor que no tiene unidades de medida, si el valor actual está por encima, por debajo o es igual que el índice de referencia que siempre tendrá el valor de 1, es decir, si $I_i > 1 o 100\,\%$, x_i es mayor que x_0, si $I_i < 1 o 100\,\%$, x_i es menor que x_0 y si $I_i = 1 o 100\,\%$, x_i es igual que x_0

En este trabajo también será frecuente hacer uso de las variaciones relativas, que no será nada más que el cociente de la resta de la medición en el período actual menos la medición de la variable en el período de referencia, dividido todo entre el valor de la variable que toma en el período de referencia:

$$\Delta I_i = \frac{x_i - x_0}{x_0}(\cdot 100)$$

Este valor permite, no solo saber que x_i es mayor o menor que x_0, sino en qué medida es esa variación:

1. Existencia: El valor del índice estará definido en el conjunto de los números reales positivos $(\mathbb{R}^+)$, es decir,
$$I_i \in \mathbb{R}^+$$

2. Identidad: En el el caso de que el período de referencia tenga el mismo valor que el período actual, el valor del índice será de 1 o del 100 % si se expresa en forma de porcentaje. Este mismo resultado siempre corresponderá al número índice de partida con el valor que se tome como referencia.
$$I_t^t = 1$$

3. Inversión: Cuando para la obtención de un número índice el período de referencia, pasa a ocupar el lugar del período actual y viceversa, es decir, cuando se intercambian entre sí los períodos se cumple que:
$$I_0^t \cdot I_t^0 = 1$$

4. Circular: Si se multiplica el índice I con período de referencia 0 y actual t por otro índice I con período de referencia t y el período actual t', el resultado de la multiplicación será un nuevo número índice con período de referencia t' y período actual 0.
$$I_0^t \cdot I_t^{t'} = I_0^{t'}$$
Siendo fácilmente demostrable a través del siguiente razonamiento:
$$\begin{aligned} I_0^t = \frac{b}{a} : \quad I_t^{t'} = \frac{c}{b} \\ I_0^t \cdot I_t^{t'} = \frac{b}{a} \cdot \frac{c}{b} = \frac{c}{a} = I_0^{t'} \end{aligned} \tag{2.2}$$
Y extendiendo al caso completo se verifica que:
$$I_0^t \cdot I_t^{t'} \cdot I_{t'}^0 = 1$$
Quedaría demostrado de la misma forma que anteriormente:
$$\begin{aligned} I_0^t = \frac{b}{a} : \quad I_t^{t'} = \frac{c}{b} : \quad I_{t'}^0 = \frac{a}{c} \\ I_0^t \cdot I_t^{t'} \cdot I_{t'}^0 = \frac{b}{a} \cdot \frac{c}{b} \cdot \frac{a}{c} = 1 \end{aligned} \tag{2.3}$$

5. Proporcionalidad: Si el período actual del índice queda multiplicado por una cantidad a la que se denota como k, el índice quedará multiplicado por la misma cantidad.

$$I_i' = \frac{k \cdot b}{a} = k \cdot I_i$$

6. Homogeneidad: El valor del índice no se verá afectado por los cambios en las unidades de medida.

2.4.2. Numeros Índices Compuestos

En el caso de que se necesite la intervención de distintas magnitudes para estudiar un fenómeno criminalístico, se requiere que entren en juego los números índice compuestos, no son más que el fruto de la media de distintos números índice simples que hacen referencia a magnitudes diferentes dentro de una misma expresión matemática, que tiene como objetivo, devolver otro número índice donde se recoja la mayor cantidad de información posible, concentrada a partir de los distintos índices simples. Una de las tareas que requiere una mayor delicadeza como paso previo a su cálculo es determinar qué magnitudes se quiere que entren en juego haciendo un estudio exhaustivo y analizando la coherencia de agrupar el conjunto de índices que hacen referencia a las magnitudes candidatas a ser consideradas.

Se distinguen dos clases de números índices compuestos: los Números Índices Compuestos No Ponderados y los Números Índices Compuestos Ponderados. Ambos serán empleadas en el desarrollo del trabajo.

Números Índices Compuestos No Ponderados

Los Números Índices Compuestos No Ponderados son los que se caracterizan por la no asignación de ningún peso o ponderación, a las magnitudes tenidas en cuenta para el cálculo de este. Se considera que la importancia de todas ellas es igual ya sea porque la subjetividad impide cuantificar cuán diferente es la importancia entre dos magnitudes o bien, la objetividad permite afirmar que una magnitud no supera a otra en intensidad de importancia. Si se quiere tener en cuenta dos magnitudes, referidas a ámbitos en la escena de la criminalidad como pueden ser los delitos de xenofobia o por ideología, desde un punto de vista objetivo resulta imposible determinar cuál de los dos tiene una mayor importancia. En cambio, si se pretende estudiar los asesinatos y los robos en una determinada ciudad, esos dos delitos, no se pueden poner al mismo nivel de importancia.

El caso más sencillo es el de considerar el Índice Compuesto No Ponderado como la media aritmética de los distintas índices simples de los que se dispone. Este índice es conocido como Índice de Sauerbeck, y como se acaba de señalar, se define como la suma de todos los índices simples, dividido entre el número de estos.

$$I_{Sbk} = \frac{I_1 + I_2 + ... + I_N}{N}(\cdot 100) = \frac{1}{N}\sum_{i=1}^{N} I_i(\cdot 100)$$

O si se expresa en forma de incremento, quedaría de la siguiente forma:

$$\Delta I_{Sbk} = \frac{\Delta I_1 + \Delta I_2 + ... + \Delta I_N}{N}(\cdot 100) = \frac{1}{N}\sum_{i=1}^{N} \Delta I_i(\cdot 100)$$

Números Índices Compuestos Ponderados

Como es de esperar, estos índices se definen de forma opuesta a los anteriores, es decir, los Números Índice Compuestos Ponderados son los que se caracterizan por la asignación de un peso a cada una de las magnitudes, denotado por w_i y llamados coeficientes de ponderación. Entrarán en juego a la hora del cálculo de este índice, al considerar, que la importancia para al menos una de las magnitudes no es la misma que para las demás, cuando la objetividad y las evidencias permitan hacer distinción sobre la diferencia de importancia de las mismas y no queden dudas sobre una posible igualdad en el peso de cada una de las magnitudes. Corresponde al segundo ejemplo de la comparación explicativa que se estableció para el punto anterior.

En este caso, si el número Índice Compuesto No Ponderado se definía como la media aritmética de los números índice simples correspondientes a cada una de las magnitudes, el Índice Compuesto Ponderado. se definirá como la media aritmética ponderada de los Índices Simples correspondientes a las magnitudes que se tengan en cuenta. Para su cálculo, basta con introducir en la fórmula expuesta anteriormente para los No Ponderados, los coeficientes de ponderación.

$$I_{ICP} = \frac{I_1 \cdot w_1 + I_2 \cdot w_2 + ... + I_N \cdot w_N}{w_1 + w_2 + ... + w_N}(\cdot 100) = \sum_{i=1}^{N} \frac{I_i \cdot w_i}{w_i}(\cdot 100)$$

Si se expresa en forma de los incrementados, quedaría del siguiente modo:

$$\Delta I_{ICP} = \frac{\Delta I_1 \cdot w_1 + \Delta I_2 \cdot w_2 + ... + \Delta I_N \cdot w_N}{w_1 + w_2 + ... + w_N}(\cdot 100) = \sum_{i=1}^{N} \frac{\Delta I_i \cdot w_i}{w_i}(\cdot 100)$$

2.5. Técnicas de Minería de Datos

La minería de datos es una rama de la estadística que está formada por un conjunto de técnicas multifuncionales que se utilizan para encontrar patrones, relaciones, comportamientos o realizar predicciones directamente desde los datos que no se pueden extraer de ellos a simple vista y que permitirán sacar conclusiones de interés para el investigador. Estos algoritmos se pueden clasificar en supervisados, cuando su objetivo es hacer una predicción a partir de los datos de los que se dispone y los algoritmos no supervisados, cuyo objetivo es descubrir relaciones y patrones dentro del conjunto de datos del que se dispone. Cada una de las técnicas es compuesta por un conjunto de algoritmos. Las técnicas más representativas son:

- Técnicas de agrupación.
- Técnicas de asociación.
- Técnicas de *clústering*.
- Árboles de decisión.
- Redes neuronales.
- Modelos de predicción numérica.

En este trabajo se utilizará el árbol de decisión J48, técnicas de asociación a través del algoritmo “a priori” y el algoritmo de *clustering* K-medias.

2.5.1. Árbol de decisión C4.5 (J48 en Weka)

Este algoritmo genera el árbol de decisión utilizando secciones de los datos que va generando progresivamente, utilizando la entropía, es decir, una medida de desorden de los datos para decidir la creación de los nodos. Cuanto menor sea el nivel de entropía de los atributos, mayor será la probabilidad de que el algoritmo genere un nodo a partir de este atributo (mayor ganancia normalizada). Para cada nodo, de lo que tratará el algoritmo es de escoger el atributo que divida de una manera más efectiva los datos basándose en el principio que se señaló anteriormente.

Se podría decir que la forma de proceder del algoritmo es la siguiente:

Se parte de los datos iniciales, a continuación, para cada atributo, el algoritmo compara la ganancia de información normalizada de la división del atributo y selecciona el atributo con la mayor ganancia normalizada. Genera un nodo de decisión a partir de la sección del atributo con mayor ganancia normalizada. Una vez se llega a este punto se repite el mismo procedimiento en los subconjuntos resultantes con las respectivas generaciones de nodos hijos que pueden ser a su vez nuevos padres o terminales si no se puede repetir este proceso.

En primer lugar, la entropía de $\vec{y}$ se define como

$$Entropía(\vec{y}) = -\sum_{i=1}^{n} \frac{|y_i|}{|\vec{y}|} log \frac{|y_i|}{|\vec{y}|}$$

Ahora, si se itera sobre todos los valores de $\vec{y}$, la entropía de i condicionada a $\vec{y}$ es:

$$Entropía(i|\vec{y}) = \frac{|y_i|}{|\vec{y}|} log \frac{|y_i|}{|\vec{y}|}$$

Por último, la ganancia quedará definida como:

$$Ganancia(\vec{y}|i) = Entropía(\vec{y} - Entropía(i|\vec{y})$$

El objetivo del algoritmo será maximizar la ganancia.

$$\text{Max: } Ganancia(\vec{y}|i)$$

2.5.2. Algoritmo a priori

Se trata de un algoritmo utilizado en minería de datos cuyo propósito es extraer reglas de asociación y patrones de comportamiento desde un conjunto de datos a partir de la búsqueda de conjuntos de observaciones frecuentes. Este algoritmo es el que permitió a las cadenas de supermercados realizar estudios para encontrar asociaciones entre artículos y, así, poder organizar las tiendas de tal forma que se agruparan en los mismos rincones artículos que se suelen comprar juntos y orientar también las campañas de *marketing*.

Este algoritmo va a tratar de medir el nivel de dependencia entre ciertos conjuntos de datos, es decir, tendrá en cuenta las probabilidades condicionadas de que una observación perteneciente a un grupo sea extraída dado que otra observación perteneciente a otro

grupo ya haya sido seleccionada.

En resumen, el algoritmo trata de encontrar reglas de asociación para determinar que, si se extrae una observación de un grupo, entonces, probablemente, también se va a extraer otra observación de otro grupo. Las reglas de asociación obtenidas se evaluarán a través de la confianza que se define como el coeficiente que está referido al número de casos que predice dicha regla correctamente y el soporte, que está definido como el coeficiente que indica el número de registros que están afectados por dicha regla de asociación.

Al umbral mínimo de soporte se le denotará por U y un conjunto X será frecuente, si el soporte de X es mayor o igual que U,

$$sop(X) \geq U$$

Por lo tanto, si un conjunto X es frecuente, todos sus subconjuntos lo serán y viceversa, si un conjunto X no es frecuente, todos sus superconjuntos, tampoco lo serán. Esto se define como el principio fundamental del algoritmo a priori.

Entonces, el algoritmo procede del siguiente modo:

Se denota por C_1 el conjunto de elementos de partida y $L_1 \subseteq C_1$ el conjunto de elementos frecuentes lo que quiere decir que $L_1 = \{i_k : sop(i_k \geq U, k = 1, 2, 3, ...\}$.

Ahora se proponen todos los conjuntos C_k con k elementos que surgen de la combinación de los elementos de L_{k-1}.

Se eliminan de C_k los elementos no frecuentes de L_k.

Por último, el algoritmo termina, cuando todos los conjuntos C_k no superan el umbral U.

2.5.3. Algoritmo k-medias

El algoritmo k-medias es un algoritmo no supervisado de *clustering* que se encarga de agrupar los datos de los que se dispone en una serie de grupos con el propósito de maximizar las diferencias entre los grupos y minimizarla dentro de ellos, por lo que la variabilidad va a jugar un papel muy grande dentro del mismo.

El algoritmo procede del siguiente modo:

En primer lugar, se seleccionan el número de *cluster* con los que se quiere que trabaje el algoritmo de forma aleatoria o a través de un criterio. El más utilizado es el criterio del codo en el que se elegirá número óptimo el punto en el que la tendencia del codo cambie, lo que implica que la variabilidad que queda a partir de ese punto puede resultar "despreciable" ya que es pequeña.

A continuación, procede a la selección de los centroides de los grupos, que se definen como los puntos en los que se encontrará el centro de cada *cluster*.

Se van recorriendo todos los datos calculando la distancia desde cada uno de los puntos al centroide para, así, asignar los puntos a los *cluster* más cercanos. Esta operación se

realiza a través de la distancia euclídea

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}$$

y evaluará cómo de buenos son los grupos a través de la función

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

a través de la cual se comprueba si el algoritmo debe parar o debe seguir funcionando.

Se recalculan los centroides y se repiten los dos pasos anteriores hasta que los resultados converjan al resultado que se espera y se finaliza el algoritmo.

Capítulo 3

Diseño de herramientas

Para el análisis de los datos de los que se dispone y para futuros análisis de datos, en el caso concreto de los gráficos de control, se ha diseñado una librería específica en R para tener una herramienta alternativa a Excel y favorecer de este modo una interpretación rápida y eficaz de los mismos.

En el caso de los números índices, se han desarrollado funciones y pequeños programas en R más específicos para los datos de los que se dispone ya que están orientados casi exclusivamente a las características que tienen los que se están manejando en esta base, aunque también serían extensibles al análisis de los que tuvieran una disposición y propiedades similares a con los que se está trabajando.

3.1. Librería CCCSA para análisis de gráficos de control

El objetivo de la creación de esta librería CCCSA, (Control Charts for Criminality Statistical Analysis), es que se pueda disponer del conjunto de herramientas necesarias que permitan estudiar e interpretar los gráficos de control en el campo de la criminalística. Esta se ha basado en el diseño e implementación de un total de cuatro funciones R que se encargan de realizar las gráficas de control más utilizadas en el campo del análisis estadístico de criminalidad. Estas funciones se pueden dividir en dos grupos. De acuerdo con el primero, la función que lo compone corresponde al gráfico de control para el número de defectos. El segundo grupo contiene un total de tres funciones que componen las gráficas de control para proporciones de defectuosos; en primer lugar, en el caso de que las muestras tienen el mismo tamaño; en segundo lugar, para cuando las muestras tienen diferentes tamaños, pero tienen poca variabilidad, y, por último, la gráfica para valores estandarizados, utilizados, cuando la variabilidad del tamaño muestral es mayor que el caso anterior.

Para el diseño en el código para el funcionamiento de esta librería se han empleado a su vez las librerías **dplyr** y **ggplot2**.

- **dplyr**: Es una librería de R que se encarga de la manipulación de datos. Es decir, perimite realizar operaciones múltiples de extracción, adición, filtro,..., entre otras, así se puede desarrollar una sencilla operación, directa y cómoda sobre los datos de los que se dispone.

- **ggplot2**: Es una librería que se encarga de ayudar a generar gráficos, representando

los datos que correspondan de una manera intuitiva, moderna y con mayor flexibilidad a la hora de configurar la estética y variables del mismo acorde a lo que se necesite en cada momento, con una personalización más específica que la que ofrecen los gráficos básicos de R.

Según el primer bloque y a la función relacionada con el **gráfico de control para el Número de Defectos**, se define la función **gcc** que se compone de los parámetros df, frq, temp, values y time. El primero se corresponde con la base de datos con la que se va a trabajar (df = data frame), "frq" se refiere a la frecuencia temporal , si es en años, será de 1, si es en cuatrimestres será 3, en trimestres 4, en meses 12 y así sucesivamente, "temp" se refiere al número de unidades totales temporales no fragmentadas, si es en años, será la suma de los años totales de los que se disponga. El parámetro values se corresponde con la variable de criminalidad que contienen los datos de interés con los que se va a trabajar, y "time", la variable que contiene a las unidades temporales. Una vez se han definido dichos parámetros, la función trabaja del siguiente modo:

En primer lugar, calcula el valor de la línea central, es decir, la suma del número de casos dividido entre el número total de períodos de los que se dispone (frq $\cdot$ temp). A continuación, el valor de los límites superior e inferior se calcula sumando y restando al valor de la línea central para cada uno de los respectivos Límites correspondientes y multiplicándolo por la raíz cuadrada del valor de la línea central. Después, pasa a calcular los límites de alerta de la misma forma que los anteriores, pero cambiando el valor del 3 por el 2 como ya se explicó en la parte teórica.

Para la representación de todos los componentes del gráfico de control, se programarán las instrucciones apoyándose en la librería de gráficos. Se van representando una a una estas componentes. En primer lugar, se programará una función que represente la línea central de color amarillo con la etiqueta "LC", dado en lenguaje ASCII. Las dos siguientes funciones, tendrán como objetivo programar las funciones que representen los límites de control, de forma análoga a la línea central, difiriendo en color, que esta vez será en rojo, con sus respectivas etiquetas "LSC" y "LIC". Como penúltimo paso, se repite el proceso anterior para los límites de alerta que tendrán un color naranja y con etiquetas "LSA" y "LIA", respectivamente. Como penúltima función dentro de este marco, se encuentra la representación de los datos correspondientes a la variable criminalística de estudio que se representará de color azul. En último lugar, se definen las funciones de forma que permiten dar al gráfico la personalización final que se considera como el título, el diseño de las líneas o del fondo. Para la construcción de estas funciones se ha apoyado en la librería **ggplot2** que permite una representación muy moderna, visual e intuitiva de los gráficos en R, los códigos de dichas funciones se podrán consultar en los anexos.

Con relación al segundo bloque, que corresponde al gráfico P, para proporciones, como se introdujo en este punto, se constituye a partir de tres funciones:

La primera se define como **gcp** que corresponde con los **gráficos de control para Proporciones cuando el número de elementos de los grupos es el mismo**; consta de cuatro parámetros: "df", "auspicious", "time", "size". El primero será común en todas ellas, ocurre lo mismo con time, auspicious se refiere al número de casos favorables y "size" hace referencia al tamaño de las muestras que será un valor numérico, ya que siempre en este caso, será el mismo valor para todas ellas. Una vez definidos los parámetros que utiliza, se podrá describir como trabaja esta función:

El mecanismo será prácticamente el mismo en todas ellas, aunque estableciendo las diferencias particulares dentro de cada una de ellas. En este caso, lo primero que se define es la variable p, que no será más que lo que corresponderá a la línea central. Se define como la suma de "auspicious" dividido entre el tamaño total de las muestras. A continuación, se definirá la variable "pi" que corresponde al valor de la proporción de la variable criminalística a lo largo del tiempo del que se dispone. Será definida como el valor de los casos favorables (auspicious) entre el tamaño de la muestra que será siempre el mismo. A continuación se procede a definir las funciones para los cálculos de los límites de control tanto superior como inferior de forma que a p se le suma o resta dependiendo del límite 3 multiplicado por la raíz cuadrada de $(\frac{p \cdot (1-p)}{\text{size}})$. Los límites de alerta se formulan análogamente a los anteriores, con el único matiz de que la raíz en estos está multiplicado por 2. A continuación en lo que se refiere a la parte de código que permite la representación gráfica en la salida que devuelve la misma, es el mismo proceso que en el anterior para la función **gcc**, con las mismas configuraciones. Para finalmente obtener la salida de la misma con la representación del gráfico de control.

La segunda función se define como **gcpn** que se utilizará en el caso de que se quiera analizar **gráficos de control para Proporciones cuando el tamaño de los grupos es diferente** y que se escogerá cuando los tamaños de muestra presentan poca variabilidad. La función está definida a través de cuatro parámetros: "df" y "time" que como ya se ha señalado serán comunes en todas las funciones, "auspicious" que también se utilizará en todas las funciones referidas a los gráficos de control p y, por último, "sizes", que, como se puede intuir, en este caso, se referirá a un vector de valores que corresponderán a los tamaños de las muestras de las que se dispone para cada valor de la variable criminalística de interés. La función procederá del siguiente modo:

En primer lugar, la función calcula el parámetro "*_n*" que será la media de todos los tamaños de muestra ("sizes") de los que se dispone. Ahora, como en la función **gcp** se procede a calcular el parámetro *p* que será la suma de "auspicious" dividido entre la suma de "sizes" desde donde se obtendrá el valor de p que corresponderá al valor de la línea central. A continuación, se calcula el vector de proporciones, dividiendo la variable que corresponde a "auspicious" entre la variable de tamaños de muestra, correspondiente a "sizes". Una vez se han calculado estos parámetros se procede al cálculo de los límites de control, sumando o restando respectivamente al valor de *p*, tres, multiplicado por la raíz cuadrada de (*p* por *1-p*), dividido entre *_n*. Para los límites de alerta, se repite el proceso, multiplicando las raices por 2 en este caso. La última parte de esta función es análoga a las anteriores y es la que se encarga de representar los parámetros que se han calculado en la gráfica a través de la librería **ggplot2** como se ha hecho en los casos anteriores, siguiendo la misma configuración. De esta manera, la función podrá devolver el gráfico de control correspondiente a los parámetros que se han tenido en cuenta.

La última función de este segundo grupo, está definida como **gcpz**, estará orientada a cuando se quieran analizar **gráficos de control para proporciones cuando el tamaño de los grupos es diferente** y cuando los tamaños de la muestra presenten una variabilidad considerable. La función se define a través de cuatro parámetros, que son los mismos que en la anterior. En este caso lo que cambiará es el proceso a partir de los datos de los que se parten, aunque compartan forman. Como se ha comentado en la parte teórica, en este caso, se llevará a cabo de una representación de los datos estandarizados. por lo tanto el funcionamiento constará del siguiente proceso:

La función empieza calculando la variable *pi* que no es más que la misma que antes, es decir, "auspicious" entre "sizes", en definitiva, el vector de proporciones. Una vez dispone la función de este vector, se calcula el valor de p de forma análoga a las anteriores. Después de que que la función dispone del vector de proporciones y del valor de p, el paso fundamental el calcular el vector de valores estandarizados al que se llamará *zi* que se calculará a través del vector de proporciones y del valor de p de la forma *(pi-p)* entre la raíz cuadrada de *(p*(1-p))/(sizes)*. Ahora, los límites de control se indicarán directamente en el apartado de programación del gráfico a través de las instrucciones utilizadas desde **ggplot2**, ya que los valores de estos límites en este tipo de gráfico siempre se mantendrán constantes, siendo -3 y 3 respectivamente; por lo tanto, lo único que habrá que hacer en la instrucción para que dibuje de la misma manera que en las anteriores los límites de control, superior e inferior asignar el valor de 3 y -3 respectivamente en vez de llamar a la función que se encarga de calcular los valores para estos mismos, como se ha hecho en los casos anteriores. Se procede del mismo modo en el caso de los límites de alerta, con los valores 2 y -2.

Cuando se necesite corregir alguno de los valores del gráfico de control debido a que se encuentren fuera de control, es decir, por encima o por debajo de los límites de control se retirarán accediendo a la base de datos que los contiene, eliminando todos los datos correspondientes a esa entrada y definiendo de nuevo los datos actualizados para que cuando se vuelvan a introducir los parámetros dentro de las funciones repitan el cálculo, pero con los datos actualizados. Este proceso se repetirá las veces que se precise hasta que finalmente, se lleve el gráfico a estado de control.

Todo sobre los códigos de las funciones y demás información que se corresponde con esta librería, se podrá consultar en el Anexo correspondiente.

3.2. Diseño de código en R para Números Índices

Para el estudio de los números índices, se ha requerido del uso de las librerías que conforman **tidyverse**, cuyo fin es agrupar todas las librerías necesarias para el análisis de datos, entre los que se encuentran **ggplot2**, **dplyr**, **tidyr**, **readr**, entre otras, que tienen como objetivo la manipulación y representación de los datos con los que se quiere trabajar.

En primer lugar, se hará las funciones para los **números índice simples**, que se programarán para cada uno de los casos que se precise. Se construirán según la Comunidad Autónoma y el ámbito de criminalidad. Los ámbitos de criminalidad se agruparán todos en el mismo gráfico. El proceso de construcción tendrá la siguiente estructura:

Primero se define un vector con los años en los que se han tomado los datos para calcular los números índices simples que se van a comparar (*year*). En segundo lugar, se define otro vector que extrae desde la base de datos con la que se trabaja, la variable de criminalidad que se desea estudiar, indicando además la comunidad autónoma deseada y el ámbito de criminalidad que se desee. Este proceso se repetirá para cada uno de los ámbitos de criminalidad ya que el objetivo es el cálculo y la representación de los números índices de todos ellos en un mismo diagrama para una comunidad autónoma determinada. El siguiente paso es agrupar todas las variables que se han calculado anteriormente, pertenecientes a cada uno de los ámbitos, para este fin, se utiliza la función *data.frame*

que permite crear una nueva base de datos con las variables que se han calculado.

El último paso será el cálculo y la representación de los números índice simples por año y ámbito en el gráfico conjunto, para esto mismo, se hará uso de la programación del gráfico a través de las funciones que ofrece el paquete **ggplot2**. En líneas generales, se van a calcular cada uno de los números índices en cada una de las variables para los ámbitos de criminalidad que se programaron con anterioridad. Para ello, a través de la función *ggplot*, indicando la base de datos con la que se quiere trabajar, siendo la que se ha creado a partir de la original y la variable del eje de abscisas (*year*), se procede a construir las líneas que corresponderán a los valores de los números índices a través de las funciones *geom_line*, donde el valor para el eje de ordenadas se calculará como el valor que toma la variable de interés criminalístico, para cada ámbito en cada uno de los años y en la comunidad autónoma que se desee, dividido entre la media de todos los valores para la misma categoría de los que se dispone y, por último, se multiplica por cien; es decir, el período de referencia se toma como el promedio de todos los períodos "actuales". Los delitos de odio por razón de xenofobia quedan asignados con el color rojo, los correspondientes por identidad de género quedan asignados con el naranja, a los delitos por razón de ideología se les asigna el azul, a los delitos de odio por razón de sexo se les establece con el color verde y a los referidos a creencias y prácticas religiosas, se les asigna el color negro.

Para el caso de los números índice compuestos, se diferencian las funciones para números índices compuestos sin ponderar y números índice compuestos ponderados.

En el caso del código para los textbfnúmeros índices compuestos sin ponderar, está orientado al estudio de los números índice, sin atender a la densidad de población. Se toman números índices simples con respecto a las comunidades autónomas, considerando la tipología de los delitos igual, por lo que se prescinde de la ponderación. En primer lugar, la función específica que se va a construir, consta de cinco parámetros: *df, time, CA, CV, y* que corresponden con la base de datos, el vector de tiempo, el vector de comunidades autónomas, el valor para la comunidad autónoma con la que se quiere operar y la variable de criminalidad de interés. Una vez se definen los parámetros, la función trabaja del siguiente modo: el primer bloque de variables se encarga de la extracción de los elementos correspondientes al ámbito de los delitos de odio en cada uno de los años que se pretenden estudiar en la comunidad autónoma que se precise. Una vez se tiene este bloque de funciones, el siguiente bloque corresponderá al cálculo de un vector que contenga los incrementos de los números índices simples para cada uno de los ámbitos de criminalidad en las dos combinaciones de años con las que se decide trabajar. El tercer bloque es el encargado de calcular los números índices compuestos que se quieren estudiar a través de la suma de los incrementos de los números índice simples dividido entre el número de estos mismos. Por último, la función irá imprimiendo uno a uno los resultados calculados en cada una de las variables que se han descrito que pertenecen a los bloques anteriores.

En el caso del código, para programar la función que obtiene los **números índice compuestos ponderados**, teniendo en cuenta los parámetros relativos a la población y densidad de población de las comunidades autónomas. En primer lugar, se genera la variable encargada de extraer los elementos de la base de datos del año 2014 a la que se denotará como "nia14". Una vez hecho esto, se construye la función "nni14" que se compone de la concatenación de tres instrucciones progresivas. Primero, selecciona el subconjunto de datos que se necesita, en este caso, "nia14", que se calculó en primer lugar;

a continuación, se indica que los datos se deben agrupar por comunidad y, en último lugar, se requiere la suma de la variable de criminalidad que se precise. Resumiendo, de lo que se encarga esta función es de obtener la suma comunidad a comunidad, del número de fenómenos ocurridos para la variable de criminalidad en el año que se precise y se repite íntegramente este paso para el año con el que se quiere comparar, esta función se denotará por "nnii20". En penúltimo lugar, se seleccionan los pesos con los que se va a ponderar, que como se han promediado, se mantienen constantes por comunidad a lo largo del tiempo. Una vez hecho esto, se procede a multiplicar los valores de la variable de criminalidad de interés por los pesos para comenzar con el cálculo de la fórmula del número índice compuesto ponderado. Ahora, solo queda programar las fórmulas, que calcularán el valor del número índice compuesto ponderado. La primera, *AIe* será la encargada de calcular el vector de incrementos para números índices simples y *AIep* será la que lleve a cabo el cálculo del valor del número índice compuesto ponderado que se está buscando.

Todo el código se puede consultar en los anexos.

Capítulo 4

Bases de datos y preprocesamiento

4.1. Almacenes de datos

En esta investigación se hará uso de dos almacenes de datos. Uno de ellos, construido a través de datos de acceso público y la otra consiste en una base de datos del Cuerpo Nacional de Policía. Sobre los que por sus características se aplicarán unos u otros métodos estadísticos ya que es distinta la información que ofrecen.

4.1.1. Almacén de datos de uso público

El primero de ellos consta de un total de 12 variables y 666 entradas. Como se acaba de señalar se ha construido a través de datos públicos disponibles en fuentes públicas como son el Portal Estadístico de Criminalidad o el INE. Para seleccionar las variables, en el Portal Estadístico de la Criminalidad, se han escogido los delitos de odio y en ellos y se han extraído los hechos conocidos, esclarecidos, victimizaciones y detenciones e investigaciones por tipo de hecho, según las comunidades autónomas, tipología penal total, desde el 2014 a 2020, seleccionando los ámbitos de xenofobia, ideología, identidad de género, sexo, creencias y prácticas religiosas; por último, se seleccionan solo las infracciones penales. Una vez se tienen todos los almacenes de datos individuales correspondientes extraídos se diseñan las variables en el almacén de datos final y se van completando con los datos correspondientes. "year" contiene los años del 2014 a 2020, "comunidad", que contiene el nombre de la comunidad autónoma donde se refiere cada dato, "ambito" que contiene los ámbitos de criminalidad que se han seleccionado y, a continuación, las variables correspondientes a las variales de criminalidad: "detinvest" (detenciones e investigaciones), "victimizaciones", "conocidos" y "esclarecidos". Una vez se han completado estas variables se añadirán cinco nuevas, referidas estas últimas a la demografía de las comunidades autónomas. En primer lugar, se añade la variable "población" que guarda los valores de la población en cada comunidad autónoma para cada año, "km2" recoge la superficie en kilómetros cuadrados de las comunidades autónomas, "denspob" recoge la densidad de población que se calcula como la población de cada comunidad autónoma en cada año dividido entre la superficie "denspob"="población"/"Km2". Por último, se calculan las dos variables restantes que harán falta y más utilizadas como son "pobpr" que se define como el promedio de la población en cada comunidad autónoma en el conjunto de todos los años y la variable "prdenskm2" que será el promedio de las densidades de población para cada comunidad autónoma en el conjunto de años de los que se dispone.

Una vez se ha construido, el siguiente paso es hacer un estudio de las variables con las que se va a trabajar y resulta muy llamativo, que para los ámbitos de ideología y sexo

en el año 2014, solo haya ceros, entonces se procederá a estudiar el motivo de esa gran cantidad de ceros y se concluye con que el motivo de esos duros es distinto para cada categoría. Se concluye con que para los delitos de ideología son datos perdidos que no están recogidos, y en el caso de los delitos por razón de sexo se concluye con que sí están recogidos, ya que hay casos en los que figura un delito y al poner la vista sobre los años siguientes se puede observar que también hay pocos delitos para el ámbito de sexo para los siguientes años, por lo que se concluye con que en este caso los datos no son faltantes. En el caso de los delitos de odio, por razón de ideología en el año 2014 se tomará la decisión de sustituirlos por el promedio de los delitos de odio por razón de ideología en cada comunidad autónoma en los años siguientes. Los delitos de odio por razón de sexo se dejan tal y como están. Una vez se completa este procedimiento, el almacén de datos está disponible para su uso.

4.1.2. Almacén de datos del Cuerpo Nacional de Policía

Tras un largo período de búsqueda de un almacén de datos criminalísticos de las fuerzas y cuerpos de seguridad del estado se ha conseguido disponer de un almacén de datos original y con datos reales del Cuerpo Nacional de Policía, del que no se pueden dar muchos más detalles ya que se trata de información confidencial y sensible al igual que el contenido de este.

Este almacén de datos consta de un total de 13 variables y 240762 entradas correspondientes a individuos que han delinquido. A continuación, se definen las variables.

En primer lugar, se tiene la variable "país", categórica, conformada por un total de 10 regiones geográficas distintas (África, América del Norte; central e insular, España, Europa, Europa del Este, Latinoamérica, Marruecos, Oriente, Oriente Próximo), más la categoría que no recoge ninguna región geográfica por no conocerse (?). Esta variable permite localizar geográficamente la procedencia de cada delincuente. La variable "edad" se trata de una variable categórica formada por cuatro grupos que dividen la edad en grupos. Por una parte, se encuentran los individuos de hasta 18 años, refiriéndose al grupo de adolescentes, a continuación, se tiene el grupo de 18 a 30 años, referido a los adultos jóvenes, el grupo de 31 a 50 refiriéndose al grupo de adultos y, por último, el grupo de adultos mayores que se recoge como los individuos mayores de 50 años. La variable "sexo" que indica el sexo de los individuos. En el caso de las variables cuantitativas, se tienen "nºdetenciones" que se refiere al número de detenciones para cada individuo. "Edad1ª" es la edad del individuo en la primera detención, "Edad2ª" es la edad del individuo en la segunda detención, "mediaentre" se refiere al número medio de días que pasan entre las detenciones, "libertad sexual" cuantifica el número de delitos contra la libertad sexual que ha cometido cada individuo, "orden público" cuantifica el número de delitos contra el orden público que ha cometido cada individuo, "patrimonio" indica el número de delitos contra el patrimonio que comete cada delincuente. "personas" indica el número de delitos contra las personas que comete cada individuo, "relaciones familiares" indica el número de delitos contra las relaciones familiares que comete cada individuo. Por último "salud pública" indica el número de delitos contra la salud pública que comete cada individuo.

Es preciso, señalar que a posteriori, se han creado tantas variables dicotómicas como variables cuantitativas referidas a la tipología de delito hay. Con el objetivo de señalar con un 1 en el caso de que se haya cometido el delito y 0 en el caso de no haberse cometido,

para poder emplearlas en los procesos de minería de datos que se decida utilizar.

Capítulo 5

Aplicaciones y resultados

5.1. Análisis y diagnosis de Criminalidad utilizando gráficos de control

Los gráficos de control pueden ser de mucha utilidad a la hora de analizar datos criminalísticos, sobre todo cuando se tiene en cuenta la variable tiempo. Estos mismos van a permitir identificar si todos los valores en cada una de las secciones temporales están bajo los límites de control y tomar medidas preventivas o correctivas en el caso de que superen los límites de alerta los límites de control respectivamente. Así como de plantear predicciones de cara a los años posteriores cuando se cumpla la premisa que los gráficos estén bajo estado de control.

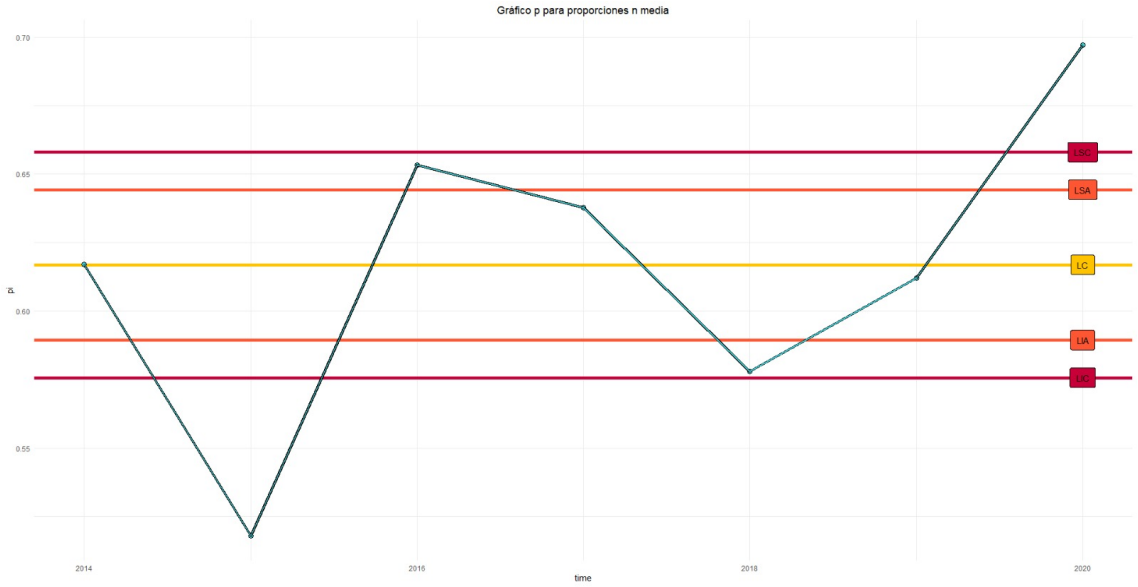
Se procederá ahora a la interpretación de los gráficos de control obtenidos a partir de la librería CCCSA.

5.1.1. Interpretación para gráficos de control de interés

Se ha decidido desarrollar estos gráficos de control considerándose como punto importante el trabajar con tasas, algo que en el ámbito criminalístico es muy útil ya que al englobar a más de una variable aporta un "plus" de información al estudio y es muy común su uso, aplicándose de esta manera el gráfico p. La tasa que se ha considerado ha sido $\frac{esclarecidos}{conocidos}$.

Con un solo ejemplo, bastará para entender la utilidad de estos gráficos aplicados a la diagnosis de la criminalidad. Se pueden hacer tantas combinaciones como el investigador considere oportunas.

Se obtiene el gráfico de control para proporciones con la técnica de n media a través de la función *gcpn* de la librería *CCCSA* para la tasa de casos esclarecidos entre casos conocidos de la suma de los distintos grupos de delitos de odio computados para toda España y se obtiene el siguiente resultado:



Del que en su análisis se pueden sacar las siguientes conclusiones:

Se observa que los puntos en los años 2015 y 2020 se encuentran fuera de control, por lo que se deberían, tomar medidas correctivas de cara a los años siguientes valorando las circunstancias que han provocado que se salgan de control y aplicando medidas para que los siguientes años evitar que sigan fuera de los límites. Por otra parte, se puede observar también que los valores de la variable en los años 2016 y 2018 se encuentran por encima de los límites de alerta, por lo que se deberían tomar medidas preventivas para conseguir que en los años posteriores no se atraviesen los límites de control. En este caso, en todos los puntos, salvo en 2020 se puede observar que estos objetivos se han conseguido, pasando de estar fuera de control en 2015 a estar por encima del límite superior de alerta en 2016 y de aquí, a pasar a estado de control en 2017, por lo que se puede considerar que estos objetivos se han conseguido. Lo mismo se puede decir del valor correspondiente a 2018 que está por debajo del límite de alerta y en 2019 se encuentra en estado de control.

Por otra parte, si se lleva el gráfico a estado de control, que se consigue eliminando los puntos que están fuera de control, se obtiene el gráfico que se adjunta en el anexo, y hallando la media de los puntos que están dentro del mismo se obtiene que es 0.6197. Este valor servirá como predicción de cara al valor que se debería de conseguir el próximo año.

5.2. Diagnosis de Criminalidad empleando Números Índice

Debido a la gran cantidad de combinaciones que se pueden obtener de los datos de los que se dispone, y como el objetivo del presente trabajo es demostrar la capacidad de utilidad de aplicación de las técnicas estadísticas como herramientas para la diagnosis de la criminalidad, la interpretación de los resultados se van a concentrar en el estudio de las comunidades de Madrid, Castilla y León y el total nacional de España, atendiendo al caso de los delitos de odio por razón de ideología a la hora de la interpretación, ya que se considera un caso de especial interés cuando del análisis de números índices simples se trate.

5.2.1. Números Índice Simples para la evolución de los ámbitos de criminalidad: Estudio comparativo de los delitos de odio en las distintas variables de criminalidad para las comunidades de Castilla y León, Madrid, Cataluña y el conjunto de España

Se aplica la función diseñada para números índices simples indicando en el valor de comunidad de Castilla y León, la de Madrid y el conjunto de España según se vaya haciendo. Estos números índice simples están calculados, con período de referencia establecido en la media de todos los valores para cada uno de los ámbitos que se estudian, con el objetivo de penalizar los años en los que no se dispone de ningún dato ya sea porque no se hayan recogido, o porque no están disponibles para esa fecha, sin dejar de considerarlos, promoviendo de esta manera, una vía resolutiva útil para su interpretación sin que se tengan que tomar otras medidas más caóticas que impidan que este proceso se lleve a cabo de manera rápida y eficaz, como puede ser la alternativa que se consideró, de empezar a tomar como referencia el año del que se dispusieran datos, pero, finalmente, se descartó esta posibilidad ya que no se consideraba el recorrido temporal completo ni las particularidades causales para cada uno de los casos, y, a la hora de hacer comparaciones, el resultado, en ciertas ocasiones podría no resultar objetivo ni cierto.

Estudio comparativo de los delitos de odio para la variable de criminalidad "Hechos Conocidos" en las comunidades de Castilla y León, Madrid, Cataluña y el conjunto de España

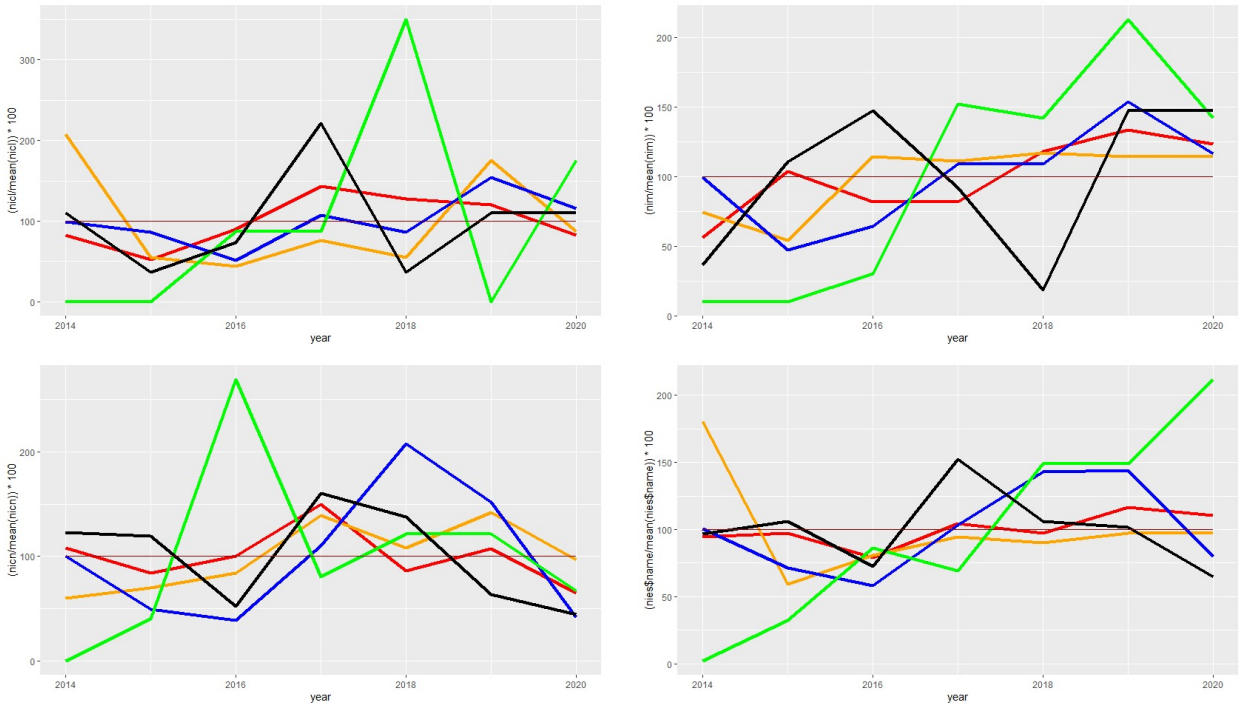


Figura 5.1: Evolución de números índices en los distintos ámbitos de delitos de odio para la variable de criminalidad "Hechos Conocidos" en las comunidades de Castilla y León, Madrid, Cataluña y el conjunto de España

Como en el caso anterior, en primer lugar se atiende al caso total de España, representado en el cuarto gráfico de 5.2 donde se observa que en 2014 toma un valor de 100.49 %, en 2015, toma un valor del 71.39 %, descendiendo ligeramente en el año 2016 hasta alcanzar un valor del 57.95 %, en el 2017 presenta una subida notable hasta alcanzar un valor de 103.18 %, continuando con la misma tendencia en 2018 alcanzando el 143.03 % y estabilizados en 2019 en este mismo valor, por último, en el año 2020 presenta un cambio de tendencia, concluyendo con un valor de 79.7 %.

En Castilla y León, la tendencia es claramente progresiva ascendente, acabando en 2020 con un ligero cambio de ella. En 2014, comienza tomando un valor de 98.77 %, en el año 2015 tienen un valor de 85.89 %, desciende ligeramente en el año 2016 hasta 51.53 %, vuelve a ascender hasta 107.36 % superando el valor de 2015, en 2017 en el año 2018 desciende con respecto a 2017 pero se incrementa con respecto a su homólogo alcanzando un valor del 85.89 %, en 2019 vuelve a incrementarse con respecto a todos los demás y adquiere una tendencia de estabilidad para los siguientes dos años, en 2019 toma un valor de 154.6 % y por último en el 2020, alcanza otra vez un valor de 115.95 %. Se puede concluir que los valores para este caso en esta comunidad no distan demasiado del número índice de referencia establecido en el 100 %, pero a diferencia del caso anterior que se refería a detenciones e investigaciones, presentan una tendencia de crecimiento.

Para la Comunidad de Madrid, como en el caso de Castilla y León, tiene una tendencia progresiva ascendente, bastante más clara que en la anterior, aunque en el 2020 presenta un cambio claro de tendencia con respecto a los años anteriores. En el 2014, comienza con un valor de 99.29 % en el año 2015 toma un valor de 47.16 %, en 2016 toma un valor de 64.53 %, en 2017 asciende hasta alcanzar 109.22 % y manteniendo en 2018 en esta misma cifra, en 2019 alcanza el valor máximo de 153.9 % y en 2020, presenta una bajada hasta el valor de 116.67 %, provocando de esta manera, un cambio de tendencia en los índices.

En el caso de Cataluña, el comportamiento general se asemeja más al del conjunto de España, en este caso, los delitos de odio conocidos en 2014 se parte de un valor del 100.26 %, en el año 2015 presentan un índice de 49.53 %, el 2016 desciende hasta 38.79 %, en el 2017 presenta una notable subida hasta alcanzar el valor de 110.4 %, alcanzando el valor máximo en el 2018 con un índice de 207.67 %, presenta un descenso en el año 2019 hasta llegar al 151.58 % y desplomándose en 2020 llegando al 41.77 %.

En el caso de los delitos conocidos en España, todos están concentrados en torno al número índice de referencia con una tendencia global creciente salvo en las categorías "Ideología" y "Creencias y prácticas religiosas" que terminan en 2020 con un comportamiento decreciente. En el caso de Castilla y León, al igual que en el caso de detenciones e investigaciones, la categoría de "Xenofobia", "Identidad de género" e "Ideología" tienen un comportamiento similar, con la única diferencia con respecto al caso anterior de que la tendencia global es progresiva ascendente, las dos categorías restantes, presentan fluctuaciones más pronunciadas pero su tendencia también es ascendente. En el caso de la Comunidad de Madrid, el comportamiento de todas sus categorías como se puede observar presenta un comportamiento similar con tendencia claramente ascendente, presentando un comportamiento de más fluctuación al igual que en el caso de Castilla y León en las categorías de "Sexo" y "Creencias y prácticas religiosas". En el caso de Cataluña, se presenta un caso bastante peculiar con respecto a los demás, puesto que las tendencias de todas las categorías son ascendentes hasta el año 2018 a partir del cual cambia para todas las categorías. Como se puede valorar en conjunto, para este caso, se observa que las

tendencias generales tienen un carácter ascendente. Realmente también se puede observar que la categoría de delitos por "Identidad" de género, en este caso evidencia de forma bastante representativa lo que pasa en cada una de las regiones que se han analizado.

Estudio comparativo de los delitos de odio por razón de ideología para la variable de criminalidad "Hechos Esclarecidos" en las comunidades de Castilla y León, Madrid, Cataluña y el conjunto de España

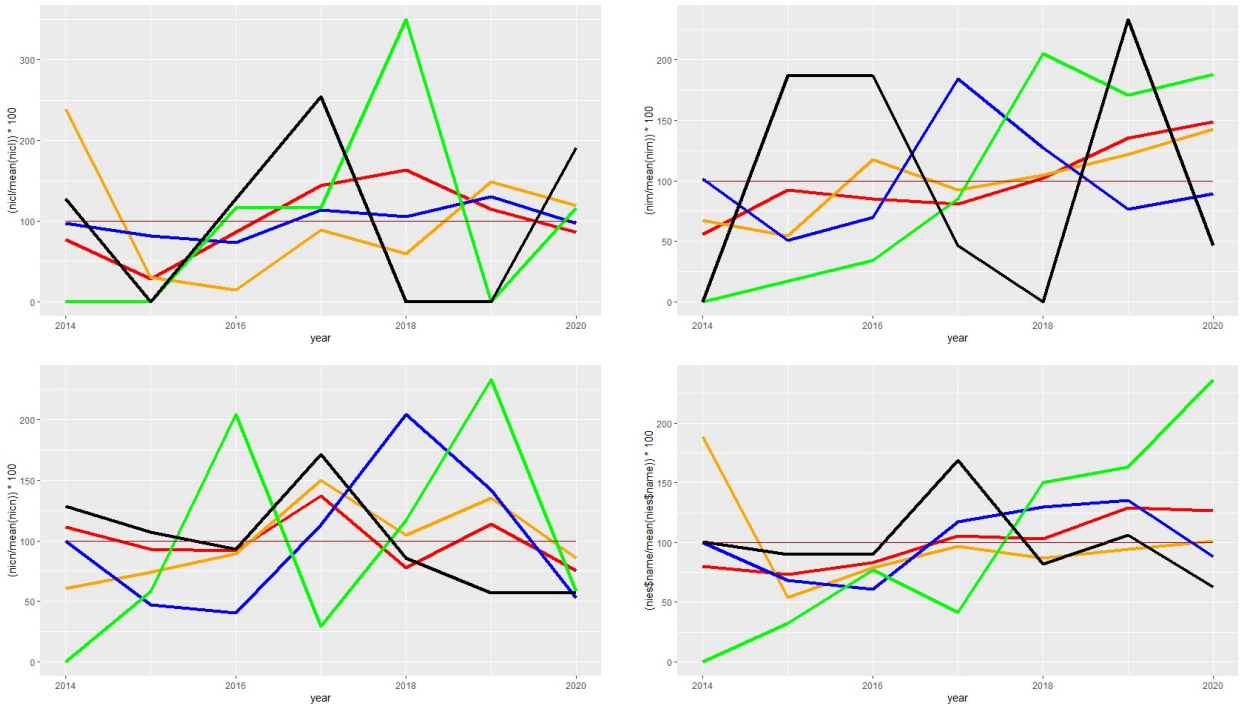


Figura 5.2: Evolución de números índices en los distintos ámbitos de delitos de odio para la variable de criminalidad "Hechos Esclarecidos" en las comunidades de Castilla y León, Madrid, Cataluña y el conjunto de España

En el caso de los delitos esclarecidos, a nivel nacional, como en los dos anteriores casos, se observa en el año 2014 que toma un valor de 99.77 %, en el año 2015 el índice es del 68.52 %, desciende en 2016 a 60.85 % y crece a partir de 2017 alcanzando un valor de 117.3 % para continuar con la tendencia alcista el año 2018 alcanzando un valor de 129.91 % presentando un paulatino crecimiento el año 2019 alcanzando un valor de 135.4 % y finalmente caer en el año 2020 con un valor de 88.25 %. El comportamiento es parecido al de los delitos conocidos para el mismo caso.

Para Castilla y León, el comportamiento también es similar a los conocidos, con una ligera tendencia a la estabilidad a partir del año 2018. En el 2014, toma un valor de 97.67 %; en el año 2015 tiene un valor de 81.4 %; desciende ligeramente, en el año 2016, hasta 73.26 %, vuelve a ascender hasta 113.95 % superando el valor de 2015; en 2017 sufre un descenso hasta 105.81 %, en el año 2018 se incrementa hasta 130.23 %, descendiendo de nuevo en 2020 alcanzando un valor del 97.67 %.

En el caso de la Comunidad de Madrid, en el 2014, toma un valor de 101.82 %; en el año 2015, el índice toma un valor de 50.91 %; en 2016, presenta una subida hasta 70 %, para ascender en 2017 alcanzando su máximo en 184.54 % para descender progresivamente

durante los dos siguientes años, tomando valores de 127.27 % y 76.36 %, respectivamente, rebasando inferiormente el número índice de referencia; por último, en 2020 presenta un ligero crecimiento hasta alcanzar 89.09 %.

Por último, para el caso de Cataluña, presenta un comportamiento parecido al de la Comunidad de Madrid. En el 2014 toma un valor de 99.76 %; en el año 2015, toma un valor de 47.34 %; en 2016, desciende hasta 40.58 %, volviendo a ascender en el 2017 hasta 113.29 %, sigue creciendo y alcanza su valor máximo de 204.59 % en 2018 y desciende en los dos siguientes tomando valores de 142.03 % en 2019, y finalizando, con un valor de 52.42 % en 2020.

Haciendo una valoración global de cada uno de los casos, en España se puede observar que todos los casos están concentrados alrededor del número índice de referencia con una tendencia que se podría valorar entre ligero crecimiento y estabilidad, destacando la categoría de "sexo" que presenta un crecimiento muy pronunciado con respecto a todas las demás. En el caso de Castilla y León, al igual que en los casos anteriores las categorías de "Xenofobia", "Identidad de género" e "Ideología" tienen un comportamiento similar presentando una tendencia entre crecimiento y estabilidad. Las dos categorías restantes presentan una fluctuación mayor con datos nulos en diversos años. En el caso de la Comunidad de Madrid, presenta un comportamiento muy heterogéneo entre las categorías siguiendo las de "Xenofobia" e "Identidad de género" un comportamiento muy similar y junto a "Ideología" presentan una tendencia general similar, por otro lado, las dos variables restantes tienen un comportamiento parecido al caso de los delitos conocidos. Por último, en el caso de Cataluña, las categorías "Xenofobia", "Identidad de género" y "Creencias y prácticas religiosas" tienen un comportamiento similar presentando un pico en el año 2017 y bajando su tendencia posteriormente, a este comportamiento, aunque no tan simétrico, se podría unir también la categoría "Ideología", la categoría restante presenta un fluctuación mucho más intensa.

5.2.2. Números Índice Compuestos para la evolución de los ámbitos de criminalidad

Para este apartado, se han empleado tanto los números índices no ponderados y los ponderados, en el caso de los no ponderados se ha considerado su utilidad cuando se pretende estudiar el número índice compuesto que proporciona la consideración de todos los ámbitos para los delitos de odio, en cambio para los índices ponderados, se ha considerado de utilidad la ponderación cuando se está calculando el índice para las distintas Comunidades Autónomas teniendo en cuenta la totalidad de los ámbitos de criminalidad, haciendo la ponderación a través de factores como pueden ser la población por comunidad autónoma o la densidad de población por comunidad autónoma entre otros.

Números Índice Compuestos no ponderados para el estudio de los ámbitos de criminalidad para los delitos de odio

En este primer caso, se ha considerado de utilidad emplear los números índice compuestos sin ponderar para estudiar a través de estos mismos la evolución de la variable de casos de delitos de odio conocidos teniendo en cuenta los ámbitos de criminalidad para los delitos de odio en las regiones que se consideren oportunas, en este caso se tomará como referencia el conjunto de España y posteriormente se comparará con Madrid, durante períodos de interés que se considerarán el año del comienzo de la recogida de datos y el

último año del que se dispone de datos, es decir, 2020 y por otro lado considerando la media de los números índices durante todos los años con respecto al año de la pandemia, para ver si se obtiene algún resultado significativo.

Se ha considerado emplear Números Índices Compuestos sin ponderar en este caso, porque es muy difícil establecer una asignación de pesos a los ámbitos con los que se está trabajando por orden de intensidad puesto que se considera igual la gravedad de cualquiera de ellos y no se dispone de ningún tribunal de expertos que sea capaz de analizar los factores que puedan intervenir y describir a cada uno de ellos que pueda determinar su mayor o menor gravedad con respecto a los demás. Y para el objetivo de este estudio, tampoco se considera pertinente valorar esta opción.

Se tiene que:

España			
Ámbito	2014	2020	$\overline{X}_{2014-2019}$
Xenofobia	416	485	(416+427+348+459+426+115):6 = 365.16
Identidad de Género	513	277	(513+168+228+268+256+278):6 = 285.17
Ideología	410	326	(410+295+237+423+585+596):6 = 424.33
Sexo	55	99	(55+20+40+32+69+69):6 = 47.5
Creencias o Prácticas Religiosas	63	45	(63+69+47+99+69+66):6 = 69.33

Cuadro 5.1: Delitos de odio en España según ámbito del delito en los años 2015, 2020 y la media desde 2014 hasta 2019

Una vez se dispone, de los datos necesarios para calcular los números índice compuestos que se quiere, se procede a su cálculo, en primer lugar, obteniendo los valores de los Números Índice Simples en incremento respectivos:

España					
Ámbito	2014	2020	$\overline{X}_{[2014,2019]}$	$\Delta Ii_{2014-2020}$	$\Delta Ii_{\overline{X}_{[2014,2019]}-2020}$
Xenofobia	416	485	365.16	16.587	0.328
Identidad de Género	513	277	285.17	-46.004	-0.028
Ideología	410	326	424.33	-20.488	-0.232
Sexo	55	99	47.5	80.000	1.084
Creencias o Prácticas Religiosas	63	45	69.33	-28.571	-0.351
			$\sum$	1.524.	0.801

Cuadro 5.2: Delitos de odio en España según ámbito del delito en los años 2015, 2020 y la media desde 2014 hasta 2019 con los respectivos cálculos de los Números Índices Simples en forma de incremento para cada una de las combinaciones que se pretenden estudiar

En la tabla se recoge el valor calculado los números Índices Simples en incremento, que permitirá calcular los Números Índice Compuestos que se buscan para realizar la comparación entre el primer y último año del período del que se dispone y en segundo lugar, la comparación de todos los años del período del que se dispone con respecto al año de la pandemia.

Los valores de los Números Índice Compuestos para el caso de la variable de criminalidad Detenciones e investigaciones, teniendo en cuenta los ámbitos de criminalidad en España son los siguientes y se calculan del siguiente modo:

$$\Delta IC_{2014-2020} = \frac{\sum \Delta Ii_{2014-2020}}{5} = 0{,}3047\,\% \tag{5.1}$$

$$\Delta IC_{[2014,2019]-2020} = \frac{\sum \Delta Ii_{\overline{X}_{[2014,2019]}-2020}}{5} = 0{,}1602\,\% \tag{5.2}$$

Se repite el mismo proceso para la Comunidad de Madrid, que permitirá hacer posteriormente la comparación.

En primer lugar, se recogen en las tablas los datos con los que se van a trabajar.

Comunidad de Madrid			
Ámbito	2014	2020	$\overline{X}_{2014-2019}$
Xenofobia	44	96	(44+81+64+64+92+104):6 = 74.833
Identidad de Género	26	40	(26+19+40+39+41+40):6 = 34.167
Ideología	40	47	(40+19+26+44+44+62):6 = 39.167
Sexo	1	14	(1+1+3+15+14+21):6 = 9.167
Creencias o Prácticas Religiosas	2	8	(2+6+8+5+1+8):6 = 5

Cuadro 5.3: Delitos de odio conocidos en la Comunidad de Madrid según ámbito del delito en los años 2015, 2020 y la media desde 2014 hasta 2019

A continuación se presentan los resultados de los cálculos, obtenidos a través de los programas diseñados en R, de los Números Índice Simples en incremento para cada uno de los ámbitos en los períodos que se han decidido estudiar.

España					
Ámbito	2014	2020	$\overline{X}_{[2014,2019]}$	$\Delta Ii_{2014-2020}$	$\Delta Ii_{\overline{X}_{[2014,2019]}-2020}$
Xenofobia	44	96	74.833	1.363	0.283
Identidad de Género	26	40	34.167	0.538	0.171
Ideología	40	47	39.167	0.175	0.2
Sexo	1	14	9.167	13	0.527
Creencias o Prácticas Religiosas	2	8	5	3	0.6
			$\sum$	18.076	1.781

Cuadro 5.4: Delitos de odio conocidos en la Comunidad de Madrid según ámbito del delito en los años 2015, 2020 y la media desde 2014 hasta 2019 con los respectivos cálculos de los Números Índices Simples en forma de incremento para cada una de las combinaciones que se pretenden estudiar

Calculando los valores de los Números Índice Compuestos simples para el caso de delitos conocidos, teniendo en cuenta los ámbitos de criminalidad en la Comunidad de Madrid, se tiene que:

$$\Delta IC_{2014-2020} = \frac{\sum \Delta Ii_{2014-2020}}{5} = \frac{18{,}076}{5} = 3{,}6152\,\% \tag{5.3}$$

$$\Delta IC_{[2014,2019]-2020} = \frac{\sum \Delta Ii_{\overline{X}_{[2014,2019]}} - 2020}{5} = \frac{1{,}781}{5} = 0{,}3562\,\% \tag{5.4}$$

5.2.3. Números Índice Compuestos ponderados para el estudio de los delitos de odio considerando las Comunidades Autónomas como variable de ponderación

También se ha considerado útil analizar los delitos de odio conocidos empleando Números Índices Compuestos ponderados, teniendo en cuenta las Comunidades Autónomas y haciendo esta ponderación a través de una serie de variables que están referidas a estas mismas como son la población en cada comunidad, o la densidad de población, como antes, comparándose en los períodos de interés que se han venido utilizando hasta el momento. Es muy útil considerar esta opción en este caso ya que al tener cada Comunidad Autónoma unas características que pueden ser medidas y determinadas de una manera objetiva, se pueden utilizar como factor de ponderación a través de las variables que se consideren útiles para hacerlas.

A continuación se presentan las dos tablas que contienen los delitos de odio conocidos para los años y períodos de estudio, agrupados por comunidades autónomas considerando las variables de ponderación *Población promediada* y *Densidad de población promediada* así como los resultados de los cálculos de los incrementos par los Número Índice Simples aplicándoles la ponderación que les corresponde a cada una de las Comunidades Autónomas, obtenidos con los programas diseñados en R.

w = Población						
Comunidad Autónoma	w_i	2014	2020	$\overline{X}_{[2014,2019]}$	$\Delta wIi_{2014-2020}$	$\Delta wIi_{\overline{X}_{[2014,2019]}-2020}$
Andalucía	8416668	253	130	135	-4091897.88	-311728.44
Aragón	1322338	57	18	28	-904757.58	-467173.31
Asturias	1036111	31	18	23	-434498.16	-236825.37
Baleares	1156163	57	24	26	-669357.53	-75075.52
Canarias	2164580	31	41	27	698251.61	1082290.00
Cantabria	583192	12	10	9	-97198.67	89721.85
Castilla y León	2441669	56	51	49	-218006.16	82488.82
Castilla la Mancha	2048690	65	41	42	-756439.38	-64525.67
Cataluña	7481450	364	228	410	-2795267.03	-3322724.44
Comunidad Valencian	4959154	91	131	93	2179847.91	2038862.24
Extremadura	1078339	12	19	14	629031.08	385121.07
Galicia	2716899	69	41	47	-1102509.74	-321353.65
Madrid	6514687	113	205	162	5303992.96	1712279.13
Murcia	1476161	26	26	23	0	192542.74
Navarra	642999	32	43	23	221030.91	567904.23
País Vasco	2171643	121	185	122	1148637.62	1112449.28
La Rioja	313801	5	3	7	-125520.40	-172590.55
Ceuta	84739	2	3	1	42369.50	169478.00
Melilla	84574	4	10	3	126861.00	169148.00
				$\sum$	-845429.9	2630288

Cuadro 5.5: Delitos de odio conocidos en España en los años 2014, 2020 y la media desde 2014 hasta 2019 junto al cálculo de los Números Índices Simples en forma de incremento para cada una de las combinaciones de estudio, para calcular los Números Índice Compuestos ponderados según la variable "Población promedio" establecida a través de las Comunidades Autónomas

Calculando los Números Índice Compuestos ponderados correspondientes:

$$\Delta ICp_{2014-2020} = \frac{\sum \Delta Ii_{2014-2020} \cdot w_i}{w_i} \cdot 100 = \frac{-845429{,}9}{46693857} \cdot 100 = -1{,}81058\,\% \qquad (5.5)$$

$$\Delta ICp_{[2014,2019]-2020} = \frac{\sum \Delta Ii_{\overline{X}_{[2014,2019]}-2020} \cdot w_i}{\sum w_i} \cdot 100 = \frac{2630288}{46693857} \cdot 100 = 5{,}63305\,\% \qquad (5.6)$$

w = Densidad de población						
Comunidad Autónoma	w_i	2014	2020	$\overline{X}_{[2014,2019]}$	$\Delta Ii_{2014-2020} w_i$	$\Delta wIi_{\overline{X}_{[2014,2019]}-2020}$
Andalucía	96.08	253	130	135	-46.710830	-3.55851
Aragón	27.7114	57	18	28	-18.960451	-9.79026
Asturias	97.7114	31	18	23	-40.975760	-22.33404
Baleares	231.6029	57	24	26	-134.085865	-15.03914
Canarias	290.6657	31	41	27	93.763134	145.33285
Cantabria	109.6029	12	10	9	-18.267143	16.86197
Castilla y León	25.9143	56	51	49	-2.313776	0.87548
Castilla la Mancha	25.7814	65	41	42	-9.519297	-0.81201
Cataluña	232.9714	364	228	410	-87.044270	-103.46922
Comunidad Valencian	213.2514	91	131	93	93.736892	87.67428
Extremadura	25.9	12	19	14	15.108333	9.25
Galicia	91.8657	69	41	47	-37.278841	-10.86583
Madrid	811.4943	113	205	162	660.685613	213.28802
Murcia	130.4986	26	26	23	0	17.02155
Navarra	61.88	32	43	23	21.271250	54.65313
País Vasco	300.1986	121	185	122	158.782715	153.78019
La Rioja	62.1986	5	3	7	-24.879429	-34.20921
Ceuta	4236.9574	2	3	1	2118.478571	8473.91428
Melilla	7047.8229	4	10	3	10571.734286	14095.6457
				$\sum$	13313.53	45592288

Cuadro 5.6: Delitos de odio conocidos en España en los años 2014, 2020 y la media desde 2014 hasta 2019 junto al cálculo de los Números Índices Simples en forma de incremento para cada una de las combinaciones de estudio, para calcular los Números Índice Compuestos ponderados según la variable "Población promedio" establecida a través de las Comunidades Autónomas

$$\Delta ICp_{2014-2020} = \frac{\sum \Delta Ii_{2014-2020} \cdot w_i}{w_i} \cdot 100 = \frac{13313{,}53}{14120{,}11} \cdot 100 = 94{,}2877\,\% \qquad (5.7)$$

$$\Delta ICp_{[2014,2019]-2020} = \frac{\sum \Delta Ii_{\overline{X}_{[2014,2019]}-2020} \cdot w_i}{\sum w_i} \cdot 100 = \frac{23068{,}22}{14120{,}11} \cdot 100 = 163{,}3714\,\% \qquad (5.8)$$

5.2.4. Conclusiones sobre los Números Índices Compuestos

Interpretación de los NICS sin ponderar:

La información que se extrae de estos casos es que, en primer lugar, para el período de 2014 a 2020 es que el número índice para la Comunidad de Madrid de 3.66152 % es superior al índice del conjunto de España siendo este último 0.3047 %, lo que indica una mayor criminalidad en lo que se refiere a los delitos de odio teniendo en cuenta los respectivos ámbitos de estudio.

En el caso del periodo que comprende la media en los años desde el 2014 al 2019 y 2020 teniendo en cuenta el año de la pandemia con respecto a los años anteriores, en este caso, sigue siendo superior el número índice en la Comunidad de Madrid, siendo de 0.3562 % con respecto al número índice 0.1602 % correspondiente a la totalidad de España, la variación que existe entre ellos es muy ligera.

En España, el índice para el primer período 0.3047 % es mayor que el referido al segundo siendo 0.1602 %. El número índice compuesto para el periodo comprendido entre el primer y el último año del que se dispone es mayor que el número índice correspondiente al período que promedia los años anteriores al 2020 y este mismo.

En el caso de Madrid, el índice para el primer período 3.6152 % es tres puntos mayor que el referido al segundo siendo 0.3562 %. El número índice compuesto para el periodo comprendido entre el primer y el último año del que se dispone es mayor que el número índice correspondiente al período que promedia los años anteriores al 2020 y el 2020.

Interpretación de los NICS ponderados:

La principal dificultad en este punto reside en escoger cuáles van a ser los factores de ponderación y qué peso se les van a asignar por lo que se precisa de un estudio previo exhaustivo donde se han establecido las cargas como la media de los valores de las dos variables en cada comunidad autónoma desde el año 2014 hasta el 2020 por lo tanto así resultan unas ponderaciones fijas y promediadas que se podrán utilizar para el cálculo de los Números Índice Compuestos de una manera rigurosa. En este caso se ha trabajado con la población y la densidad de población. Resulta interesante discutir cual podría ser el mejor factor de ponderación. La ventaja de ponderar con respecto a la densidad de población es que considera la superficie; la ponderación será mayor en las Comunidades donde mayor concentración de población haya, por lo tanto es una ventaja directa para analizar la criminalidad de forma más objetiva puesto que penaliza las comunidades autónomas más grandes y da más importancia a las que mayor concentración poblacional tienen, por lo tanto en las zonas donde mayor concentración poblacional haya, mayor serán los niveles de criminalidad. Entonces, la variable de ponderación que se considera más acertada es la densidad de población.

En primer lugar, utilizando como variable de ponderación la población en las comunidades autónomas, para el período de 2014 a 2020 el número índice tiene un valor de -1.810583 es inferior al índice para el periodo promedio de 2014 a 2019 comparado con el año 2020 siendo este 5.63305 %.

Ahora, utilizando como variable de ponderación la densidad de población en las co-

munidades autónomas, para el período de 2014 a 2020 el número índice tiene un valor de 94.2877 % es inferior que el periodo promedio de 2014 a 2019 comparado con el año 2020 teniendo un valor de 163.3714 %.

Si se compara ahora el número índice en el primer período para la primera variable de ponderación y la segunda se puede apreciar una diferencia enorme, lo que implica la gran diferencia que existe entre el empleo de un factor de ponderación u otro. Lo mismo pasa el segundo período.

5.3. Análisis estadístico de criminalidad sobre datos del CNP

5.3.1. Análisis descriptivo de las variables

En el conjunto de las variables de las que se dispone se pueden diferenciar una serie de variables cualitativas y cuantitativas. De las cualitativas se puede extraer la siguiente información:

En la variable *país* se observa que la nacionalidad predominante con respecto a las demás en los delitos es la Española, ahora bien, si se computa el conjunto de delitos para las personas con nacionalidad extranjera se observa que los delitos suman una menor cantidad sobre el total, siendo los latinoamericanos los que recogen un mayor porcentaje de delitos en la población extranjera y los norteamericanos, la curiosidad, por ejemplo, sería ahora, comparar todos los datos con respecto a la cantidad de población española y extranjera para observar cual de ellas tiene una mayor proporción de delincuencia. (Figura 9.1)

Para la variable *edad*, se puede observar que en el grupo en el que más delitos se concentran es en el grupo de *19 - 30* y es destacable también que el siguiente grupo con mayor concentración de delitos es el de menores de edad hasta 18 años, siendo menor en el grupo de *>=50 años*. Por lo tanto, esto puede generar una pregunta: ¿La edad influye en el hecho de cometer delitos? A priori, sin hacer ningún tipo de contraste se puede comentar que, efectivamente, se observa una posible relación entre la edad y el hecho de delinquir. Parece que a medida que los grupos están alejados de las edades más jóvenes se delinque menos. (Anexo figura 9.2)

En el caso de la última variable categórica *sexo*, Los hombres tienen una concentración de delitos casi diez veces superior a las mujeres. Aunque luego se tendría que valorar el número de detenciones totales. (Anexo figura 9.3)

Ahora, para el caso de las variables cuantitativas se puede extraer a primera vista la siguiente información:

En el caso del número de detenciones (*nº detenciones*) la media de detenciones es de 5.80 por individuo, casi 6, una cifra considerablemente elevada, lo que induce que la reincidencia es alta, con mínimo de 1 y máximo de 134, por lo que el rango será de 131 percibiéndose una amplia variabilidad que se confirma en la desviación típica o varianza con un valor de 4.624 y 21.378. Su distribución de frecuencias está considerablemente desplazada hacia la izquierda y presenta un apuntamiento considerable. (Anexo figura 9.12)

Para la edad en la primera detención (*edad 1ª*), la edad es de 26.61 años, con un mínimo de 14 y un máximo de 90 años. Por último señalar que la desviación típica y la varianza tienen valores de 9.868 y 97.380 respectivamente. Su distribución está ligeramente desplazada hacia la izquierda y su apuntamiento está ligeramente elevado. (Anexo figura 9.13)

La edad media a la que se produce la segunda detención es de 33.73 años, 7 años superior que en el caso de la media para la edad en la primera detención, es decir la diferencia de entre las medias de edades en la primera y la segunda detención es de 7 años. Con un mínimo de 14 y un máximo de 91 años. Por último señalar que la desviación típica y la varianza tienen valores de 10.702 y 114.924 respectivamente, es decir, la variabilidad de los datos es superior en el caso de la edad en la segunda detención. Su distribución está ligeramente desplazada hacia la izquierda y su apuntamiento está ligeramente elevado. (Anexo figura 9.14)

En el caso del tiempo entre la primera y la segunda detención (*mediaentre* la media de días entre la primera y la segunda detención es de 695.757, con un mínimo de 0 días, es decir, que en algunos casos, las dos detenciones se han llevado a cabo el mismo día, y un máximo de 9287,5 días. La desviación típica y la varianza tienen valores de 713,4 y 508955,7 respectivamente. La distribución está ligeramente desplazada a la izquierda y apuntamiento notablemente elevado. 9.15

Las siguientes variables se estudiarán conjuntamente. La media de los delitos para libertad sexual (Figura 9.16), contra el orden público (Figura 9.17), contra las personas (Figura 9.18), contra las relaciones familiares (Figura 9.19)y la salud pública (Figura 9.20), no llegan a 2. La única que es superior a 2 es la media de patrimonio que es igual a 3.46. Todas ellas tienen su mínimo en 1. los máximos son respectivamente 13, 28, 11, 25, 22 y 92. La variabilidad en todas ellas es similar, salvo en la variable que se refiere a los delitos por causa de patrimonio que es muy superior a las anteriores. Algo que se podría intuir a partir de la media de esta y su máximo con respecto a las demás. Figura 9.4

Las salidas de SPSS a partir de las que se han interpretado las variables se adjuntan en los anexos.

5.3.2. Contrastes útiles para las variables de interés

Ahora se podrían plantear las siguientes preguntas:

¿Existen diferencias significativas del número de detenciones dependiendo de la nación de procedencia?

¿Existen diferencias significativas en el número de detenciones entre la gente autóctona y la gente extranjera?

¿Existen diferencias significativas del número de detenciones dependiendo del grupo de edad?

¿Existen diferencias significativas entre la edad a la que se comete el primer y/o el segundo delito dependiendo del sexo?

¿Existen diferencias significativas entre la edad a la que se comete el primer y/o el segundo delito dependiendo del grupo de edad?

Cada una de estas preguntas se podrá resolver planteando y resolviendo distintos contrastes ANOVA. En el segundo caso, habrá que recodificar la variable de forma que los grupos distintos a España, queden agrupados en uno que conglomere las naciones restantes, bajo la categoría de extranjeros.

Para el primer caso, realizando el contraste ANOVA, como el p-valor es < 0.05, para un $\alpha = 5\,\%$ y para el conjunto de datos con el que se está trabajando, se puede concluir que existen diferencias para el número de detenciones según la procedencia. Ahora, a través de los contrastes a posteriori se observan los grupos en los cuáles existen esas diferencias. África-China, África-España, África-Latinoamérica, África-Marruecos, África-Oriente, África-Oriente Próximo, América del Norte; Central o insular-España, China-España, China-Europa, China-Europa del Este, China-Marruecos, España-Europa, España-Europa del Este, España-Latinoamérica, España-Marruecos, España-Oriente, España-Oriente Próximo, Europa-Latinoamérica, Europa-Oriente, Europa-Oriente Próximo, Europa del Este-Latinoamérica, Europa del Este-Oriente, Europa del Este-Oriente Próximo, Latinoamérica-Marruecos, Latinoamérica- Oriente Próximo, Marruecos-Oriente, Marruecos-Oriente Próximo, Oriente-Oriente Próximo. (Anexo figuras 9.5 y 9.6)

Para el segundo caso, realizando el contraste ANOVA, como el p-valor es < 0.05, para un $\alpha = 5\,\%$ y para el conjunto de datos con el que se está trabajando, se puede concluir que existen diferencias para el número de detenciones dependiendo de si el sujeto es autóctono o extranjero. (Anexo figura 9.10)

Para el tercer caso, realizando el contraste ANOVA, como el p-valor es < 0.05, para un $\alpha = 5\,\%$ y para el conjunto de datos con el que se está trabajando, se puede concluir que existen diferencias para el número de detenciones dependiendo del grupo de edad. Siendo significativas todas las combinaciones que se pueden visualizar en los contrastes a posteriori. (Anexo figuras 9.7 y 9.8)

Para el cuarto caso, realizando el contraste ANOVA, como el p-valor es < 0.05, para un $\alpha = 5\,\%$ y para el conjunto de datos con el que se está trabajando, se puede concluir que existen diferencias para la edad en la que se comete el primer delito dependiendo del sexo.

Para el cuarto caso, realizando el contraste ANOVA, como el p-valor es > 0.05, para un $\alpha = 5\,\%$ y para el conjunto de datos con el que se está trabajando, al no ser el contraste significativo, se puede concluir que no existen diferencias para la edad en la que se comete el segundo delito dependiendo del sexo. (Anexo figura 9.9)

Para el quinto caso, realizando el contraste ANOVA, como el p-valor es < 0.05, para un $\alpha = 5\,\%$ y para el conjunto de datos con el que se está trabajando, se puede concluir que existen diferencias para la edad en la que se comete el primer y el segundo delito dependiendo de la edad. Siendo significativos todos los contrastes a posteriori significativos, es decir, existen diferencias entre todos los grupos entre sí. (Anexo figura 9.7)

Las salidas de SPSS a partir de las que se han interpretado las variables se adjuntan en los anexos.

5.3.3. Algunas técnicas de minería de datos

En las técnicas de minería de datos, se considera interesante desarrollar un proceso de clasificación a través de un árbol de decisión, de asociación a través del algoritmo a priori y de agrupación utilizando el algoritmo k-medias. En el estudio de la criminalidad es interesante el uso de los árboles de decisión puesto que permiten encontrar relaciones entre variables que pasan desapercibidas a simple vista. Lo que puede aportar una información muy valiosa a los investigadores.

En este caso se utilizará Weka, el *software* para minería de datos desarrollado desde la universidad de Waikato en Australia, desde donde se pueden aplicar un gran abanico de técnicas de minería de datos. El único inconveniente es que los archivos de datos tienen que estar en un archivo *.arff* que previamente se haya preparado desde un *.txt* en el formato que acepta este programa. Por lo tanto, el primer paso será transformar los datos del *.csv* al formato de lectura de datos que admite Weka.

Una vez se ha hecho esto, se modificarán las variables que hacen referencia a los delitos, recodificándolas todas de forma binaria, asignando el 0 en el caso de que no se haya cometido el delito y un 1 en el caso de que sí. Una vez se ha hecho esto, se procede al desarrollo de los respectivos árboles de decisión. El algoritmo que se decide usar es el J48 que es uno de los más utilizados este tipo de estudios.

Árbol de decisión J48

En primer lugar lo que se pretende es clasificar a los delincuentes en lo que se refiere a delitos contra el patrimonio, ya que como España es un pais caracterizado por su patrimonio cultural y artístico, se considera una variable que puede resultar destacable su estudio, en función de la nacionalidad, el sexo, el grupo de edad, si se han cometido delitos contra el orden público y contra la salud pública.

El resultado es un árbol de clasificación que ha conseguido clasificar correctamente un 74.509 % de la información por lo que se le considera un modelo de clasificación muy robusto, dejando de esta manera un 25.491 % de la información sin clasificar, en el que salen del nodo original, cuatro nodos hijos, perteneciendo cada uno a un grupo de edad. El primer nodo es correspondiente a los detenidos de hasta 18 años cometen delitos contra el patrimonio

Para los que se encuentran en el grupo de 19 a 30 años, si no cometen delitos de orden público, si no cometen tampoco delitos contra la salud pública, entonces tampoco cometen delitos contra el patrimonio. En el caso de que si cometan delitos contra el patrimonio, entonces, si el número de detenciones es menor o igual a tres y son varones, entonces no cometen delitos contra el patrimonio, en el caso de que sean mujeres, si los cometen. Ahora, si el número de delitos es mayor que tres, cometen delitos contra el patrimonio. En el caso de que sí cometan un delito contra el orden público y el número de delitos es menor o igual a tres, implica que no cometen delitos contra el patrimonio. en el caso de que sean menor a tres y no cometan delitos contra la salud pública, entonces, cometerán delitos contra el patrimonio. En caso contrario, si el número de detenciones es menor o igual a cuatro, no cometen delitos contra el patrimonio y si por el contrario, el número de detenciones es mayor que cuatro, cometen delito contra el patrimonio.

En el caso del tercer nodo, que se refiere al grupo de edad comprendido entre 31 y 50 años, si son hombres y el número de detenciones es menor o igual a tres, entonces no cometerán delitos contra el patrimonio, si son mayores a tres, no cometen delitos de orden público, el número de detenciones es menor o igual a cuatro y no se cometen delitos contra la salud pública, entonces se cometerán delitos contra el patrimonio, en caso de que sí se cometan delitos contra la salud pública, no se cometerán delitos contra el patrimonio. En el caso de que sí se cometan delitos contra el orden público y el número de detenciones sea menor o igual a cuatro, entonces, no se cometen delitos contra el patrimonio; en el caso de que sean las detenciones mayores que cuatro, se cometerán delitos contra el patrimonio. En el caso de las mujeres, si no han cometido delitos contra el orden público ni la salud pública, cometen delitos contra el patrimonio. En el caso de que sí hayan cometido delito contra la salud pública y el número de delitos sea menor o igual a cuatro, entonces no cometerán delito contra el patrimonio; en el caso de que los delitos sean mayores a cuatro, entonces sí cometerán delitos contra el patrimonio. En el caso de que sí hayan cometido delitos contra el orden público y el número de detenciones sea menor o igual a cuatro, entonces no cometerán delito contra el patrimonio; al contrario, si el número de delitos es mayor que cuatro, entonces cometerán delito contra el patrimonio.

Por último, poniendo la mirada sobre el último nodo primario, que está referido al grupo de edad de cincuenta o más años, si son hombres, no cometen delito contra el patrimonio. Si son mujeres y no han cometido delito contra la salud pública, y tampoco han cometido delitos contra el orden público, entonces, cometerán delitos contra el patrimonio; si han cometido delitos contra el orden público, no cometerán delitos contra el patrimonio. En el caso de que hayan cometido delitos contra la salud pública, entonces no cometen delitos contra el patrimonio. En la interpretación se tratará de obviar las reglas redundantes.

(Anexo figuras 11.1, 11.2, 11.3)

Toda esta información que se extrae es muy útil una vez se entrega a los servicios policiales ya que facilita comprender el comportamiento de los delincuentes y desarrollar planes de acción según el grupo en el que estén enmarcados que a su vez oriente al tipo de acción criminal que se va a desarrollar y poderse de esta manera subsanar previamente a que se haya cometido.

Asociación: Algoritmo a priori

El algoritmo a priori, como se comentaba en la teoría, tratará de extraer relaciones entre las variables que no se pueden extraer a simple vista solo con observar los datos. Se ha decidido trabajar con todas las variables cualitativas que componen la base de datos y se configuran los parámetros del algoritmo del siguiente modo: El umbral mínimo de soporte se establecen en 0.1, el tipo de métrica se escogerá la confianza, como se ha señalado en la teoría ya que es la más utilizada en los algoritmos de asociación y se seleccionan 15 reglas para extraerse. Se ejecuta y se obtiene que el soporte mínimo ha tenido un valor de 0.5 agrupando un total de 120380 instancias, la confianza mínima ha sido de 0.9 y el número de iteraciones del algoritmo han sido un total de 10. Además también se extrae que se han encontrado 8 patrones de tamaño 1, 22 patrones de tamaño 2, 16 patrones de tamaño 3 y un patrón de tamaño 4. Una vez extraída toda esta información del algoritmo se procede a entender las mejore reglas de asociación que ha conseguido extraer el algoritmo.

La primera regla se refiere a que cuando se cometen delitos contra el patrimonio pero

no se cometen delitos contra la libertad de las personas (para 136955 instancias) entonces no se cometen delitos contra la libertad sexual (para 130261 instancias), con una confianza de 0.95≈1

Cuando se cometen delitos contra el patrimonio pero no se cometen delitos contra las relaciones familiares (para un total de 129754 instancias) entonces tampoco se cometen delitos contra la libertad sexual (para 123389 instancias), con una confianza de 0.95≈1.

Cuando se cometen delitos contra el patrimonio (para un total de 172957 instancias), entonces, no se cometerán delitos contra la libertad sexual (para 163749 instancias), con una confianza de 0.95≈1.

Cuando se cometen delitos contra el patrimonio pero no se cometen delitos contra la salud pública para un total de 140686 instancias entonces tampoco se cometen delitos contra la libertad sexual para 132943 instancias, con una confianza de 0.95≈1.

Si se es hombre y se ha cometido delito contra el patrimonio (para un total de 152397 instancias), entonces, no se cometerán delitos contra la libertad sexual (para 143580 instancias), con una confianza de 0.93≈1.

Si se es Español (para 148796 instancias), no se cometerá delitos contra la libertad sexual (para 140095 instancias) .

Si se es Español y hombre (para 131543 instancias), no se cometerá delitos contra la libertad sexual (para 123463 instancias) .

Si se es hombre y no se ha cometido delito contra las relaciones familiares (para 1138520 instancias), no se cometen delitos contra la libertad sexual (para 129287 instancias).

Si se es hombre y no se ha cometido delito contra las personas (para 167256 instancias) no se cometen delitos contra la libertad sexual (para 155996 instancias).

Ahora, se decide discretizar las variables cuantitativas y realizar el mismo algoritmo, eliminando las variables dicotomizadas, la variable grupo de edad y operando con las nuevas solamente. Los parámetros se mantienen igual que para el caso anterior. Se han encontrado 30 patrones de tamaño 1, 73 patrones de tamaño 2, 50 patrones de tamaño 3 y 10 patrón de tamaño 4 y 1 de tamaño 5. Algunas de las reglas más significativas obtenidas son las siguientes:

Si la edad a la última detención está comprendida en el grupo de 26.5 a 60.5 años y se ha cometido más de un delito contra las relaciones familiares, entonces se es hombre, con una confianza de 0.95≈1.

Si la edad a la primera detención está comprendida en el grupo de 15.5 a 26.5 años y se ha cometido algún delito contra las personas, entonces se es hombre, con una confianza de 0.94≈1.

Si tiene la nacionalidad Española y ha cometido delito contra las relaciones familiares menos de dos veces, entonces, se es hombre, con una confianza de 0.93≈1

Si tiene la nacionalidad Española y ha cometido al menos un delito contra las personas, entonces se es hombre, con una confianza de 0.93≈1

Es una técnica muy útil que permite encontrar relaciones entre las variables con un nivel de significación alto, que garantice su efectividad, que a primera vista no se observan, que puede facilitar a su vez a los servicios de investigación policial adecuar un plan de acción correcto que permita desarrollar el nivel de eficacia en sus actuaciones.

Las salidas del programa se pueden consultar en los anexos.

Clustering: Algoritmo k-medias

En primer lugar, se procede a seleccionar todas las variables con las que se quiere trabajar en este algoritmo, que, en este caso, serán todas las variables numéricas originales. Se excluirán las variables categóricas y dicotómicas que se crearon posteriormente. El siguiente paso será la selección del número de *cluster* con los que se va a trabajar que no se va a desarrollar de un modo aleatorio, sino que se va a determinar a partir del método del codo a través de R y se seleccionan 4. A partir de aquí, se ejecuta el algoritmo y se extrae la siguiente información.

De la tabla de salida, se puede interpretar que el primer grupo aglutina un total del 16 % de la información, el segundo, un 31 %, el tercero, un 24 % y, por último, el cuarto, formado por un 29 % de la información. También se observa que la mayoría de los detenidos son individuos masculinos de nacionalidad española comprendidos entre 19 y 30 años.

En el primer grupo, los individuos que predominan son los de nacionalidad latinoamericana, con una edad comprendida entre 19 y 30 años, varones tienen un número medio de detenciones de 5, la edad media a la primera detención es de 27 años, a la segunda, de 28, la edad media entre las detenciones es la más baja de todos los grupos con un valor de 382 días. En el segundo grupo, los individuos que predominan son los de nacionalidad española, con una edad comprendida entre 19 y 30 años, varones tienen un número medio de detenciones de 7, la edad media a la primera detención es de 24 años, a la segunda, de 33, la edad media entre las detenciones es de 916 días. En el tercer grupo, los individuos que predominan son los de nacionalidad española, con una edad igual o menor a 18 años, varones tienen un número medio de detenciones de 5, la edad media a la primera detención es de 16 años, a la segunda, de 24, la edad media entre las detenciones es de 683 días. En el cuarto grupo, los individuos que predominan son los de nacionalidad española, con una edad comprendida entre 31 y 50 años, varones tienen un número medio de detenciones de 5, la edad media a la primera detención es de 39 años, a la segunda, de 45, la edad media entre las detenciones es de 641 días.

Se sacan las siguientes conclusiones:

En el primer y segundo *cluster* hay una mayor cantidad de orientales, en el primero hay más chinos que en los otros dos. En los tres hay poca concentración de americanos.

En el caso de la variable para grupos de edad se observa que el primer *cluster* aglomera las edades de 31 a 50 y los mayores de 50 años, el segundo *cluster* agrupa las edades de 19 a 30 años, por último, el tercero, agrupa los menores de 18.

Según el número de detenciones, el primer *cluster* es el que presenta los delincuentes con un menor número de detenciones, hasta algo más de 68.5; el segundo, hasta casi 134

detenciones y el tercero, un valor entre los de los dos *cluster* anteriores.

La edad a la primera detención en el primer *cluster* abarca desde los 20 hasta los 90 y es el que mayor concentración tiene; el segundo va de 14 a 25, aproximadamente, y en el tercero de 14 a 18, aproximadamente.

La edad a la segunda detención pasa lo mismo que en el caso anterior la única diferencia es el segundo y tercer grupo alcanzan valores más altos.

Ahora en relación a las variables de criminalidad, el número máximo de delitos de orden público cometido por cada individuo es mayor en el tercer grupo siguiéndole el segundo y, por último, el primero. Para los delitos de patrimonio, sucede lo mismo que en el caso anterior. En los delitos contra las personas, el número máximo de delitos cometido por cada individuo en el grupo 2 es menor que en el grupo 3 y a su vez es menor que en el grupo 1. Para los delitos en las relaciones familiares sucede lo mismo que para los delitos contra el orden público. Por último, con relación a los delitos contra la salud pública el número máximo de delitos cometido por cada individuo, en el grupo 2 es mayor que en el grupo 3 y a su vez en este es mayor que en el grupo 1.

Será de utilidad para analizar el comportamiento en grupo de cada uno de los conglomerados y analizar las propiedades que tienen en común cada uno de ellos para adecuar las medidas oportunas necesarias por parte de las Fuerzas y Cuerpos de Seguridad del Estado.

Capítulo 6

Conclusiones

En este trabajo se ha pretendido demostrar la utilidades de las distintas técnicas estadísticas más adecuadas aplicadas sobre los datos que lo necesiten.

En primer lugar se ha conseguido desarrollar una librería de R que permite una fácil y rápida obtención de los gráficos de control alternativa al método de Excel, que permite interpretar los gráficos de un modo eficaz y al detalle.

En segundo lugar, en relación a la base de datos que se ha creado a partir de fuentes de carácter público, formada por una variable temporal, otra relacionada con la comunidad autónoma, cuatro más relacionadas con esta y cuatro variables de criminalidad, se ha observado que las técnicas estadísticas más adecuadas para la aplicación son; en primer lugar los gráficos de control, ya que existe una variable temporal, permiten evaluar si las variables de criminalidad o sus combinaciones se encuentran bajo control estadístico y, si no es así, aplicar las medidas correctivas necesarias para que en los próximos años lo estén o anticiparse cuando los valores se encuentran sobrepasando los límites de alerta para que no rebasen los límites de control. En segundo lugar, también se cree conveniente la utilidad de un análisis con números índice simples que involucren el tiempo con las variables de criminalidad y de localidad y puedan hacerse comparaciones entre sí y, por otro lado, números índices compuestos, tanto sin ponderar, involucrando los ámbitos de calidad, como ponderados, a partir de la comunidad autónoma y las variables relacionadas con la comunidad autónoma que permiten realizar una ponderación. Lo que va a permitir una comparación posterior entre los períodos que se consideran de interés y las "localizaciones" que han sido seleccionadas.

Por otro lado, según la base de datos del Cuerpo Nacional de Policía, al tener relación directa todas las variables, lo primero que se ha considerado hacer es un análisis descriptivo para obtener un resumen de la información que se va a encontrar y, así, saber a lo que se va a enfrentar uno a la hora de analizar los datos y tener una visión general. A continuación, se cree conveniente realizar una serie de contrastes de hipótesis para responder a preguntas que es frecuente que se planteen y puedan aportar más información sobre los datos que se dispone. Por último, se considera de utilidad aplicar distintas técnicas de minería de datos de asociación, para encontrar relaciones entre las variables que a priori no se ven, clasificación, a través de árboles de decisión que permiten seccionar los datos para encontrar patrones que permitan clasificar los mismos y obtener una información que ni puede ser observada a simple vista y, por último, los métodos de *clustering* que permiten dividir en grupos los datos según una serie de características comunes que compartan.

Toda la información extraída de estos métodos será de gran utilidad para las autoridades de las Fuerzas y Cuerpos de Seguridad del Estado encargadas de realizar los protocolos para reducir o mantener bajo control la criminalidad o diseñar planes prácticos para erradicar esta misma y todo gracias a toda la información que se obtiene a partir de la aplicación de los métodos estadísticos ya sea directa o indirectamente.

Capítulo 7

Abstract

Scientific research applied to the criminalistic field group a very wide set of multidisciplinary techniques that cover very diverse scientific fields (such as chemistry, medicine, computer science, mathematics, etc.) aimed at different purposes within the same field. In this case, statistics will be used to take stock of the value of statistical techniques for the study of crime. The statistics will serve as a tool that will allow extracting and understanding information about the data with which they work in order to provide information to the members of the "Fuerzas y Cuerpos de seguridad del Estado", leaving it to the statistical researchers to decide to apply the most appropriate techniques according to the data available to optimize the quality of the results.

Although the main objective of this research is not to carry out an investigation on some data to draw conclusions about them, but rather it is about assessing and demonstrating the degree of usefulness of different statistical techniques when making a correct diagnosis of crime, adapting them to the type of data that is available to be analyzed, since it will not always be possible to apply the same techniques on one type of data as on others, in addition to the design and implementation in code of techniques for its analysis; it has been necessary to follow the majority of the points mentioned above when carrying out the work. Conclusions about what the results of the statistical analyses provide would take a back seat. In all the work, the process described above has been followed to a greater or lesser extent since for each set of data or rather, for each phase of the work, a problem to be solved has been established, based on previous knowledge, a series of tools have been selected such as statistical techniques themselves and the *software* that have been considered appropriate for each of the two phases In addition to other knowledge that has been used, once this has been available, it has been tried to give light to these objectives through a reflective thought, it has been decided which are the methods that best suit each case.

The application of statistical techniques in the work has been carried out in two phases as noted above. In the first, which corresponds to the application of techniques for the diagnosis of crime in data of public origin, previous knowledge was available on the statistical techniques that were planned to be applied, of the R programming language on which the tools that allow analyzing the data have been designed. In the second phase, referring to the study of the data of the National Police Corps, the previous knowledge available is included in all the statistical tools, the *software* with which it is going to work, in this case, with SPSS and Weka and the concepts related to crime that must be taken into account to understand the data available. From this point we will try to solve the objectives set by assessing from a reflective thought what will be the most appropriate

techniques to apply on the data, in this case, descriptive statistical techniques, hypothesis contrasts and a classification tree, after having developed the pre-operational study and having the conviction that the techniques that will be applied on the data will be appropriate.

A summary of the statistical techniques that have been used in this research are the next:

Measures of central tendency bring together all the statisticians in charge of providing all the information about the central values of the sample that is available, and are representative of the whole of it. Dispersion measures that refer to a series of statisticians from which information about the variability of the available data is extracted. The measures of form are statistical that refer to the description of the distribution followed by the data available without the need for their graphical representation. Bar chart, graph formed by the number of rectangles corresponding to each category or element of the study variable and height corresponding to the frequency of each of the categories on an axis of Cartesian coordinates. Inferential statistics techniques:

Contrasts of hypothesis, statistical method belonging to the inferential statistics in which there are two hypotheses, the first, denoted by "H0" is known as Null Hypothesis, the starting point of the contrast, which will be accepted as long as there is not enough evidence not to do so and "H1" is the Alternative Hypothesis to the null which will be accepted in the event that there is significant evidence that allows rejecting the Null Hypothesis.

Control charts:

are statistical tools that are used for the analysis of qualitative or quantitative data and take into account the temporal order of events. They are used to define a goal for an operation you want to perform, to help achieve it and fundamentally to determine if that goal has been achieved or not.

Index numbers:

"abstract statistical measure (without unity), designed to show the changes of a variable or group of variables with respect to time, geographical location or other characteristic" (LÓPEZ MARTÍN, 1982). The initial period is called the base or reference period, always representing the value of 1 or 100 % if it is counted as a percentage and the period to be compared is called the current period.

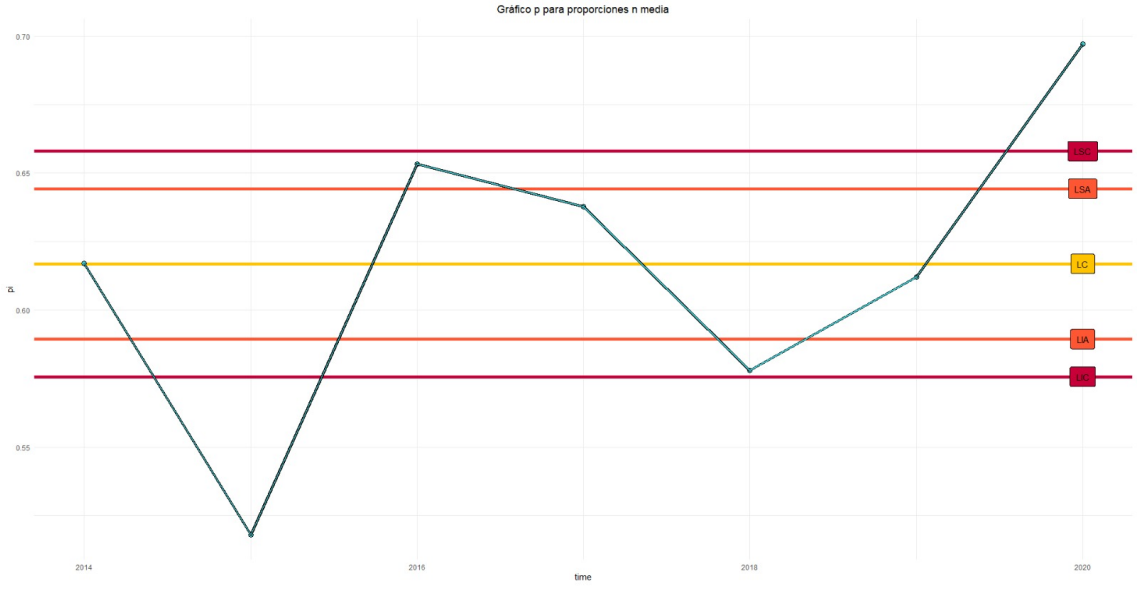
Data mining techniques:

Data mining is a branch of statistics that is formed by a set of multifunctional statistical techniques that are used to find patterns, relationships, behaviors or make predictions in the data that cannot be extracted from them with the naked eye and that will allow to draw conclusions of interest to the researcher. These algorithms can be classified as supervised, when their objective is to make a prediction from the data available and the unsupervised algorithms, whose objective is to discover relationships and patterns within the data set that is available.

For the analysis of the data available and for future data analysis, in the specific case of control charts, a specific library has been designed in R to have an alternative tool to Excel and thus favor a quick and effective interpretation of them.

In the case of index numbers, functions and small programs have been developed in R more specific for the data available since they are oriented almost exclusively to the characteristics of the data that are being handled in this database, although they would also be extensible to the analysis of data that had a disposition and properties similar to those that are being worked.

The objective of the creation of the library is that it is possible to have the set of necessary tools that allow to study and interpret the Control Graphs in the field of criminalistics. This library has been based on the design and implementation of a total of four R functions that are responsible for making the most used control graphs in the field of statistical crime analysis. These functions can be divided into two groups. According to the first, the function that composes it corresponds to the control graph for the number of cases. The second group groups a total of three functions that make up the control graphs of the proportions, firstly, in the case that the samples have the same size, secondly for when the samples have different sizes, but have little variability and finally, the graph for standardized values, used, when the variability is greater than the previous case. For the design in the code for the operation of this library, the libraries **dplyr** and **ggplot2** have been used in turn.



In this research, two data stores will be used. One of them, built through publicly accessible data and the other consists of a database of the National Police Corps. On which due to their characteristics some or other statistical methods will be applied since the information they offer is different.

The first database consists of a total of 12 variables and 666 entries. With the variables “year”, “community”, “ambito”, “detinvest”, “victimizations”, “known”, “enlightened’, “population”, “km2”, “denspob” and “pobpr”.

Once it has been constructed , it concludes that for crimes of ideology, the missing

data are lost data that are not collected, and in the case of crimes based on sex it is concluded that they are collected, since there are cases in which a crime appears and when looking at the following years it can be observed that there are also few crimes for the field of sex for the following years, so it is concluded that in this case the data are not missing. In the case of hate crimes based on ideology in 2014, the decision will be made to replace them with the average of hate crimes based on ideology in each autonomous community in the following years. Sex-based hate crimes are left as they are. Once this procedure is complete, the dataset is available for use.

According to the second database, there are the variables: "country", "age", "sex", "number of detentions", "Age1", "Age2nd", "mediaentre", "sexual freedom", "public order", "heritage", "people", "family relations" and "public health".

After, as many dichotomous variables have been created as quantitative variables referring to the type of crime there is. With the aim of indicating with a 1 in the case that the crime has been committed and 0 in the case of not having been committed, to be able to use them in the data mining processes that it is decided to use.

According to the dataset that has been created from public sources, consisting of a temporary variable, another related to the autonomous community, four more related to that autonomous community and four crime variables, it has been observed that the most appropriate statistical techniques for the application are, in the first place, the control graphs, since there is a temporal variable, which allow to evaluate if the crime variables or their combinations are under statistical control and if not, apply the necessary corrective measures so that in the coming years they are or anticipate when the values are exceeding the alert limits so that they do not exceed the control limits. Secondly, it is also considered convenient the usefulness of an analysis with simple index numbers calculated, with a reference period established in the average of all the values for each of the areas studied, with the aim of penalizing the years in which no data is available either because they have not been collected, or because they are not available by that date, while still considering them, promoting in this way, a useful way for its interpretation without having to take other more chaotic measures that prevent this process from being carried out quickly and efficiently, such as the alternative that was considered, to begin to take as a reference the year for which data were available, but finally this possibility was ruled out since the complete time route and the causal particularities for each of the cases were not considered. that involve time with the variables of crime, and locality and comparisons can be made with each other and on the other hand, composite index numbers, non pondered, to study through them the evolution of the variable of known hate crime cases taking into account the areas of criminality for hate crimes in the regions that are considered appropriate, in this case the whole of Spain will be taken as a reference and subsequently compared with Madrid, during periods of interest that will be considered the year of the beginning of the data collection and the last year for which data are available, that is, 2020 and on the other hand considering the average of the index numbers during all the years with respect to the year of the pandemic. Ad pondered, considering the Autonomous Communities and making this weighting through a series of variables that are referred to them such as the population in each community, or the population density, as before, comparing in the periods of interest that have been used so far. It is very useful to consider this option in this case since each Autonomous Community has characteristics that can be measured and determined in an objective way, they can be used as a weighting factor. Both unweighted, involving the areas of quality, and weighted, from the autonomous community and the

variables related to the autonomous community that allow a weighting. This will allow a later comparison between the periods that are considered of interest and the "locations" that have been selected.

On the other hand, for the database of the National Police Corps, having a direct relationship of all the variables, the first thing that has been considered to do is a descriptive analysis to obtain a summary of the information that will be found and also know what one is going to face when analyzing the data and having an overview. Next, it is considered convenient to make a series of contrasts of hypotheses to answer questions that are frequent that are raised and can provide more information about the data that are available. Finally, it is considered useful to apply different techniques of association data mining, to find relationships between the variables that which are not seen, classification, through decision trees that allow sectioning the data to find patterns that allow them to be classified and obtain information that can not be observed with the naked eye, and finally, the methods of *clustering* that allow data to be divided into groups according to a series of common characteristics that they share.

All the information extracted from these methods will be very useful for the authorities of the National Police Corps in charge of carrying out the protocols to reduce or keep under control crime or design practical plans to eradicate this same and all thanks to all the information obtained from the application of statistical methods either directly or indirectly.

Capítulo 8

Bibliografía

Referencias

ALBAJAR, R. A., y MARTÍN, Q. M. (2006). *Estadística* (Vol. 4). CISE.

Aplicación de minería de datos para la exploración y detección de patrones delictivos en argentina, author=F.Valenga, I.Perversi, E.Fernández, H.Merino, D.Rodríguez, P.Britos, R.García-Martínez, journal=XIII Congreso Argentino de Ciencias de la Computació, pages=13. (s.f.).

CABERO MORÁN María Teresa, P. C. E., y Quintín, M. M. (s.f.). *Bases científicas de la investigación y diagnosis de la criminalidad*. CNP.

Calón, E. C. (1949). Los nuevos métodos científicos de investigación criminal y los derechos de la persona. *Anuario de Derecho Penal y Ciencias Penales*(1), 37–54.

Díaz, J. M. R. (2021). Linear models. *Apuntes de Modelos Lineales*.

en Percepción Remota y Sistemas de Información Espacial, S. L. (2014). Memorias. *XVI Simposio Internacional*, 10.

Korting, T. S. (2006). C4. 5 algorithm and multivariate decision trees. *Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil*, 22.

LÓPEZ MARTÍN, L. (1982). Los números índices. *Revista de didáctica de las matemáticas*(3), 81–96.

Ministerio del Interior, G. d. E. (s.f.). Hechos conocidos. *Metadata conocidos portal estadístico de criminalidad*, 5.

Perversi, I. (2007). Aplicación de minería de datos para la exploración y detección de patrones delictivos en argentina. , 117.

Singh, J., Ram, H., y Sodhi, D. J. (2013). Improving efficiency of apriori algorithm using transaction reduction. *International Journal of Scientific and Research Publications*, *3*(1), 1–4.

Torche, A. (1998). *Contabilidad nacional, números índices: desestacionalización y trimestralización*. Pontificia Universidad Católica de Chile, Instituto de Economía, Oficina de ....

Capítulo 9

Anexos I: Salidas de SPSS

país

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	?	363	,2	,2	,2
	ÁFRICA	9026	3,7	3,7	3,9
	AMÉRICA DEL NORTE; CENTRAL E INSULAR	104	,0	,0	3,9
	CHINA	936	,4	,4	4,3
	ESPAÑOLA	148796	61,8	61,8	66,1
	EUROPA (MENOS E)	7334	3,0	3,0	69,2
	EUROPA DEL ESTE	23798	9,9	9,9	79,1
	LATINOAMÉRICA	25190	10,5	10,5	89,5
	MARRUECOS	18383	7,6	7,6	97,2
	ORIENTE	465	,2	,2	97,4
	ORIENTE PRÓXIMO	6365	2,6	2,6	100,0
	Total	240760	100,0	100,0	

Figura 9.1: Frecuencias país

edad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	<=18	56651	23,5	23,5	23,5
	> 50	6041	2,5	2,5	26,0
	19 - 30	113105	47,0	47,0	73,0
	31 - 50	64963	27,0	27,0	100,0
	Total	240760	100,0	100,0	

Figura 9.2: Frecuencias edad

sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	?	663	,3	,3	,3
	H	213634	88,7	88,7	89,0
	M	26463	11,0	11,0	100,0
	Total	240760	100,0	100,0	

Figura 9.3: Frecuencias sexo

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. Desviación	Varianza	Asimetría		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Desv. Error	Estadístico	Desv. Error
nºdetenciones	240760	3	134	5,80	4,624	21,378	4,320	,005	35,574	,010
edad1ª	240760	14	90	26,61	9,868	97,380	1,144	,005	1,420	,010
edad2ª	240760	14	91	33,73	10,720	114,924	,707	,005	,372	,010
mediaentre	240760	,076923077	9287,500000	695,7568436	713,4113106	508955,698	2,554	,005	10,816	,010
libertadsexual	15931	1	13	1,28	,714	,509	4,108	,019	27,977	,039
lib_sex_bin	240760	0	1	,07	,249	,062	3,491	,005	10,184	,010
ordenpúblico	70419	1	28	1,55	1,049	1,100	4,029	,009	34,484	,018
orde_púb_bin	240760	0	1	,29	,455	,207	,912	,005	-1,168	,010
patrimonio	172957	1	92	3,46	3,749	14,052	4,577	,006	39,377	,012
patrimonio_bin	240760	0	1	,72	,450	,202	-,971	,005	-1,057	,010
personas	49880	1	11	1,31	,683	,467	3,074	,011	14,012	,022
personas_bin	240760	0	1	,21	,405	,164	1,445	,005	,088	,010
relacionesfamiliares	80154	1	25	1,77	1,136	1,290	2,478	,009	12,353	,017
rel_fam_bin	240760	0	1	,33	,471	,222	,709	,005	-1,497	,010
saludpública	47807	1	22	1,67	1,189	1,413	3,236	,011	18,671	,022
salud_púb_bin	240760	0	1	,20	,399	,159	1,511	,005	,284	,010
N válido (por lista)	106									

Figura 9.4: Estadísticos descriptivos

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
nºdetenciones	Entre grupos	60231,351	9	6692,372	316,473	,000
	Dentro de grupos	5083405,550	240387	21,147		
	Total	5143636,901	240396			
edad1ª	Entre grupos	307593,635	9	34177,071	355,570	,000
	Dentro de grupos	23105798,80	240387	96,119		
	Total	23413392,43	240396			
edad2ª	Entre grupos	647101,587	9	71900,176	640,591	,000
	Dentro de grupos	26981119,28	240387	112,240		
	Total	27628220,87	240396			

Figura 9.5: Tabla ANOVA 1

Comparaciones múltiples

Variable dependiente		(I) país	(J) país	Diferencia de medias (I-J)	Desv. Error	Sig.	Intervalo de confianza al 95% Límite inferior	Intervalo de confianza al 95% Límite superior
nºdetenciones	HSD Tukey	AFRICA	AMÉRICA DEL NORTE; CENTRAL O INSULAR	,841	,454	,700	-,59	2,28
			CHINA	,613*	,158	,004	,11	1,11
			ESPAÑA	-,935*	,050	,000	-1,09	-,78
			EUROPA	-,135	,072	,694	-,36	,09
			EUROPA DEL ESTE	-,131	,057	,386	-,31	,05
			LATINOAMÉRICA	,415*	,056	,000	,24	,59
			MARRUECOS	-,202*	,059	,022	-,39	-,02
			ORIENTE	,742*	,219	,024	,05	1,43
			ORIENTE PRÓXIMO	-1,188*	,075	,000	-1,43	-,95
		AMÉRICA DEL NORTE; CENTRAL O INSULAR	AFRICA	-,841	,454	,700	-2,28	,59
			CHINA	-,228	,475	1,000	-1,73	1,28
			ESPAÑA	-1,775*	,451	,003	-3,20	-,35
			EUROPA	-,975	,454	,493	-2,41	,46
			EUROPA DEL ESTE	-,972	,452	,492	-2,40	,46
			LATINOAMÉRICA	-,425	,452	,995	-1,85	1,00
			MARRUECOS	-1,043	,452	,384	-2,47	,39
			ORIENTE	-,099	,499	1,000	-1,68	1,48
			ORIENTE PRÓXIMO	-2,029*	,455	,000	-3,47	-,59
		CHINA	AFRICA	-,613*	,158	,004	-1,11	-,11
			AMÉRICA DEL NORTE; CENTRAL O INSULAR	,228	,475	1,000	-1,28	1,73
			ESPAÑA	-1,548*	,151	,000	-2,02	-1,07
			EUROPA	-,748*	,160	,000	-1,25	-,24
			EUROPA DEL ESTE	-,744*	,153	,000	-1,23	-,26
			LATINOAMÉRICA	-,198	,153	,956	-,68	,29
			MARRUECOS	-,815*	,154	,000	-1,30	-,33
			ORIENTE	,128	,261	1,000	-,70	,95
			ORIENTE PRÓXIMO	-1,802*	,161	,000	-2,31	-1,29
		ESPAÑA	AFRICA	,935*	,050	,000	,78	1,09
			AMÉRICA DEL NORTE; CENTRAL O INSULAR	1,775*	,451	,003	,35	3,20
			CHINA	1,548*	,151	,000	1,07	2,02
			EUROPA	,800*	,055	,000	,63	,97
			EUROPA DEL ESTE	,804*	,032	,000	,70	,91
			LATINOAMÉRICA	1,350*	,031	,000	1,25	1,45
			MARRUECOS	,732*	,036	,000	,62	,85
			ORIENTE	1,676*	,214	,000	1,00	2,35
			ORIENTE PRÓXIMO	-,254*	,059	,001	-,44	-,07
		EUROPA	AFRICA	,135	,072	,694	-,09	,36
			AMÉRICA DEL NORTE; CENTRAL O INSULAR	,975	,454	,493	-,46	2,41
			CHINA	,748*	,160	,000	,24	1,25
			ESPAÑA	-,800*	,055	,000	-,97	-,63
			EUROPA DEL ESTE	,004	,061	1,000	-,19	,20
			LATINOAMÉRICA	,550*	,061	,000	,36	,74
			MARRUECOS	-,068	,064	,988	-,27	,13
			ORIENTE	,876*	,220	,003	,18	1,57
			ORIENTE PRÓXIMO	-1,054*	,079	,000	-1,30	-,80

	EUROPA DEL ESTE	AFRICA	,131	,057	,386	-,05	,31
		AMÉRICA DEL NORTE; CENTRAL O INSULAR	,972	,452	,492	-,46	2,40
		CHINA	,744*	,153	,000	,26	1,23
		ESPAÑA	-,804*	,032	,000	-,91	-,70
		EUROPA	-,004	,061	1,000	-,20	,19
		LATINOAMÉRICA	,546*	,042	,000	,41	,68
		MARRUECOS	-,071	,045	,858	-,21	,07
		ORIENTE	,873*	,215	,002	,19	1,55
		ORIENTE PRÓXIMO	-1,057*	,065	,000	-1,26	-,85
	LATINOAMÉRICA	AFRICA	-,415*	,056	,000	-,59	-,24
		AMÉRICA DEL NORTE; CENTRAL O INSULAR	,425	,452	,995	-1,00	1,85
		CHINA	,198	,153	,956	-,29	,68
		ESPAÑA	-1,350*	,031	,000	-1,45	-1,25
		EUROPA	-,550*	,061	,000	-,74	-,36
		EUROPA DEL ESTE	-,546*	,042	,000	-,68	-,41
		MARRUECOS	-,618*	,045	,000	-,76	-,48
		ORIENTE	,326	,215	,886	-,35	1,01
		ORIENTE PRÓXIMO	-1,604*	,065	,000	-1,81	-1,40
	MARRUECOS	AFRICA	,202*	,059	,022	,02	,39
		AMÉRICA DEL NORTE; CENTRAL O INSULAR	1,043	,452	,384	-,39	2,47
		CHINA	,815*	,154	,000	,33	1,30
		ESPAÑA	-,732*	,036	,000	-,85	-,62
		EUROPA	,068	,064	,988	-,13	,27
		EUROPA DEL ESTE	,071	,045	,858	-,07	,21
		LATINOAMÉRICA	,618*	,045	,000	,48	,76
		ORIENTE	,944*	,216	,001	,26	1,63
		ORIENTE PRÓXIMO	-,986*	,067	,000	-1,20	-,77

	ORIENTE	AFRICA	-,742*	,219	,024	-1,43	-,05
		AMÉRICA DEL NORTE; CENTRAL O INSULAR	,099	,499	1,000	-1,48	1,68
		CHINA	-,128	,261	1,000	-,95	,70
		ESPAÑA	-1,676*	,214	,000	-2,35	-1,00
		EUROPA	-,876*	,220	,003	-1,57	-,18
		EUROPA DEL ESTE	-,873*	,215	,002	-1,55	-,19
		LATINOAMÉRICA	-,326	,215	,886	-1,01	,35
		MARRUECOS	-,944*	,216	,001	-1,63	-,26
		ORIENTE PRÓXIMO	-1,930*	,221	,000	-2,63	-1,23
	ORIENTE PRÓXIMO	AFRICA	1,188*	,075	,000	,95	1,43
		AMÉRICA DEL NORTE; CENTRAL O INSULAR	2,029*	,455	,000	,59	3,47
		CHINA	1,802*	,161	,000	1,29	2,31
		ESPAÑA	,254*	,059	,001	,07	,44
		EUROPA	1,054*	,079	,000	,80	1,30
		EUROPA DEL ESTE	1,057*	,065	,000	,85	1,26
		LATINOAMÉRICA	1,604*	,065	,000	1,40	1,81
		MARRUECOS	,986*	,067	,000	,77	1,20
		ORIENTE	1,930*	,221	,000	1,23	2,63

Figura 9.6: Contrastes a posteriori paises

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
nºdetenciones	Entre grupos	185945,474	3	61981,825	3007,918	,000
	Dentro de grupos	4961072,248	240756	20,606		
	Total	5147017,722	240759			
edad1[a]	Entre grupos	20054458,19	3	6684819,395	474657,464	,000
	Dentro de grupos	3390677,489	240756	14,083		
	Total	23445135,67	240759			
edad2[a]	Entre grupos	15429584,24	3	5143194,748	101170,424	,000
	Dentro de grupos	12239298,22	240756	50,837		
	Total	27668882,46	240759			

Figura 9.7: Tabla ANOVA 2

Comparaciones múltiples

Variable dependiente		(I) edad	(J) edad	Diferencia de medias (I-J)	Desv. Error	Sig.	Intervalo de confianza al 95% Límite inferior	Límite superior
nºdetenciones	HSD Tukey	<=18	19 - 30	1,584*	,023	,000	1,52	1,64
			31 -50	2,332*	,026	,000	2,27	2,40
			> 50	2,985*	,061	,000	2,83	3,14
		19 - 30	<=18	-1,584*	,023	,000	-1,64	-1,52
			31 -50	,748*	,022	,000	,69	,81
			> 50	1,401*	,060	,000	1,25	1,55
		31 -50	<=18	-2,332*	,026	,000	-2,40	-2,27
			19 - 30	-,748*	,022	,000	-,81	-,69
			> 50	,653*	,061	,000	,50	,81
		> 50	<=18	-2,985*	,061	,000	-3,14	-2,83
			19 - 30	-1,401*	,060	,000	-1,55	-1,25
			31 -50	-,653*	,061	,000	-,81	-,50
	Bonferroni	<=18	19 - 30	1,584*	,023	,000	1,52	1,65
			31 -50	2,332*	,026	,000	2,26	2,40
			> 50	2,985*	,061	,000	2,82	3,15
		19 - 30	<=18	-1,584*	,023	,000	-1,65	-1,52
			31 -50	,748*	,022	,000	,69	,81
			> 50	1,401*	,060	,000	1,24	1,56
		31 -50	<=18	-2,332*	,026	,000	-2,40	-2,26
			19 - 30	-,748*	,022	,000	-,81	-,69
			> 50	,653*	,061	,000	,49	,81
		> 50	<=18	-2,985*	,061	,000	-3,15	-2,82
			19 - 30	-1,401*	,060	,000	-1,56	-1,24
			31 -50	-,653*	,061	,000	-,81	-,49

Figura 9.8: Contrastes a posteriori

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
nºdetenciones	Entre grupos	7581,275	1	7581,275	354,558	,000
	Dentro de grupos	5133783,843	240095	21,382		
	Total	5141365,118	240096			
edad1ª	Entre grupos	22412,332	1	22412,332	230,303	,000
	Dentro de grupos	23365207,31	240095	97,317		
	Total	23387619,64	240096			
edad2ª	Entre grupos	2,096	1	2,096	,018	,893
	Dentro de grupos	27595828,77	240095	114,937		
	Total	27595830,87	240096			

Figura 9.9: Tabla ANOVA 3

ANOVA

nºdetenciones

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	45359,980	1	45359,980	2140,633	,000
Dentro de grupos	5101657,742	240758	21,190		
Total	5147017,722	240759			

Figura 9.10: Tabla ANOVA 4

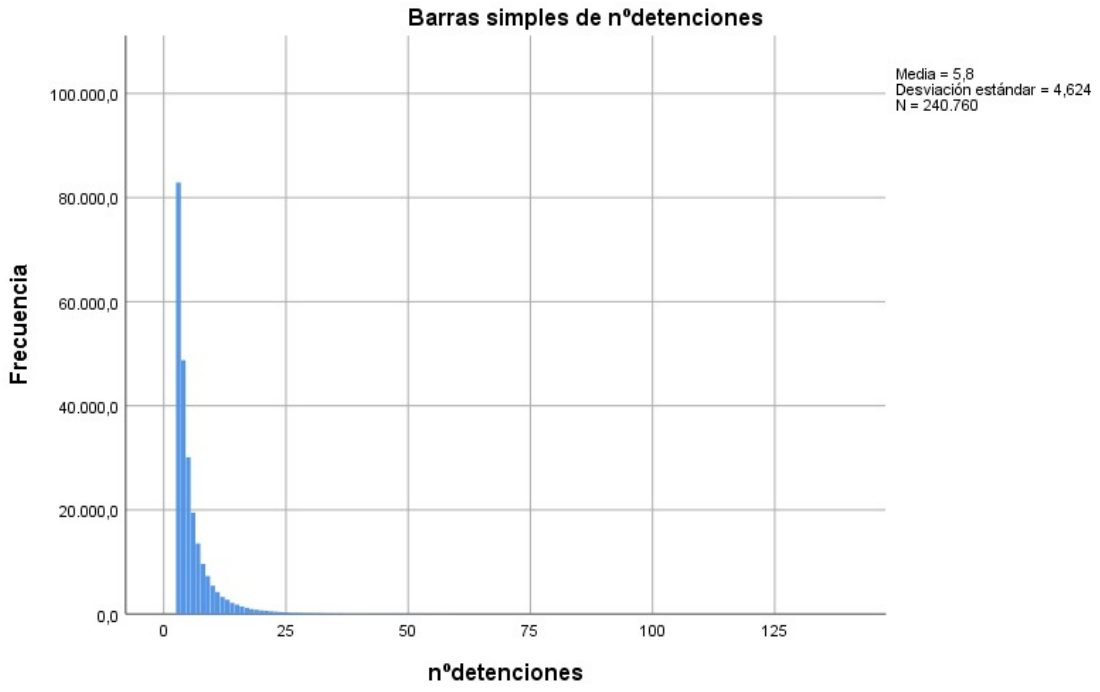


Figura 9.11: Distribuccion detenciones

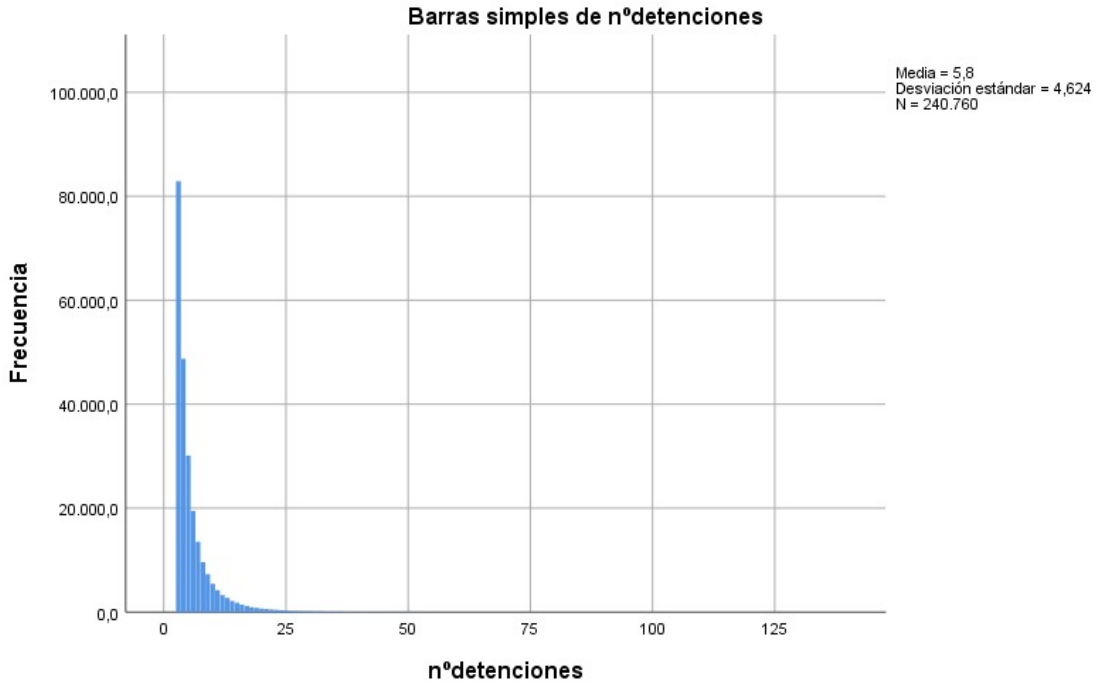


Figura 9.12: Distribuccion detenciones

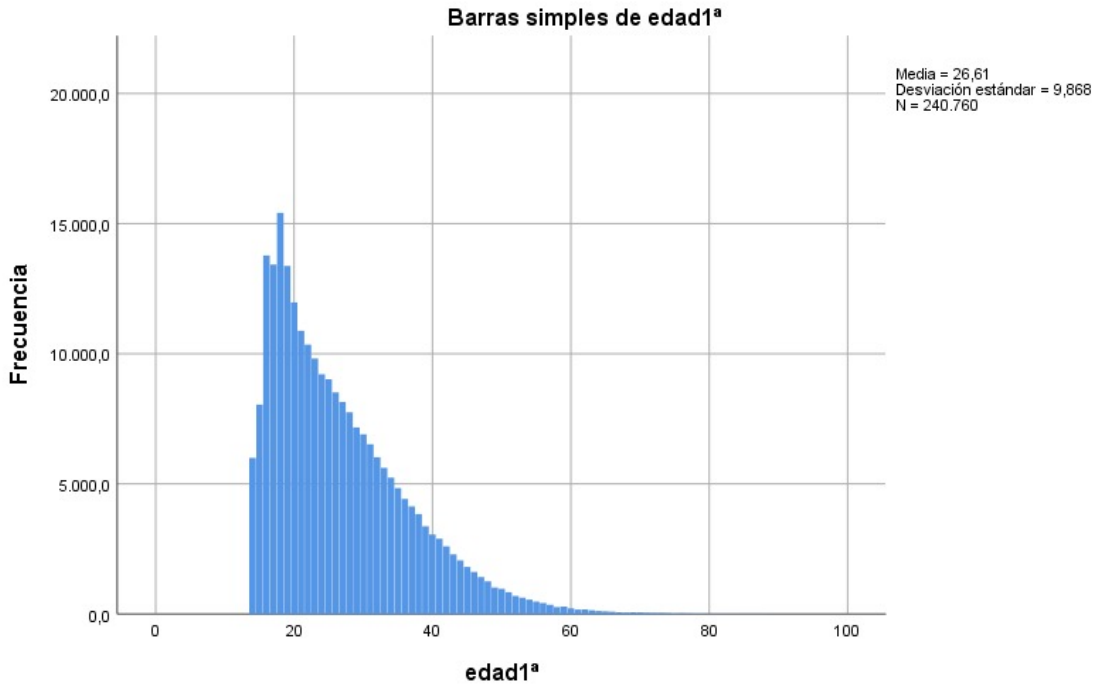


Figura 9.13: Distribución edad1

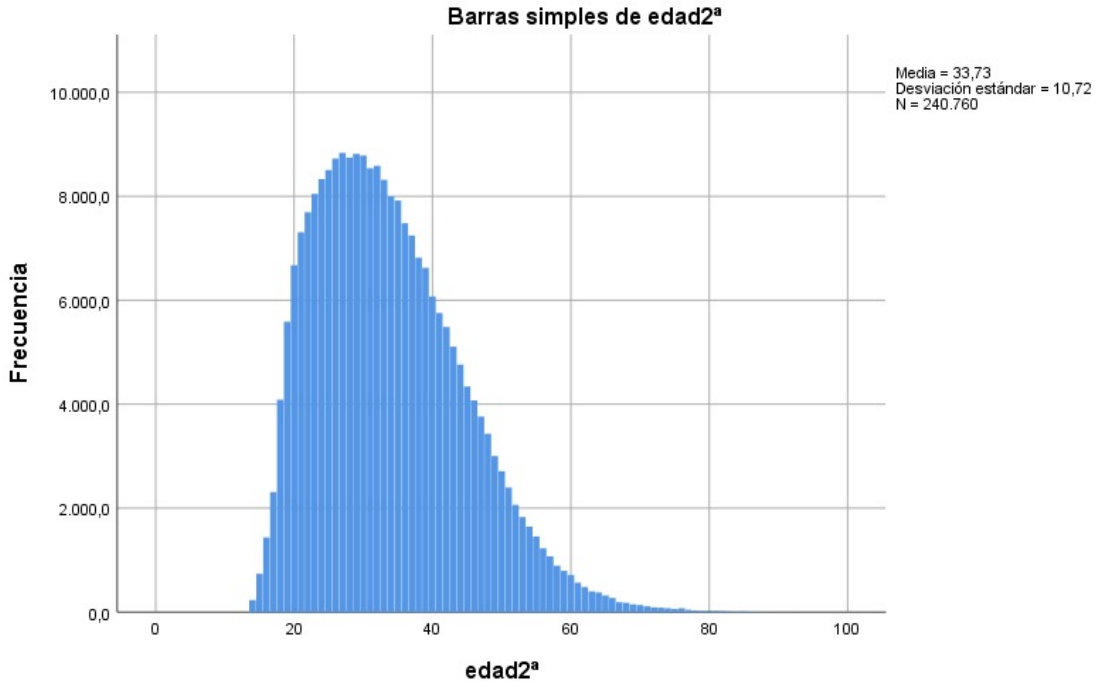


Figura 9.14: Distribución edad2

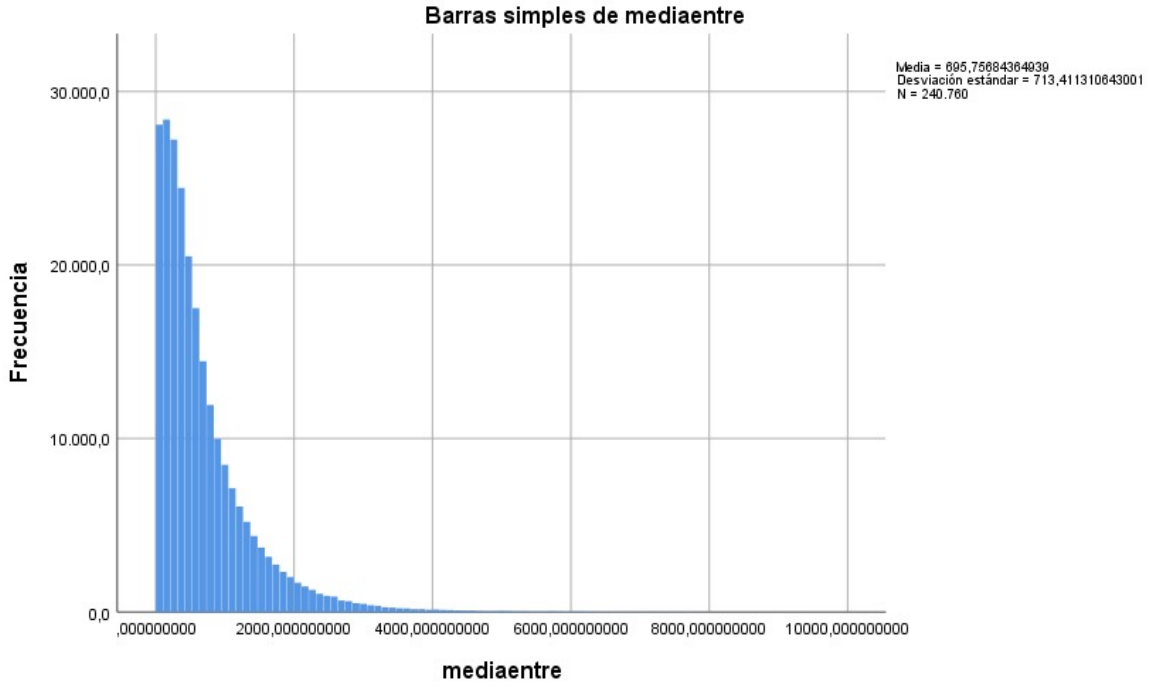


Figura 9.15: Distribución media entre

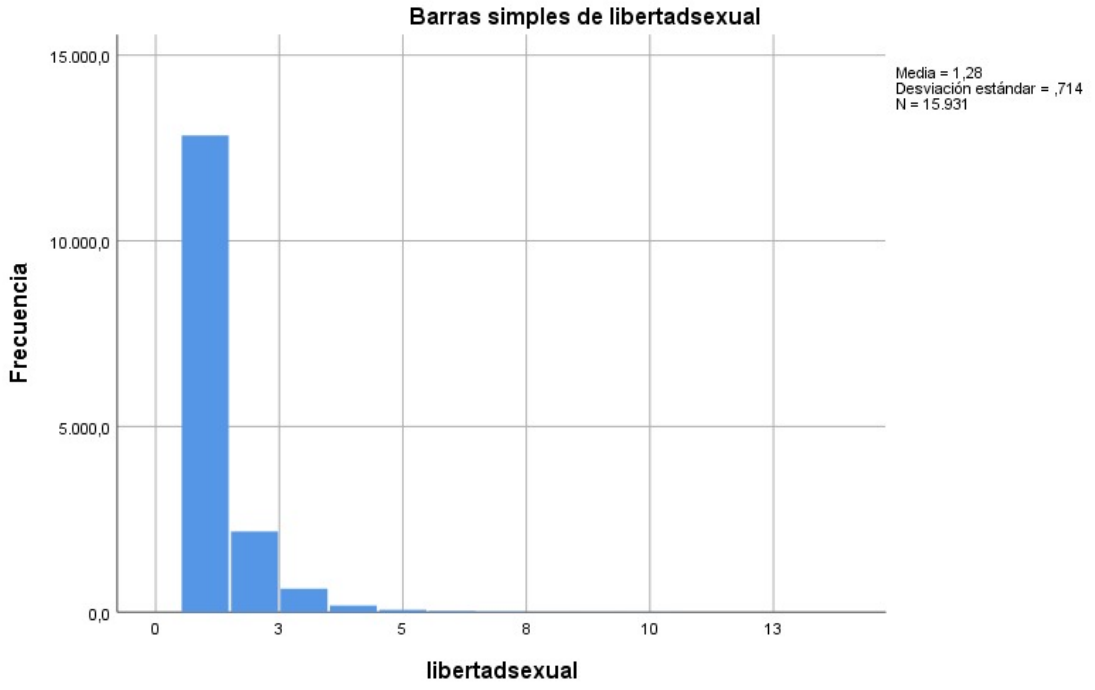


Figura 9.16: Distribución delitos libertad sexual

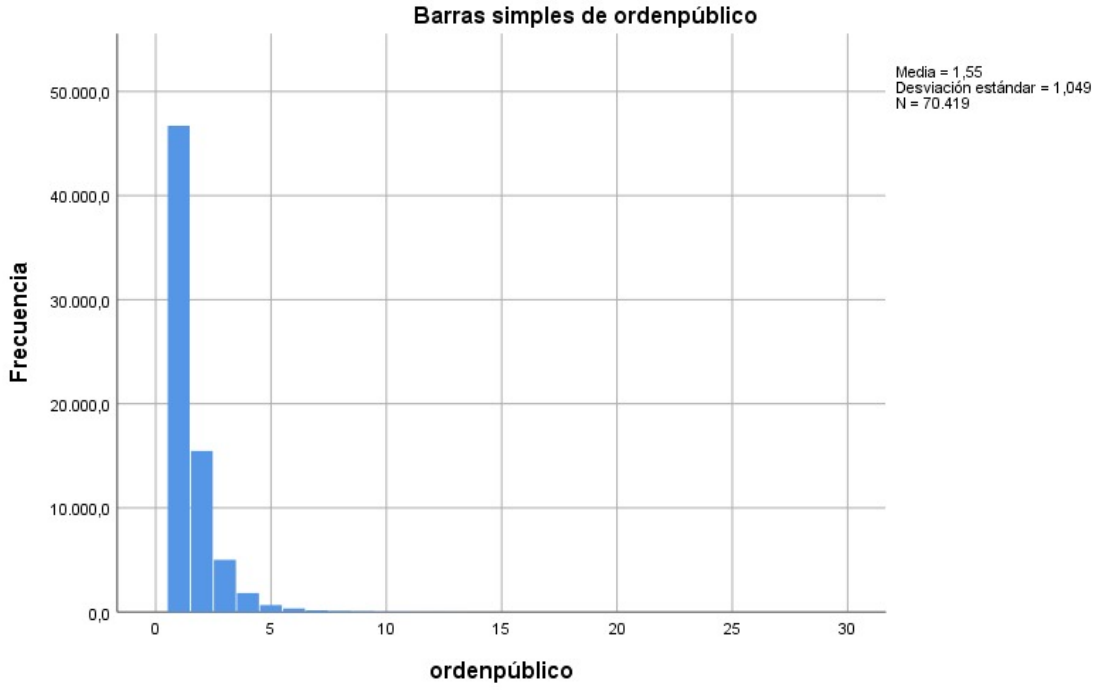


Figura 9.17: Distribución delitos contra el orden público

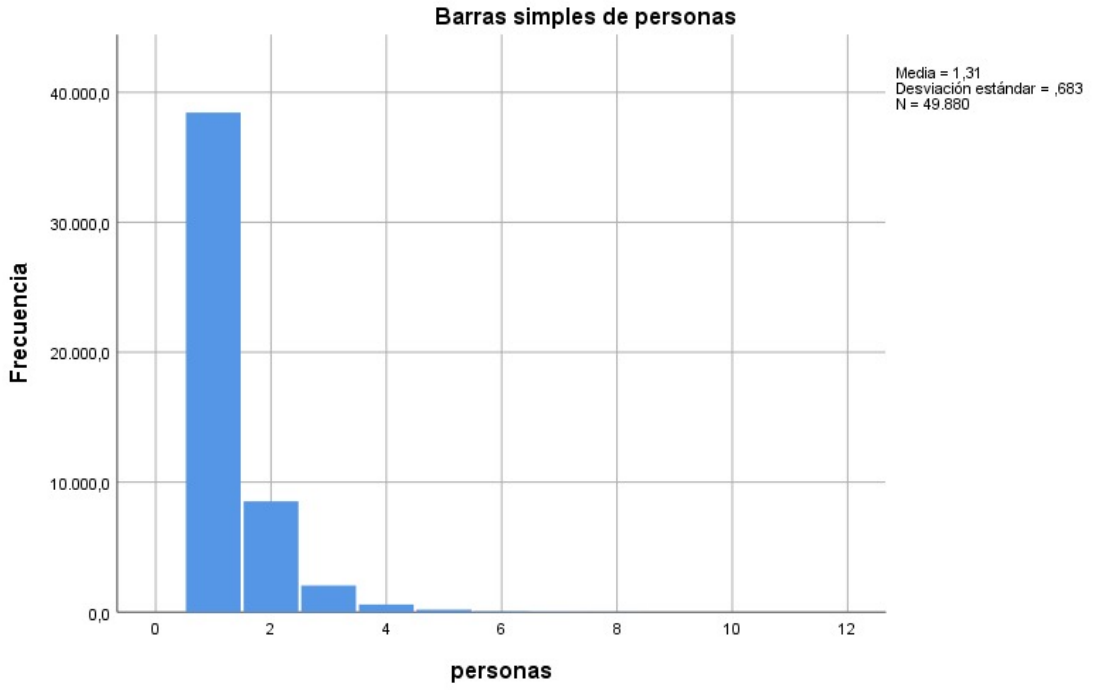


Figura 9.18: Distribución delitos contra la libertad de las personas

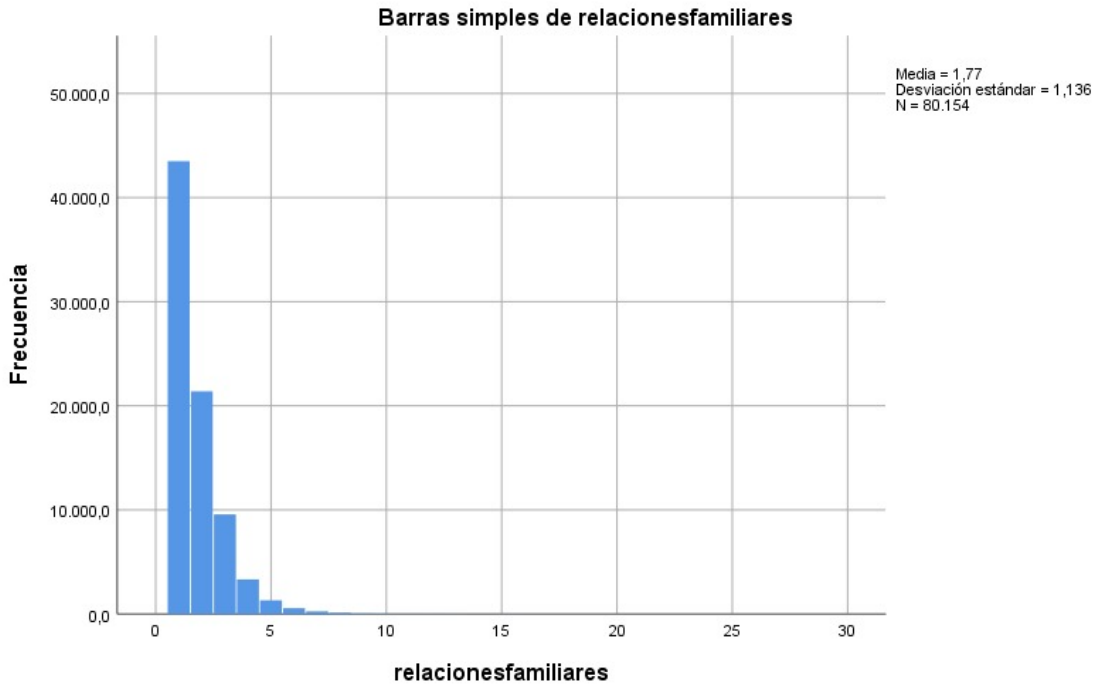


Figura 9.19: Distribución delitos contra las relaciones familiares

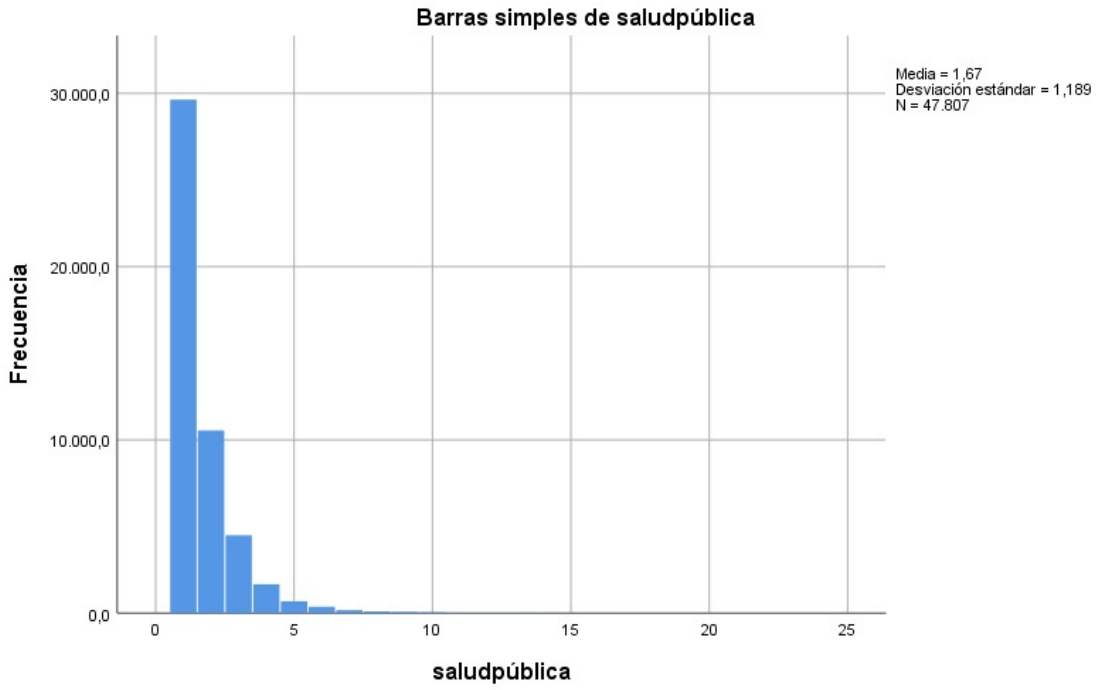


Figura 9.20: Distribución delitos contra la salud pública

Capítulo 10

Anexos II: Código de R

```
## Gráfico C
gcc <- function(df, frq, temp, values, time){
  c <- sum(values)/(temp*frq)
  LSC <- c+3*sqrt(c)
  LIC <- c-3*sqrt(c)
  LSA <- c+2*sqrt(c)
  LIA <- c-2*sqrt(c)
  ggplot(df, aes(x = time, y = values, group = 1)) +
    geom_hline(yintercept = c, size = 2, col = "#FFC300") + geom_label(
      label="LC",
      x=max(time),
      y=c,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FFC300"
    ) +
    geom_hline(yintercept = LSC, size = 2, col = "#C70039") + geom_label(
      label="LSC",
      x=max(time),
      y=LSC,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#C70039"
    ) +
    geom_hline(yintercept = ifelse(LIC < 0, 0, LIC), size = 2, col = "#C70039")
    + geom_label(
      label="LIC",
      x=max(time),
      y=LIC,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#C70039"
    ) +
    geom_hline(yintercept = LSA, size = 2, col = "#FF5733") + geom_label(
```

```
      label="LSA",
      x=max(time),
      y=LSA,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FF5733"
    ) +
    geom_hline(yintercept = ifelse(LIA < 0, 0, LIA), size = 2, col = "#FF5733") +
    geom_label(
      label="LIA",
      x=max(time),
      y=LIA,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FF5733"
    ) +
    geom_point(color = "black", size = 3) +
    geom_point(color = "#3DCFD5", size = 2) +
    geom_line(color = "black", size = 1.5) +
    geom_line(color = "#3DCFD5", size = 1) +
    scale_color_manual(values = c("#DAF7A6")) +
    theme_minimal() +
    ggtitle("Gráfico C para número de casos") +
    theme(plot.title=element_text(hjust=0.5))

}

gcc(TS2, 12, 19, TS2$victimas, TS2$mesrec)

## Gráfico P

### Gráfico P ni iguales

gcp <- function(df, auspicious, size, time){
  p <- sum(auspicious)/(size*length(auspicious))
  pi <- auspicious/size
  LIC <- p-(3*sqrt(((p*(1-p))/size)))
  LSC <- p+(3*sqrt(((p*(1-p))/size)))
  LIA <- p-(2*sqrt(((p*(1-p))/size)))
  LSA <- p+(2*sqrt(((p*(1-p))/size)))
  ggplot(df, aes(x = time, y = pi, group = 1)) +
    geom_hline(yintercept = p, size = 2, col = "#FFC300") + geom_label(
      label="LC",
      x=max(time),
      y=p,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
```

```
      fill="#FFC300"
    ) +
    geom_hline(yintercept = LSC, size = 2, col = "#C70039") + geom_label(
      label="LSC",
      x=max(time),
      y=LSC,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#C70039"
    ) +
    geom_hline(yintercept = LIC, size = 2, col = "#C70039") + geom_label(
      label="LIC",
      x=max(time),
      y=LIC,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#C70039") +
    geom_hline(yintercept = LSA, size = 2, col = "#FF5733") + geom_label(
      label="LSA",
      x=max(time),
      y=LSA,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FF5733"
    ) +
    geom_hline(yintercept = LIA, size = 2, col = "#FF5733") + geom_label(
      label="LIA",
      x=max(time),
      y=LIA,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FF5733") +
    geom_point(color = "black", size = 3) +
    geom_point(color = "#3DCFD5", size = 2) +
    geom_line(color = "black", size = 1.5) +
    geom_line(color = "#3DCFD5", size = 1) +
    scale_color_manual(values = c("#DAF7A6")) +
    theme_minimal() +
    ggtitle("Gráfico p para proporciones con tamaños de muestra iguales") + theme(plot

}

gcp(PEQ, PEQ$disconformes, 50, PEQ$muestra)

### Grafico P con ni distintos
```

```
#### Grafico p con n media. Tamaños con variabilidad baja

gcpn <- function(df, auspicious, sizes, time){
  n_ <- mean(sizes)
  p <- sum(auspicious)/sum(sizes)
  pi <- auspicious/sizes
  LSC <- p+3*sqrt((p*(1-p)/n_))
  LIC <- p-3*sqrt((p*(1-p)/n_))
  LIA <- p-2*sqrt((p*(1-p))/n_)
  LSA <- p+2*sqrt((p*(1-p))/n_)
  ggplot(df, aes(x = time, y = pi, group = 1)) +
    geom_hline(yintercept = p, size = 2, col = "#FFC300") + geom_label(
      label="LC",
      x=max(time),
      y=p,
      label.padding = unit(0.55, "lines"),
      label.size = 0.1,
      color = "black",
      fill="#FFC300"
    ) +
    geom_hline(yintercept = LSC, size = 2, col = "#C70039") + geom_label(
      label="LSC",
      x=max(time),
      y=LSC,
      label.padding = unit(0.55, "lines")
      label.size = 0.1,
      color = "black",
      fill="#C70039"
    ) +
    geom_hline(yintercept = LIC, size = 2, col = "#C70039") + geom_label(
      label="LIC",
      x=max(time),
      y=LIC,
      label.padding = unit(0.55, "lines"),
      label.size = 0.1,
      color = "black",
      fill="#C70039") +
    geom_hline(yintercept = LSA, size = 2, col = "#FF5733") + geom_label(
      label="LSA",
      x=max(time),
      y=LSA,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FF5733"
    ) +
    geom_hline(yintercept = LIA, size = 2, col = "#FF5733") + geom_label(
      label="LIA",
      x=max(time),
      y=LIA,
```

```
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FF5733") +
    geom_point(color = "black", size = 3) +
    geom_point(color = "#3DCFD5", size = 2) +
    geom_line(color = "black", size = 1.5) +
    geom_line(color = "#3DCFD5", size = 1) +
    scale_color_manual(values = c("#DAF7A6")) +
    theme_minimal() +
    ggtitle("Gráfico p para proporciones n media") +
    theme(plot.title=element_text(hjust=0.5))

}

gcpn(PMD, PMD$D, PMD$n, PMD$m)

#### Grafico p. Tamaños con variabilidad considerable

gcpz <- function(df, auspicious, sizes, time){
  pi <- auspicious/sizes
  p <- sum(auspicious)/sum(sizes)
  zi <- (pi-p)/sqrt((p*(1-p))/(sizes))
  ggplot(df, aes(x = time, y = zi, group = 1)) +
    geom_hline(yintercept = 0, size = 2, col = "#FFC300") + geom_label(
      label="LC",
      x=max(time),
      y=0,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#FFC300"
    ) +
    geom_hline(yintercept = 3, size = 2, col = "#C70039") + geom_label(
      label="LSC",
      x=max(time),
      y=3,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#C70039"
    ) +
    geom_hline(yintercept = -3, size = 2, col = "#C70039") + geom_label(
      label="LIC",
      x=max(time),
      y=-3,
      label.padding = unit(0.55, "lines"),
      label.size = 0.35,
      color = "black",
      fill="#C70039") +
```

```
    geom_point(color = "black", size = 3) +
    geom_point(color = "#3DCFD5", size = 2) +
    geom_line(color = "black", size = 1.5) +
    geom_line(color = "#3DCFD5", size = 1) +
    scale_color_manual(values = c("#DAF7A6")) +
    theme_minimal() +
    ggtitle("Gráfico p para proporciones n media") +
    theme(plot.title=element_text(hjust=0.5))

}

gcpz(PMD, PMD$D, PMD$n, PMD$m)
```

Código para los números índices

```
# Números índice simples

nicl <- subset(ni$esclarecidos, ni$comunidad == 7 & ni$ambito == 1)
nicl2 <- subset(ni$esclarecidos, ni$comunidad == 7 & ni$ambito == 2)
nicl3 <- subset(ni$esclarecidos, ni$comunidad == 7 & ni$ambito == 3)
nicl4 <- subset(ni$esclarecidos, ni$comunidad == 7 & ni$ambito == 4)
nicl5 <- subset(ni$esclarecidos, ni$comunidad == 7 & ni$ambito == 5)
df <- data.frame(nicl, nicl2, nicl3, nicl4, nicl5, year)

ggplot(df, aes(x=year)) +
  geom_line(aes(y = (nicl/mean(nicl))*100), color = "red", size = 1.5) +
  geom_line(aes(y = (nicl2/mean(nicl2))*100), color="orange", size = 1.5) +
  geom_line(aes(y = (nicl3/mean(nicl3))*100), color="blue", size = 1.5) +
  geom_line(aes(y = (nicl4/mean(nicl4))*100), color="green", size = 1.5) +
  geom_line(aes(y = (nicl5/mean(nicl5))*100), size = 1.5) +
  geom_line(aes(y = 100), color = "darkred")

# Números índice compuestos sin ponderar

nicr <- function(df, time, y){
  nia14 <- subset(y, time == 2014)
  nia15 <- subset(y, time == 2015)
  nia16 <- subset(y, time == 2016)
  nia17 <- subset(y, time == 2017)
  nia18 <- subset(y, time == 2018)
  nia19 <- subset(y, time == 2019)
  nia20 <- subset(y, time == 2020)
  AIa1 <- ((nia20-nia14)/nia14)*100
  AIa2 <- ((nia20-nia19)/nia19)*100
  AIaC1 <- sum(AIa1)/5
  AIaC2 <- sum(AIa2)/5
  print(nia14)
  print(nia15)
  print(nia16)
  print(nia17)
```

```
  print(nia18)
  print(nia19)
  print(nia20)
  print(AIa1)
  print(AIa2)
  print(AIaC1)
  print(AIaC2)
}

# Números índice compuestos ponderados

nia14 <- subset(nip, nip$year == 2014)
nni14 <-  nia14 %>%
  group_by(comunidad) %>%
  summarise(suma=sum(conocidos))
nia20 <- subset(nip, nip$year == 2020)
nni20 <-  nia20 %>%
  group_by(comunidad) %>%
  summarise(suma=sum(conocidos))
nnpond <- nip %>%
  group_by(comunidad) %>%
  summarise(pond=mean(pobpr))
AIe <- ((nni20$suma-nni14$suma)/nni14$suma)*nnpond$pond;AIe
AIep <- (sum(AIe)/sum(nnpond$pond))*100; AIep

#Comparación antes y después pandemia

niax <- nip[nip$year < 2020,]
nnisum <-  niax %>%
  group_by(comunidad, year) %>%
  summarise(suma=sum(conocidos))
nnimean <- nnisum %>%
  group_by(comunidad) %>%
  summarise(media=mean(suma))
nnimean <- nnimean$media[!is.na(nnimean$media)]
nia20 <- subset(nip, nip$year == 2020)
nni20 <-  nia20 %>%
  group_by(comunidad) %>%
  summarise(suma=sum(conocidos))
nnpond <- nip %>%
  group_by(comunidad) %>%
  summarise(pond=mean(denskm2))
AIe <- ((nni20$suma-nnimean)/nnimean)*nnpond$pond;AIe
AIep <- (sum(AIe)/sum(nnpond$pond))*100; AIep
```

Capítulo 11

Anexos III: Salidas de Weka

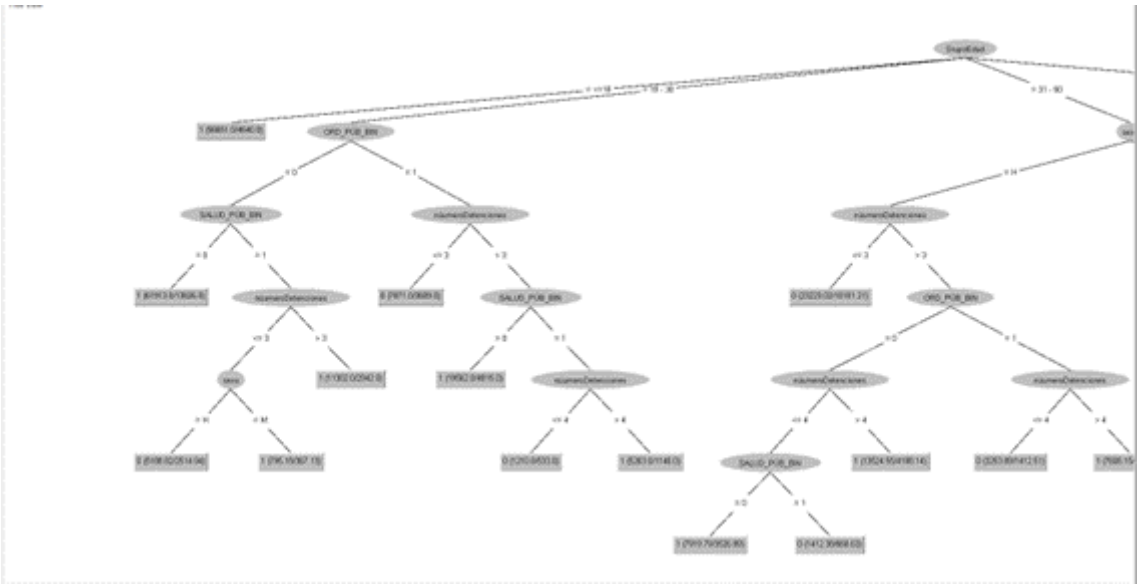


Figura 11.1: Árbol de decisión primera parte

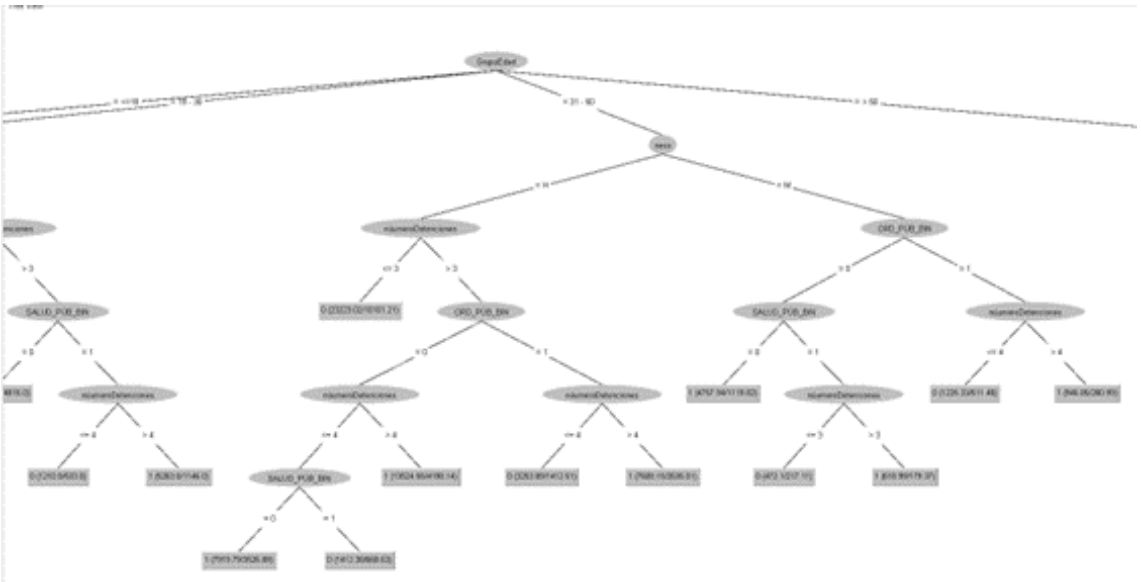


Figura 11.2: Árbol de decisión segunda parte

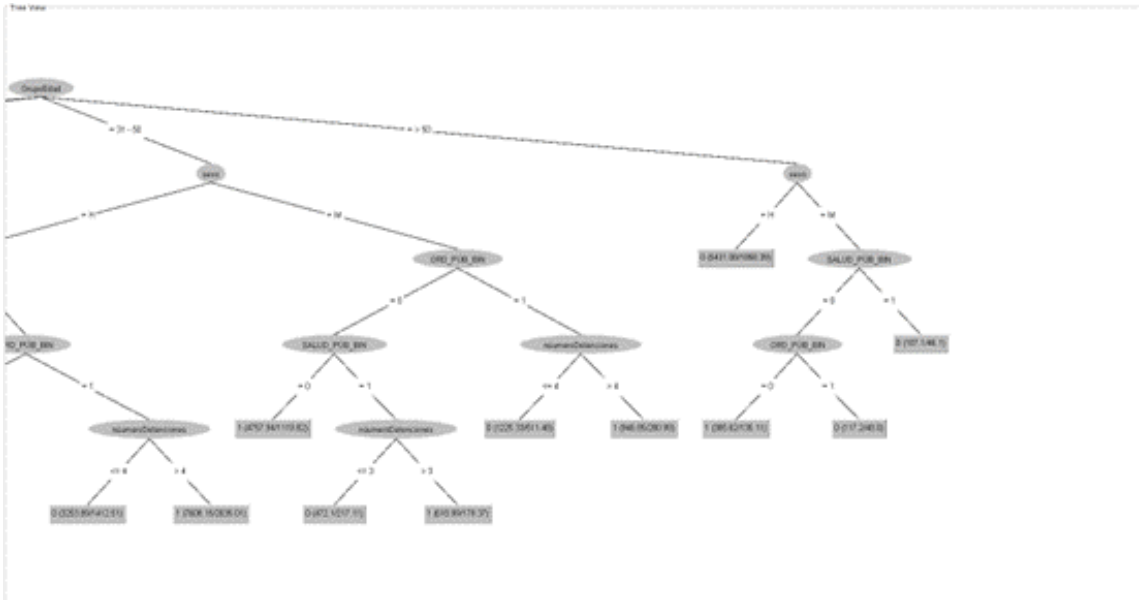


Figura 11.3: Árbol de decisión tercera parte

```
    J48 pruned tree
------------------

GrupoEdad = <=18: 1 (56651.0/4640.0)
GrupoEdad = 19 - 30
|   ORD_PÚB_BIN = 0
|   |   SALUD_PÚB_BIN = 0: 1 (61913.0/13826.0)
|   |   SALUD_PÚB_BIN = 1
|   |   |   núumeroDetenciones <= 3
|   |   |   |   sexo = H: 0 (5188.82/2514.94)
|   |   |   |   sexo = M: 1 (795.18/367.13)
|   |   |   núumeroDetenciones > 3: 1 (11302.0/2942.0)
|   ORD_PÚB_BIN = 1
|   |   núumeroDetenciones <= 3: 0 (7871.0/3609.0)
|   |   núumeroDetenciones > 3
|   |   |   SALUD_PÚB_BIN = 0: 1 (19562.0/4815.0)
|   |   |   SALUD_PÚB_BIN = 1
|   |   |   |   núumeroDetenciones <= 4: 0 (1210.0/533.0)
|   |   |   |   núumeroDetenciones > 4: 1 (5263.0/1146.0)
GrupoEdad = 31 - 50
|   sexo = H
|   |   núumeroDetenciones <= 3: 0 (23223.02/10101.21)
|   |   núumeroDetenciones > 3
|   |   |   ORD_PÚB_BIN = 0
|   |   |   |   núumeroDetenciones <= 4
|   |   |   |   |   SALUD_PÚB_BIN = 0: 1 (7919.79/3526.89)
|   |   |   |   |   SALUD_PÚB_BIN = 1: 0 (1412.38/668.63)
|   |   |   |   núumeroDetenciones > 4: 1 (13524.55/4198.14)
|   |   |   ORD_PÚB_BIN = 1
|   |   |   |   núumeroDetenciones <= 4: 0 (3253.89/1412.51)
|   |   |   |   núumeroDetenciones > 4: 1 (7608.15/2635.01)
```

```
|   sexo = M
|   |   ORD_PÚB_BIN = 0
|   |   |   SALUD_PÚB_BIN = 0: 1 (4757.94/1119.82)
|   |   |   SALUD_PÚB_BIN = 1
|   |   |   |   núumeroDetenciones <= 3: 0 (472.1/217.11)
|   |   |   |   núumeroDetenciones > 3: 1 (618.99/179.37)
|   |   ORD_PÚB_BIN = 1
|   |   |   núumeroDetenciones <= 4: 0 (1225.33/511.48)
|   |   |   núumeroDetenciones > 4: 1 (946.85/280.99)
GrupoEdad = > 50
|   sexo = H: 0 (5431.08/1858.39)
|   sexo = M
|   |   SALUD_PÚB_BIN = 0
|   |   |   ORD_PÚB_BIN = 0: 1 (385.62/135.11)
|   |   |   ORD_PÚB_BIN = 1: 0 (117.2/48.0)
|   |   SALUD_PÚB_BIN = 1: 0 (107.1/46.1)

Number of Leaves  :  24

Size of the tree :  45


Time taken to build model: 1.24 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      179387               74.5086 %
Incorrectly Classified Instances     61373               25.4914 %
Kappa statistic                      0.3129
Mean absolute error                  0.3445
Root mean squared error              0.4151
Relative absolute error             85.1389 %
Root relative squared error         92.2909 %
Total Number of Instances           240760

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,411    0,124    0,565      0,411    0,476      0,320    0,737     0,503     0
0,876    0,589    0,791      0,876    0,832      0,320    0,737     0,868     1
Wd. Avg.    0,745    0,458    0,728      0,745    0,731      0,320    0,737     0,765

=== Confusion Matrix ===

      a      b   <-- classified as
  27854  39949 |      a = 0
  21424 151533 |      b = 1
```

11.1. Algoritmo a priori

```
=== Run information ===

Scheme:       weka.associations.a priori
              -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     datosTablaPlanaConGruposMD-weka.filters.unsupervised.attribute
              .Remove-R5-weka.filters.unsupervised.attribute.Remove-R7,9,11,13-weka
              .filters.unsupervised.attribute.Remove-R11,13-weka.filters
              .unsupervised.attribute.Remove-R5-6-weka
              .filters.unsupervised.attribute.Remove-R4
Instances:    240760
Attributes:   9
              Nacionalidad
              GrupoEdad
              sexo
              LIB_SEX_BIN
              ORD_PÚB_BIN
              PATR_BIN
              PER_BIN
              REL_FAM_BIN
              SALUD_PÚB_BIN
=== Associator model (full training set) ===


a priori
=======

Minimum support: 0.5 (120380 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 22

Size of set of large itemsets L(3): 16

Size of set of large itemsets L(4): 1

Best rules found:

 1. PATR_BIN=1 PER_BIN=0 136955 ==> LIB_SEX_BIN=0 130261
 <conf:(0.95)> lift:(1.02) lev:(0.01) [2368] conv:(1.35)

 2. PATR_BIN=1 REL_FAM_BIN=0 129754 ==> LIB_SEX_BIN=0 123389
 <conf:(0.95)> lift:(1.02) lev:(0.01) [2220] conv:(1.35)
```

```
 3. PATR_BIN=1 172957 ==> LIB_SEX_BIN=0 163749
 <conf:(0.95)> lift:(1.01) lev:(0.01) [2236] conv:(1.24)

 4. PATR_BIN=1 SALUD_PÚB_BIN=0 140686 ==> LIB_SEX_BIN=0 132943
 <conf:(0.94)> lift:(1.01) lev:(0.01) [1566] conv:(1.2)

 5. sexo=H PATR_BIN=1 152397 ==> LIB_SEX_BIN=0 143580
 <conf:(0.94)> lift:(1.01) lev:(0.01) [1267] conv:(1.14)

 6. Nacionalidad=ESPAÑOLA 148796 ==> LIB_SEX_BIN=0 140095
 <conf:(0.94)> lift:(1.01) lev:(0) [1144] conv:(1.13)

 7. REL_FAM_BIN=0 160606 ==> LIB_SEX_BIN=0 150433
 <conf:(0.94)> lift:(1) lev:(0) [454] conv:(1.04)

 8. PER_BIN=0 190880 ==> LIB_SEX_BIN=0 178747
 <conf:(0.94)> lift:(1) lev:(0) [497] conv:(1.04)

 9. Nacionalidad=ESPAÑOLA sexo=H 131543 ==> LIB_SEX_BIN=0 123163
 <conf:(0.94)> lift:(1) lev:(0) [324] conv:(1.04)

10. sexo=H REL_FAM_BIN=0 138520 ==> LIB_SEX_BIN=0 129287
<conf:(0.93)> lift:(1) lev:(-0) [-67] conv:(0.99)

11. ORD_PÚB_BIN=0 PER_BIN=0 139291 ==> LIB_SEX_BIN=0 129965
<conf:(0.93)> lift:(1) lev:(-0) [-109] conv:(0.99)

12. sexo=H PER_BIN=0 167256 ==> LIB_SEX_BIN=0 155996
<conf:(0.93)> lift:(1) lev:(-0) [-192] conv:(0.98)

13. ORD_PÚB_BIN=0 170341 ==> LIB_SEX_BIN=0 158633
<conf:(0.93)> lift:(1) lev:(-0) [-436] conv:(0.96)

14. PER_BIN=0 SALUD_PÚB_BIN=0 153195 ==> LIB_SEX_BIN=0 142663
<conf:(0.93)> lift:(1) lev:(-0) [-395] conv:(0.96)

15. sexo=H 213634 ==> LIB_SEX_BIN=0 198751
<conf:(0.93)> lift:(1) lev:(-0) [-746] conv:(0.95)

    === Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100
              -periodic-pruning 10000 -min-density
              2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance
              -R first-last" -I 500 -num-slots 1 -S 10
Relation:     datosTablaPlanaConGruposMD-weka.filters.unsupervised.attribute.
              Remove-R9,11,13,15,17,19
Instances:    240760
Attributes:   13
              Nacionalidad
              GrupoEdad
```

```
              sexo
              núumeroDetenciones
              edadPrimeraDetención
              edadÚltimaDetención
              mediaDiasEntreDetenciones
              LIBERTAD SEXUAL
              ORDEN PÚBLICO
              PATRIMONIO
              PERSONAS
              RELACIONES FAMILIARES
              SALUD PÚBLICA
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 15
Within cluster sum of squared errors: 116841.3675110625

Initial starting points (random):

Cluster 0: ÁFRICA,'31 - 50',H,4,34,37,371,1.284414,1.546727,3.45849,1.314114,
1.774497,1.674797
Cluster 1: ESPAÑOLA,<=18,H,9,17,36,839.75,1.284414,1.546727,9,1.314114,
1.774497,1.674797
Cluster 2: ESPAÑOLA,<=18,H,24,14,19,81.826087,1.284414,1.546727,23,
1.314114,1.774497,1
Cluster 3: ESPAÑOLA,'31 - 50',H,10,32,45,521.111111,1.284414,1.546727,7,
1.314114,1,1.674797

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                                Full Data          0          1
                                        (240760.0) (38953.0) (75785.0)
=============================================================================
Nacionalidad                              ESPAÑOLA LATINOAMÉRICA   ESPAÑOLA
GrupoEdad                                 19 - 30    19 - 30    19 - 30
sexo                                            H          H          H
núumeroDetenciones                         5.7994     5.0051     5.9694
edadPrimeraDetención                       26.6113    24.4235    23.8023
edadÚltimaDetención                          33.73    27.9123    33.1602
mediaDiasEntreDetenciones                695.7568   382.1625   915.9688
LIBERTAD SEXUAL                             1.2844      1.277      1.285
ORDEN PÚBLICO                               1.5467     1.5263     1.5543
```

PATRIMONIO	3.4585	3.0328	3.3716
PERSONAS	1.3141	1.3033	1.3107
RELACIONES FAMILIARES	1.7745	1.7501	1.7644
SALUD PÚBLICA	1.6748	1.6598	1.6822

Attrib	2	3
	(56647.0)	(69375.0)
ESPAÑOLA	ESPAÑOLA	
<=18	31 - 50	
H	H	
7.2476	4.8773	
16.4273	39.2236	
24.8873	44.8393	
683.3986	641.3675	
1.2796	1.2918	
1.5517	1.5458	
4.4004	3.0233	
1.345	1.2986	
1.7411	1.8265	
1.6627	1.685	

Time taken to build model (full training data) : 1.7 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	38953	(16%)
1	75785	(31%)
2	56647	(24%)
3	69375	(29%)