Universidad Complutense de Madrid

Facultad de Estudios Estadísticos

Máster en minería de datos e inteligencia de negocios

Proyecto de fin de máster
"Anonimización de datos continuos utilizando análisis factorial"

Palmira Aldeguer, Junio 2016

Tutora: Aida Calviño Martínez

Índice de contenidos

1. Introducción

En esta era de la Sociedad de la Información se generan ingentes cantidad de datos y esta tendencia está creciendo, siendo cada vez más el volumen de datos disponibles. Además, se está demandando desde la sociedad que exista una política de transparencia con los datos públicos de manera que se pueda explotar la información disponible.

No obstante, existe determinada información que, por su sensibilidad, (por ejemplo datos personales o confidenciales) no se pueden publicar ya que se podrían estar incumpliendo ciertos derechos fundamentales.

En concreto, en el Reglamento (Ce) No 223/2009 Del Parlamento Europeo y del Consejo se encuentra el siguiente párrafo: "*Los datos sobre unidades estadísticas individuales podrán difundirse en forma de fichero de uso público consistente en registros anónimos elaborados de manera que no se identifique a la unidad estadística, ni de forma directa ni indirecta, teniendo en cuenta todos los medios pertinentes que un tercero pueda utilizar razonablemente.*"

Para cubrir estas dos necesidades (preservar la privacidad y publicar datos) se han desarrollado diferentes técnicas englobadas en lo que se ha denominado anonimización o secreto estadístico. Esta necesidad dual se puede resumir en que se pueda conocer información pero protegiendo la identidad.

El presente proyecto trata de explotar la técnica estadística de análisis factorial para conseguir este fin de anonimizar datos que deban ser publicados y protegidos al mismo tiempo.

2. Definición del problema

En el apartado anterior se ha presentado el problema dual entre proteger y publicar pero, incluso la identificación de los datos a proteger es, en sí mismo, motivo de estudio. Esto se debe a que existen diferentes consideraciones a tener en cuenta.

En primer lugar se define "qué es identificar". Como indica la propia palabra, es reconocer a una persona o empresa (en adelante se abreviará utilizando únicamente el término individuo para ambos casos).

Este reconocimiento de un individuo se puede deber a diferentes grados de conocimiento:

- Datos identificativos o identificadores. Puede existir un dato que directamente se pueda asociar con un individuo. En este punto se clasificaría el DNI o el nombre completo.

- Datos Identificables o casi-identificadores: Es un conjunto de datos que, unidos, permiten reconocer a un individuo. Puede tratarse de una sola variable o de varias. A diferencia de los identificadores, no tienen una relación directa y unívoca con el individuo, pero se puede establecer una regla de asociación basada en conocimiento. Por ejemplo, teniendo en cuenta sólo la información sobre beneficios del año anterior, un conjunto de datos de empresas se podría asociar con el registro público donde se publica esa información y reconocer a las empresas incluidas en el conjunto de datos. En otro caso, podrían tener que utilizarse varias variables, como la unión del ámbito geográfico, el número de empleados y la superficie para conocer a qué empresa corresponde alguna de las observaciones.

Dependiendo del grado de esfuerzo para reconocer un individuo con base en determinadas variables, será más importante asegurar la protección de dichas variables.

Además, existe una clasificación de la propia información según su confidencialidad, lo que implica el grado de seguridad que se necesita aplicar que es:

- Información sensible: Aquella que por su naturaleza necesita ser protegida (salario, religión, salud, etc.).
- Información no sensible: Aquella que no necesita ser protegida de manera especial por tratarse de información no confidencial de los individuos.

En este caso también se puede ponderar el grado de protección necesario, siendo mayor cuanto más sensible se clasifique la información.

Estas dos clasificaciones responden a los riesgos subyacentes que se listan a continuación:

- Riesgo de reconocimiento de un individuo (identificador e identificable).
- Riesgo de divulgación de información o exposición de un individuo (sensible o no-sensible). Este riesgo es complementario al de reconocimiento ya que, dado un reconocimiento de un individuo que utilice un subconjunto de variables, el resto de variables serán las que amplíen el conocimiento respecto a ese individuo.

Al no ser excluyentes, se puede ponderar el grado de seguridad que se necesita aplicar a las variables. Para el caso de identificadores, normalmente se utiliza un mecanismo de eliminación de la información ya que la probabilidad de reconocimiento es del 100%; para proteger el resto de variables, se utilizan diferentes técnicas según los datos sean continuos o categóricos (ordinales o nominales).

Compañía	Superficie	Empleados	Beneficios
A	790	60	250.137
B	1200	75	428.423
C	90	40	200.575
D	540	34	155.286
E	760	45	189.482
F	560	50	250.528

Tabla 1. Ejemplo de datos de empresas

La Tabla 1 muestra un ejemplo de conjunto de datos en el cual la variable *Compañía* sería un identificador; la variable *beneficios* podría ser un identificable, en tanto se publique ese dato en las memorias anuales de las empresas; y la conjunción de *superficie* y *empleados* podría también ser utilizado para reconocer una empresa, por ejemplo, en el caso de poca superficie y muchos empleados.

Existe una casuística específica que no se va a tratar en este trabajo que es el de las tablas de magnitudes y de frecuencias que también tienen los riesgos antes mencionados aunque de una manera más indirecta ya que no se trata de observaciones individuales sino agregadas. Para estos dos casos se trabaja con el concepto de k-anonimidad, mediante el cual se pretende indicar que un valor agregado determinado está en riesgo si los individuos que se incluyen en esa suma son menos de un valor umbral k. Tanto la problemática como los algoritmos aplicados cambian respecto al escenario inicial con el que se va a trabajar, por lo que en el resto de apartados se centrará toda la narrativa en el análisis del escenario de observaciones individuales (también denominados microdatos).

3. Objetivo

El objetivo de este análisis es el estudio de un nuevo mecanismo de protección de información sensible basado en la técnica de análisis factorial. Como preludio al análisis posterior, se puede indicar que se ha elegido esta técnica estadística en base a estudios anteriores que utilizan análisis de componentes principales (que es un caso concreto del factorial) así como por sus cualidades de mantener la relación entre las variables. Gracias a la extracción de factores comunes e independientes, además de mantener la relación entre variables, permite aplicar diferentes modificaciones univariantes. Esta propiedad elimina la complejidad existente en algunas otras técnicas que tienen que incluir cálculos multivariantes para no modificar la relación entre variables.

Tal como se ha visto en el apartado anterior, existen multitud de escenarios de qué información se debe proteger, por lo que el estudio se basará en la creación de dos escenarios básicos:

- Protección de todas las variables.
- Protección de un subconjunto de variables identificadas como sensibles.

Por tanto, el objetivo final es comparar la técnica de análisis factorial como mecanismo de anonimización con otros mecanismos existentes para los escenarios seleccionados y poder evaluar la bondad de esta técnica propuesta. Para evaluar la bondad se utilizan los conceptos de utilidad y seguridad:

- Seguridad: indica el riesgo de identificación
- Utilidad: indica el parecido de los datos resultantes respecto a los originales y, por tanto, cuan útil son los resultados que se obtengan sobre los datos publicados

En apartados posteriores se detallarán las métricas que se utilizarán para medir estas dos características.

4. Estructura del proyecto

El trabajo realizado se ha estructurado en las siguientes fases:

- Estudio del estado del arte en cuanto a la problemática existente y los estudios previos para su solución.
- Definición de criterios de comparación: donde se identifican los parámetros para comparar los métodos y elegir las métricas homogéneas por las que medirlos. En este punto se han desarrollado los aspectos de seguridad y utilidad que son los que rigen el problema inicialmente descrito.
- Descripción del análisis factorial, sus principales propiedades y su aplicabilidad al proceso de anonimización.
- Comparación y elección del algoritmo de modificación: en este punto se eligen varios algoritmos de modificación de los factores y se analizan los resultados de cada uno de ellos para elegir el algoritmo que mejores resultados obtiene según los criterios establecidos.
- Elección de los parámetros del método elegido: tras la selección del algoritmo se vuelve a analizar, para ese algoritmo concreto, todas las alternativas de modificación mediante un análisis más profundo. Este análisis incluye la evaluación de todas las alternativas de modificación de factores así como de los residuos.

- Comparación con otros métodos: en esta fase se eligen otros algoritmos identificados en el estado del arte como los buenas alternativas para compararlos con el que se ha elegido en los pasos anteriores. Se podrán comparar y determinar cuál es el mejor y en qué situaciones.
- Caso de protección selectiva. Hasta este momento se ha asumido que se querían proteger todas las variables continuas de un conjunto de datos, pero, dependiendo de la clasificación de la información que contenga cada variable, puede existir la necesidad de sólo proteger algunos de los datos o incluso querer protegerlos con diferente grado de utilidad y seguridad. En este apartado se analiza el caso específico de protección selectiva con los mismos métodos que para el caso general.
- Concusiones y trabajo futuro: Por último se las conclusiones generales y algunas líneas de actuación en las que se podría profundizar en trabajos posteriores.

5. Estado del arte

En los últimos años la disciplina SDC (Statistical Disclosure Control) ha tenido un gran auge pero sigue siendo una disciplina reciente por lo que, a pesar de existir diferentes métodos para minimizar el riesgo de identificación y exposición, todavía queda campo para investigar diferentes técnicas que mejoren los resultados obtenidos hasta el momento.

Cabe comentar que existe un factor diferencial en cada conjunto de datos que puede hacer que un determinado algoritmo sea superior a otros para un escenario en concreto. En el caso del análisis factorial, éste extrae nuevas variables inherentes a las iniciales que son independientes, pero si el grado de correlación entre las variables es prácticamente nulo las ventajas de aplicar esta técnica podrían verse diluidas ya que serían similares a utilizar otros algoritmos directamente.

En función de la naturaleza de los datos que se quieran proteger con la anonimización, también se distingue entre variables: categóricas, continuas y ordinales; donde cada una de ellas tiene sus propios algoritmos de anonimización.

A causa de la no existencia de un mecanismo general que funcione en todas las circunstancias, se han desarrollado diferentes técnicas estadísticas cada una de las cuales potencia alguna de las variables medidas (seguridad o utilidad). Estas técnicas pasan a describirse a continuación.

Una clasificación de los métodos de anonimización podría ser:

- Sin modificación: Aquellos métodos que no modifican los valores de las observaciones sino que se basan en la agrupación o eliminación de cierta información.
- Con modificación: Aquellos métodos en que no se cambia el número de variables ni observaciones pero se protege la información mediante algoritmos que la alteran de manera que, o bien no se pueda identificar a qué individuo corresponde una observación, o bien si se puede identificar los valores en sí mismos, al estar modificados, se disminuye el riesgo de exposición.
- Datos sintéticos e híbridos: Aquellos métodos que implican la generación de nueva información sintética similar a la información de origen.

Por último, cabe comentar que, para la modificación de datos o la generación de nuevos, se ha encontrado en la documentación de referencia utilizada muchos estudios de métodos de anonimización que unen varias de las técnicas comentadas que mejoran los resultados de los métodos individuales.

Métodos sin modificación

En estos métodos no se modifica la información original sino que se eliminan parcialmente detalles de la información original. En este enfoque se está eliminando información, por lo que la completitud del conjunto utilizado se puede ver penalizada si se elimina información relevante o se queda sesgada la muestra. Otra manera de explicar estos algoritmos sería que no se modifican los datos per sé sino la agrupación, disminuyendo bien el número de variables o el número de observaciones disponibles.

Algunos de los métodos existentes se describen en el cuadro siguiente

Método	Variables	Descripción	Comentarios
Muestra	Categóricas	Se genera una muestra de los datos iniciales.	Se puede obtener información sesgada si no se utiliza una técnica de muestreo adecuada.
Agrupación global	Continuas y Categóricas	Se agrupan variables para formar una nueva o se crea una nueva categoría combinada formada por varios valores.	Se obtiene un conjunto de datos con menor granularidad que el inicial.
Agrupación superior e inferior	Continuas y Categóricas	Caso especial de agrupación sólo con los valores extremos.	

Método	Variables	Descripción	Comentarios
Eliminación local	Categóricas	Se cambian por NA los datos de las variables (o combinación de ellas) que se consideran en peligro por tener muy poca representatividad.	Al estar eliminando datos se pierde completamente cierta información.

Tabla 2. Métodos sin alteración de datos

Métodos con modificación

En este caso se opta por modificar los valores para mantener toda la información disponible aunque modificada para que no sea exactamente igual a la original. De esta manera además de disminuir el riesgo de identificar el individuo correspondiente a una observación dada también se decrementa la exposición asociada.

Método	Variables	Descripción	Características
Inclusión de ruido	Continua	Se incluye un cierto porcentaje de ruido.	Dependiendo de la cantidad de ruido incluida se puede estar protegiendo muy poco o disminuyendo la utilidad.
Redondeo	Continua	Se redondean los valores dentro de un rango.	Se pierden las pequeñas diferencias de las observaciones.
Micro-agregación	Continua y Ordinal	Similar al redondeo pero teniendo en cuenta el entorno. Se calcula una medida de centralidad (como la media o la mediana) para valores similares de manera que el histograma ordenado se queda como una escalera aunque de peldaños de tamaño variable.	Se pierden las diferencias de las observaciones (en mayor o menor medida según cuantos valores se agrupen) aunque el redondeo que se realiza está suavizado con la función de distribución.
Permutación (swap)	Continua y Ordinal	Se cambian de lugar todos los valores de las variables a proteger.	Las relaciones entre las variables se pueden perder ya que se está cambiando aleatoriamente alguna variable.
Permutación acotada	Continua y Ordinal	Similar a la permutación pero se intercambian los valores dentro de un rango.	Se pueden llegar a perder las relaciones pero menos que en el caso de permutación ya que el

Método	Variables	Descripción	Características
			cambio está acotado a valores similares (en el rango).
Re-muestreo	Continua	Utilizar la propiedad de que un muestreo aleatorio tiende a mantener las propiedades del conjunto inicial. Dentro de este caso estaría la técnica bootstrap.	Se pierde la diversidad del conjunto inicial ya que pueden aparecer valores repetidos y no estar algunas observaciones del conjunto inicial.
PRAM (post-randomization)	Categórica	Se basa en una función de probabilidad para reclasificar en un grupo incorrecto.	N/A
MASSC	Categórica	Método combinado que usa cuatro pasos: micro aglomeración, sustitución, sub-muestreo y calibración.	N/A

Tabla 3. Métodos con alteración de datos

Todos los métodos anteriores se basan en la alteración de los valores de los datos iniciales, de manera que no se pueda revertir el cambio y obtener los datos originales de una manera directa. En contraposición con los métodos donde no se alteran los valores, aquí se opta por modificar, no la estructura (tipo de variable, número de observaciones o número de variables), sino que se alteran los valores internamente de los datos, sin alterar la estructura que los soporta.

Estos métodos de anonimización con modificación, se caracterizan en su mayoría por tener ciertos parámetros que determinan el grado de alteración y, por tanto, afectan de manera directa o inversa a la seguridad y, al contrario, a la utilidad.

Hay que tener en cuenta que, al tener variables potencialmente correlacionadas en los datos originales, los datos de las variables no son independientes del resto y que, por tanto, si se opta por modificar de manera independiente cada variable con algún algoritmo univariante se podría estar rompiendo las correlaciones de variables y la utilidad se vería muy afectada. Por este motivo, se opta por métodos multivariante que tienen como contrapartida que los tiempos computacionales son más elevados que para sus homónimos univariante y que, en algunos casos, no están tan maduros o no existen tantas opciones de modificación como en el caso univariante.

A continuación se describen con más profundidad los métodos indicados en la tabla anterior para variables continuas ya que es el ámbito del presente proyecto.

Inclusión de ruido

Este método de alteración de datos trata de incluir un factor aleatorio para alterar los datos numéricos de manera que no se conozca el valor real del conjunto inicial.

Este valor de aleatoriedad está acotado con un valor máximo controlado por el parámetro de ruido máximo. En el caso de que el ruido máximo sea pequeño, se tendrá poca protección pero mucha utilidad y, conforme se vaya incrementando el valor del máximo ruido, se incrementará la seguridad pero decrementará la utilidad del conjunto de datos resultante.

Retomando el comentario general de utilizar métodos multivariante, para el caso de generación de ruido, sólo existe la posibilidad en R de crear una distribución normal por lo que están muy limitadas las alternativas.

Para poder incluir ruido sin que se alteren algunos de los momentos de los datos resultantes y, por tanto, se mantengan la media y la varianza, se puede incorporar un factor de corrección del resultado tras la incorporación del ruido que recupere las propiedades de media y varianza del conjunto inicial.

Redondeo

Mediante el redondeo se opta por una pérdida de información acotada al factor de redondeo que se decida. Este factor de redondeo es el parámetro que afecta a los resultados de aplicar este método; ante valores de redondeo pequeños se tendrá mucha utilidad pero poca seguridad y, conforme se amplíe el factor de redondeo, la seguridad aumentará pero la utilidad disminuirá.

A continuación se muestra el resultado de aplicar un redondeo a miles de euros (1.000) sobre la variable *Beneficios*.

Compañía	Superficie	Empleados	Beneficios	Beneficios
A	790	60	250.137	250.000
B	1200	75	428.423	428.000
C	90	40	200.575	201.000
D	540	34	155.286	155.000
E	760	45	189.482	189.000
F	560	50	250.528	251.000

Tabla 4. Ejemplo de redondeo

La diversidad de valores se pierde al dejar sólo la información de miles de euros en la variable *Beneficios*. El redondeo máximo que se realiza en este caso es de 500 euros en cada una de las observaciones.

Si se opta por un redondeo de 100.000 se tendría una pérdida de hasta 50.000 euros de sensibilidad en las mediciones. Como se ve en la siguiente tabla, con este redondeo se llegan a asimilar los beneficios de la empresa C, D y E.

Compañía	Superficie	Empleados	Beneficios	Beneficios'
A	790	60	250.137	300.000
B	1200	75	428.423	400.000
C	90	40	200.575	200.000
D	540	34	155.286	200.000
E	760	45	189.482	200.000
F	560	50	250.528	300.000

Tabla 5. Ejemplo de redondeo con 100.000

Existe una línea sutil respecto a la diferencia entre el método de agrupación global (en que no se modifican los datos) y redondeo (donde sí se modifican). Se ha indicado que los métodos sin modificación no alteran los valores sino que alteran la estructura de los mismos. En este caso concreto, con la agrupación total se podría conseguir también un escalonado de los datos, pero para conseguirlo, en lugar de redondear los propios valores, se definiría una nueva clase categórica que uniese los intervalos y que esas categorías significasen un intervalo: por contra, en el caso del redondeo se alteran directamente los valores sin cambiar el tipo de dato utilizado.

Micro-agregación

El caso de la micro-agregación es similar al redondeo pero teniendo en cuenta la función de distribución de los datos. Para tener en cuenta esta consideración, se utiliza un parámetro k que indica cuántos valores se van a utilizar para unificar la variable.

En el ejemplo utilizado, si se supone que se quiere modificar sólo la variable *beneficios* sin tener en cuenta el resto de variables con un parámetro k=2 obtendríamos

Compañía	Superficie	Empleados	Beneficios	Beneficios'
A	790	60	250.137	225.356
B	1200	75	428.423	339.476
C	90	40	200.575	225.356
D	540	34	155.286	172.384
E	760	45	189.482	172.384
F	560	50	250.528	339.476

Tabla 6. Ejemplo de microagregación

En este caso la empresa D y E se unificarían con el valor medio de ambas; y de forma similar con las empresas C/A y F/B. De esta manera se quedan 3 valores diferentes en lugar de los 6 que había inicialmente.

Este método genera una pérdida de información que depende de los datos sobre los que se utilice y, como se ha visto, hay conjuntos para los que si sus observaciones son similares quedan muy cercanos a su valor original, pero también otros (F/B) que distan mucho de los datos iniciales.

Para utilizar este método con todas las variables continuas de un conjunto habría que aplicar métodos multivariantes tanto para ordenar las observaciones como para realizar los cálculos de agrupación o unificación.

Permutación

Este método (también denominado swap) simplemente cambia la posición de los datos de las observaciones de manera aleatoria de manera que un valor puede cambiar de posición y, por tanto, cambia la observación a la que pertenecía en el conjunto original. Al tener un factor de aleatoriedad también puede darse el caso de que un valor permanezca en la misma posición y observación, aunque la probabilidad será menor que la de cambiar.

Se ha comentado que los métodos han de tener en cuenta que, si se cambian los valores de cada variable de manera independiente, se pueden romper las relaciones entre ellas. En este método es esto lo que sucede ya que se cambian las variables de manera independiente o, como máximo eligiendo un subconjunto de variables para cambiar en bloque. Si se opta por cambiar en bloque varias variables se mantendrán las relaciones entre ellas pero no con el resto y, de manera homóloga al resto de métodos, cuanto más grande es el subconjunto elegido para cambiar en bloque, mayor será la utilidad pero menor será la seguridad.

Permutación acotada

Este método (conocido en inglés como rankswap) surge como variación del método de permutación simple de manera que los valores se intercambian pero dentro de un intervalo máximo. Este intervalo máximo es el parámetro que permite controlar que se mantengan las relaciones entre las variables ya que se cambiarán valores relativamente cercanos.

Si se supone que se quiere cambiar el valor de la variable beneficios con un intervalo máximo de 2 de diferencia - que significa los vecinos hasta grado 2 - podría obtenerse el siguiente resultado:

Compañía	Superficie	Empleados	Beneficios	Beneficios'	Reordenación
A	790	60	250.137	428.423	4->5
B	1200	75	428.423	250.528	6->4
C	90	40	200.575	155.286	3->2
D	540	34	155.286	189.482	1->3
E	760	45	189.482	200.575	2->1
F	560	50	250.528	250.137	5->6

Tabla 7. Ejemplo de permutación acotada

Los resultados de este método, al incorporar un rango máximo son mejores, en cuanto a utilidad, que en el caso de permutación simple, pero incluyen un incremento computacional importante ya que se tiene que ordenar el conjunto de datos de manera previa para poder cambiar de posición los valores dentro de los N vecinos más cercanos.

Re-muestreo

El remuestreo se basa en generar una muestra seleccionando datos del conjunto original permitiendo la repetición de las observaciones. Un método ampliamente utilizado es bootstrap que se basa en la propiedad de convergencia del límite del valor medio de manera que, al tratarse de una muestra aleatoria sobre el conjunto original, el conjunto final no estará lejos del original.

En este caso, la pérdida de información está relacionada con las observaciones que no aparecen en la muestra. Por otro lado, el grado de protección está relacionado con las muestras que se repiten y con las que no aparecen.

Métodos sintéticos

La alternativa de crear datos completamente sintéticos se basa en utilizar la función de distribución para generar un nuevo conjunto de datos que sea similar al inicial aunque completamente nuevo. Para ello se utilizan métodos paramétricos de aproximación para simular la función de distribución mediante una función continua. Este proceso se realiza mediante un suavizado (smooth) de la función de distribución ya que ésta es una función escalonada.

A efectos de mostrar cómo se comporta este método se ha generado un conjunto con una combinación de observaciones aleatorias normales y uniformes. Al aplicar diferentes métodos de suavizado se obtiene el siguiente gráfico:

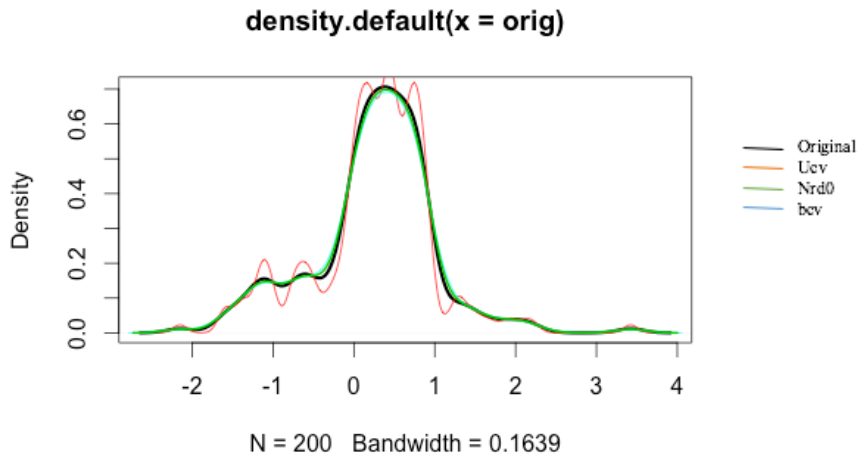


Figura 8. Función de densidad

Además, existen métodos sintéticos de generación de datos más complejos como los basados en imputación múltiple que no se han analizado en el presente proyecto.

Métodos combinados

Los métodos combinados son aquellos que incluyen varias técnicas de modificación de las descritas anteriormente. Se podría considerar en cierta medida que el presente proyecto implica la creación de un nuevo método combinado de anonimización que utiliza el análisis factorial junto con otros métodos de los anteriormente comentados.

En este apartado se van a explicar sólo aquellos que se han elegido para comparar con el análisis factorial.

Microagregación con ruido

En este método se combinan la microagregación (que proporciona seguridad mediante la unificación de los datos) con la inclusión de ruido (para devolver parte de la variabilidad perdida con la microagregación).

La microagregación se realiza con funciones multivariantes teniendo en cuenta todas las variables existentes para no romper las correlaciones existentes. Para calcular la microagregación multivariante, se utilizan funciones de distancia para determinar qué observaciones y variables son las más cercanas. Si el problema fuese univariante o las variables fuesen independientes se podría calcular sólo con la variable a proteger de manera independiente, pero en el caso genérico existen correlaciones con otras variables y se han de intentar respetar las mismas al generar las modificaciones.

A continuación, se introduce ruido cuya magnitud depende de la disminución de la variabilidad derivada del primer paso, a saber, la microagregación.

Análisis de componentes principales con permutación

En este método se utiliza análisis de componentes principales (PCA) para extraer las relaciones entre las variables y poder realizar modificaciones manteniendo inalterada la relación. Por tanto, en primera instancia se extraen los componentes principales que resumen las variables existentes para, a continuación, proceder a modificar los datos de acuerdo a un método simple. En este caso, se modifican los componentes con la permutación de los datos (swap), de acuerdo con los resultados obtenidos en la bibliografía consultada (véase Calviño 2015).

Este análisis es el que se ha tomado como base para profundizar en la posibilidad de extraer los factores, que es un método más general que el análisis de componentes principales que podría mejorar los resultados obtenidos.

Los pasos principales de este método son:

- Se calculan los componentes principales antes de modificarlos por lo que se mantienen las correlaciones entre los componentes tras la modificación.
- Se pueden aplicar técnicas univariantes individualmente sobre cada componente ya que son ortogonales entre ellos. En concreto se aplica la permutación (swap).
- Se recuperan las variables a partir de los componentes modificados.

6. Criterios para comparar los métodos

Tal como se ha indicado anteriormente, para poder comparar los métodos de protección, se utilizan dos conceptos básicos: seguridad y utilidad.

- En la **seguridad** se mide cuánto se han protegido los datos originales, es decir, cuánto se parecen los datos modificados con los originalesy, por tanto, cuánto se puede descubrir sobre alguien con base en los datos modificados. Un método será mejor cuanto más proteja los datos; lo que implica que será mejor cuanto menos se parezcan los datos originales a los modificados.
- En la **utilidad** se mide cuánto se parecen los datos modificados a los originales, es decir, cuán útil son las conclusiones que se obtienen de analizar los datos modificados. En este caso, será mejor el método que consiga que los datos modificados sean lo más parecidos posible a los originales.

Estas dos medidas son contrarias ya que por un lado se persigue que sean muy diferentes los datos originales de los modificados (seguridad) y, por el otro, se quiere que sean lo más parecidos posible (utilidad). En esta intrínseca contradicción se mueven los métodos de protección de datos y, aquellos que son muy buenos protegiendo, son menos adecuados para mantener la utilidad del conjunto resultante, y viceversa.

Todos los valores del resultado de las métricas utilizadas están en el intervalo [0,1], donde el 1 representa el peor valor para esa métrica y el 0 es el mejor valor.

Seguridad

Para medir la seguridad se han elegido las métricas - que se explican más adelante - que comparan la similitud entre los datos originales y los modificados, de manera que, cuanto más se parecen los conjuntos, menos seguridad se está aportando mediante la modificación. Tal como se ha comentado, se han utilizado métricas definidas en el intervalo [0,1] donde 0 es la mayor protección, y 1 es la menor protección (ninguna, es decir que no hay ninguna modificación significativa; si dos conjuntos son idénticos el valor es 1). Estas métricas miden el riesgo de re-identificación, es decir, cuan fácil es poder saber la identidad de un individuo a partir de la información del conjunto modificado. Si se puede establecer una correspondencia unívoca entre ambos conjuntos, la seguridad será mínima aunque los datos no sean exactamente iguales.

- Overall disclosure risk (ODR): mide si un valor modificado está dentro de un intervalo cercano a su correspondiente valor original tomando como referencia la desviación típica.
- Interval disclosure: métrica utilizada en la literatura que también utiliza intervalos para medir si un valor modificado queda más o menos cerca del original pero en este caso se basa en las observaciones más cercanas.
- Outliers: Medición de cuánto se protegen los valores outliers existentes en un conjunto.

Todas estas medidas, intentan determinar cuánto se parecen los conjuntos utilizando una aproximación de los valores, no la forma general, sino focalizando en cada una de las observaciones y buscando su seguridad individual. Esto se puede realizar mirando cada observación, cada intervalo o centrándose en los outliers, pero no dejan de ser métricas para evaluar el riesgo de poder identificar a un individuo a partir de los valores modificados.

Overall Disclosure Risk (ODR)

En esta métrica, para cada valor (observación y variable) se establece un intervalo alrededor del cual se busca el correspondiente modificado. Si se encuentra el valor modificado, la seguridad es menor que si no se encuentra en dicho intervalo. El intervalo se calcula en relación a la desviación típica multiplicada por un parámetro k que determina cuán estricto o sensible se quiere que sea esta métrica.

Este cálculo se realiza para todos los valores del conjunto inicial y se suma para cuántos ha sido localizado el valor modificado en el intervalo correspondiente. A continuación, se normaliza el resultado para obtener un número entre 0 y 1.

Esta métrica está desarrollada en el paquete de R sdcMicro en la función dRisk por lo que se ha utilizado esa implementación con el parámetro k por defecto.

Interval Disclosure

De acuerdo a la bibliografía (véase Domingo-Ferrer y Torra (2004)), en esta medida se escogen varios intervalos alrededor de los valores modificados y se determina si para cada valor su correspondiente valor original está dentro de esos intervalos. Al hacerse de manera iterativa sobre intervalos de tamaños se puede ir midiendo el grado riesgo de identificación. Para el presente trabajo, se han utilizado intervalos crecientes que representan el conjunto de observaciones vecinas desde el 1% al 10% del número total de observaciones.

Al igual que en el caso anterior, se normaliza el valor en el intervalo de 0 a 1.

Outliers

Esta métrica se ha creado en este proyecto para tener en cuenta la distribución de los outliers. Un outlier es una observación atípica que está alejada de los valores medios. Para el presente proyecto, se ha considerado como outlier, toda observación que se aleja de los cuartiles 1 y 3 una distancia superior a 1.5 veces el rango intercuartil (RIC). Esto se corresponde con una observación menor a Q1-1.5RIC o superior a Q3+1.5RIC.

Estos puntos atípicos se han de proteger en especial ya que su propia existencia podría permitir reconocer una observación y, por tanto, que esa observación estuviese en riesgo. Por ejemplo, en una lista de empresas donde haya una empresa con ingresos muy superiores al resto, aunque el valor se modifique puede ser insuficiente si se es capaz de relacionar el nuevo valor con el original; aunque el valor no fuese exacto, podría ser suficiente para identificar ese valor atípico y, por tanto a la empresa. En este aspecto hay publicaciones específicas que tratan la protección de los outliers (véase Mateo et al (2005)).

Con esta idea de que los outliers son especialmente sensibles y hay que medir su protección de manera específica, se ha creado esta medida.

Para calcular cuánto se protegen los outliers existentes, se han calculado los outliers para cada una de las variables que se quieren proteger y, sobre este conjunto inicial, se valida

si siguen existiendo los mismos outliers en el conjunto protegido. También se tiene en cuenta si están en la misma posición o si, aunque sigan estando en el listado de outliers, han cambiado de posición. En el peor de los casos, esta medida indica que todos los outliers se mantienen en la misma posición y, en el mejor de los casos, no se repiten outliers entre el conjunto original y el modificado.

El algoritmo de cálculo se podría resumir en:

- Se obtienen todos los outliers de las variables en el conjunto inicial
- Se obtienen todos los outliers de las variables en el conjunto modificado
- Para cada outlier del conjunto inicial
    - Si existe ese outlier también en los outliers modificados se suma 0.5
    - Si, tras ordenarlos, además está en la misma posición se suma 0.5

Finalmente se normaliza el resultado primero dentro de cada variable (dividiendo entre el número de outliers en esa variable) y, posteriormente, la suma de los resultados parciales (dividiendo entre el número de variables con outliers).

Con este algoritmo se obtiene un valor acotado entre 0 y 1, donde 0 es el mejor valor (no coincide ningún outlier) y 1 el peor (se mantienen todos los outliers y en la misma posición).

Utilidad

En cuanto a la utilidad, ésta se mide por el grado de parecido entre los conjuntos. Esta similitud no es que los datos sean absolutamente idénticos, sino que mantengan el mayor número de propiedades del conjunto inicial. Para ello, se utilizan métricas que tienen en cuenta esas propiedades o características entre las que se encuentran: media, varianza, momentos, etc.

Como se han elegido tres medidas de la seguridad, se van a utilizar también tres medidas de la utilidad:

- PIL1: Esta métrica mide las propiedades relativas al momento 1 y 2, es decir compara la media y la varianza de los dos conjuntos.
- PIL2: En este caso se tienen en cuenta los momentos 3 y 4, es decir, la simetría y la forma de las colas de las distribuciones.
- Propensity Scores: Esta métrica sirve para medir hasta qué punto se pueden diferenciar los dos conjuntos utilizando sólo los propios datos y sus relaciones.

PIL1

El nombre PIL es la abreviatura de "Probabilistic Information Loss" y, en la bibliografía consultada, agrupa 6 indicadores de características estadísticas a preservar, de manera que, cuantas más propiedades se mantengan, mayor es la utilidad.

En este caso, se ha separado en dos métricas para que fuese homogénea la medición de seguridad y utilidad. La separación se ha realizado según si el indicador estaba cerca del centro de la función de distribución o de la forma en general (incluidos en el PIL1) o si estaba más lejos del centro (incluidos en el PIL2). Además, de esta manera se separan los indicadores sobre las colas para tener una medida pareja a la métrica de seguridad de los outliers.

Por tanto, en el PIL1 se han utilizado indicadores para las propiedades:

- Media
- Varianza
- Covarianza

Tal como se ha indicado, estas características están focalizadas en el centro de la función de distribución y en propiedades generales.

La diferencia entre las medidas de los conjuntos se rige por la fórmula:

$$Z = \frac{\hat{\Theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\Theta})}}$$

Según la fórmula anterior, se ha de utilizar la varianza de la función original de la que se ha extraído la muestra para el conjunto original. Para ello, se estima la varianza de la población mediante los momentos de la muestra. Dependiendo de qué PIL se esté calculando, se calcula la varianza de acuerdo al momento correspondiente. En su formulación general se expresa mediante la fórmula:

$$\mathrm{var}\, m_r = \frac{1}{n}(\mu_{2r} - \mu_r^2 + r^2 \mu_2 \mu_{r-1}^2 - 2r\, \mu_{r-1} \mu_{r+1})$$

Por último, en todas las métricas se ha obtenido un valor en el intervalo [0,1] para poder comparar los resultados aunque sean métodos de anonimización diferentes, o incluso si se aplican sobre conjuntos diferentes. Para ello, en este caso, se puede calcular, a partir del valor anterior, cuál es la probabilidad (respecto a una función normal) de ese valor obtenido, que quedará en el intervalo 0 a 1.

$$pil(\hat{\Theta}) = 2 \cdot P(0 \leq Z \leq \frac{|\hat{\theta} - \theta|}{\sqrt{Var(\hat{\Theta})}})$$

PIL2

En esta métrica se han unificado aquellas propiedades relacionadas con los extremos de la función de distribución, así como con la evolución. De esta manera se calculan valores relacionados con:

- Cuartiles: se calcula con base en los cuartiles. Se han cogido todos los cuartiles existentes en intervalos de 0.05% del tamaño del conjunto.
- Simetría: utilizando el cálculo del momento 3.
- Kurtosis: utilizando el cálculo del momento 4.

Se han de tener las mismas consideraciones para el cálculo que en el caso anterior de cómo calcular las varianzas y las probabilidades de los resultados.

Propensity Scores

Con esta métrica, se mide el parecido entre el conjunto de datos inicial y el modificado en términos de relación entre las variables. De esta manera, se evalúa si se puede distinguir el conjunto de datos modificado respecto del inicial.

Para utilizar este método se unen los dos conjuntos de datos (inicial y modificado) y se crea una nueva variable binaria que representa si una observación pertenece al conjunto inicial o al modificado. A continuación, se efectúa una clasificación de las observaciones en 2 conjuntos.

Para la implementación de este método se ha optado por utilizar regresión logística por simplicidad y porque así se sugiere por los autores (véase Woo et al (2009)), pero se podría haber utilizado alguna otra técnica de clasificación no supervisada para agrupar en 2 conjuntos (por ejemplo árboles o redes neuronales). En la regresión logística, además de utilizar las variables directas, se han calculado las interacciones simple y los cuadrados de las variables; así se consigue representar las relaciones entre las variables hasta grado 2. En este aspecto también podría haberse continuado incluyendo interacciones, pero los tiempos de computación habrían aumentado y, al aumentar las variables sintéticas, se podría llegar a forzar una separación completa entre las observaciones no natural (como en el caso de que el número de variables llegase a superar el número de observaciones).

Como la base del análisis factorial es extraer las relaciones subyacentes entre las variables, la anonimización mediante los factores ofrece buenos resultados en cuanto a esta medida ya que conserva muy bien la relación entre las variables, tal y como se verá más adelante.

7. Metodología utilizada

Tras haber descrito cuál es el problema y las diferentes alternativas, se centrará el estudio en el escenario que se va a analizar en el presente trabajo, que es la protección de variables continuas de microdatos utilizando el análisis factorial (junto con alguna técnica de modificación de las indicadas anteriormente).

Propiedades del Análisis factorial

Se ha elegido el análisis factorial siguiendo la investigación del artículo de Calviño (2015) donde se explora la anonimización con componentes principales. El análisis de componentes principales es un caso particular de análisis factorial, por lo que se espera poder generalizar el método y mejorar en algunos aspectos al original, por tratarse de un método más general y, por tanto, con más opciones de parametrización.

El análisis factorial se basa en el cálculo de las variables subyacentes que generan las variables conocidas. Esto quiere decir, que en un conjunto de datos se conocen determinadas variables, pero que, en muchas ocasiones, éstas representan diferentes puntos de vista de un mismo origen, por lo que se podrían calcular las variables originales que han provocado esos puntos de vista. Estas variables originales se denominan factores. El análisis factorial, por tanto, pretende encontrar los factores subyacentes que explican las variables que se observan en el conjunto de datos. Para esta generación de los factores se parte de la premisa de que aquellas variables altamente correlacionadas derivan de un factor subyacente, por lo que para la generación se utiliza la matriz de covarianzas de manera que se agrupan en factores las correlaciones que existen entre las variables.

En el modelo factorial se asume que existe una relación lineal entre las variables que se observan y los factores (que no se pueden ver directamente en el conjunto de datos) de la forma siguiente:

$$\mathbf{X}_{(p\times 1)} - \boldsymbol{\mu}_{(p\times 1)} = \underset{(\mathbf{p}\times\mathbf{m})}{\mathbf{L}}\ \underset{(\mathbf{m}\times\mathbf{1})}{\mathbf{F}} + \underset{(\mathbf{p}\times\mathbf{1})}{\boldsymbol{\varepsilon}}$$

Toda variable observada (X) podría formularse como función lineal de los factores de donde se deriva (factores comunes: F) mas un residuo (factores específicos:ε). Siendo L la matriz de cargas que indica la representatividad de un factor en una variable, y siendo μ el vector de medias.

Con estas premisas, se pueden calcular los factores de manera que sean incorrelados entre sí (ortogonales) y generen las variables de manera lineal añadiendo los residuos.

Teniendo en cuenta este punto de partida, los factores no están relacionados, son incorrelados, y sólo las variables conocidas son las que tienen unas relaciones a causa de cómo se han generado respecto de las primeras.

Esta propiedad es muy interesante, ya que al calcular factores que son independientes entre sí, se pueden alterar libremente sin afectar las relaciones entre las variables del conjunto. Lo que implica que, si el método de modificación de los factores empleado mantiene las medias y las varianzas de los factores originales, el vector de medias y la matriz de varianzas-covarianzas de los datos modificados coincidirá con los de los datos originales.

Las principales propiedades del método se enumeran a continuación:

- Los factores son incorrelados entre sí, es decir, son ortogonales.
- Por la propiedad anterior, se pueden realizar modificaciones univariantes de cada factor de manera individual sin tener en cuenta el resto de factores ya que no se alterarán las relaciones entre las variables.
- Un factor puede estar involucrado en el cálculo de varias variables.
- Cuando se cambia un factor se están cambiando todas las variables que se generan a partir de él.
- Por lo tanto, con un cambio univariante, se pueden estar modificando muchas de las variables que se tengan que proteger.
- Se pueden rotar los factores (ya que son ortogonales) sin alterar las variables derivadas; sólo se altera la proporción de cada factor dentro de una variable.
- Mediante rotaciones se puede ponderar (maximizar, minimizar o lo que se considere) el efecto de las modificaciones en las variables a proteger.

En cuanto a las métricas, se puede derivar, en general, que:

- Propensity Scores: se mantiene muy cercana a valores óptimos de utilidad ya que mantiene la relación de las variables y, por tanto, es muy difícil discernir la separación entre el conjunto inicial y el modificado con base en los valores y sus relaciones.
- PIL1, PIL2: las propiedades estadísticas de las funciones de distribución, no están aseguradas ya que se están modificando los factores inherentes y, por tanto, esto se plasma en las variables analizadas. El poder preservar estas características depende más del método de modificación que de las propiedades del análisis factorial. Por ejemplo una modificación de los factores con ruido implicará que

se modificará en función de la cantidad de ruido incluido; eso sí, la relación entre variables no se verá afectada.

- Seguridad: ninguna de las 3 medidas está asegurada que vaya a tener valores especialmente significativos (ni buenos ni malos), dependerá de qué factores se alteren y cómo.

En los siguientes apartados se va a analizar cómo se pueden alterar los factores y el efecto que tienen sobre las métricas utilizadas para la comparación.

8. Alternativas de modificación

Para anonimizar utilizando análisis factorial en primer lugar se ha de elegir el número de factores. Se han realizado pruebas con todos los factores posibles (desde 1 hasta el máximo) pero, por facilidad en la explicación y visualización en los gráficos, se ha optado por incluir sólo los resultados asociados a 7 factores (sobre un total de 13 variables continuas y 834 observaciones) del conjunto de datos Tarragona (incluido en el paquete sdcMicro de R). Para este caso la variabilidad explicada es del 87% y, por tanto, los residuos contienen sólo el 13% restante. Además, incluir más factores no aporta grandes diferencias en cuanto al resultado de las modificaciones como se verá en el siguiente apartado donde se ahonda en el método de modificación elegido.

Tras calcular los factores se ha de proceder a modificar los factores según alguna técnica de modificación de las indicadas anteriormente. En este apartado se han elegido varias técnicas de modificación de factores y se van a comparar para elegir la que mejores resultados aporta. Para la elección de la técnica de modificación se han utilizado los criterios descritos en el apartado anterior siguientes: PIL1, PIL2, Interval Diclosure, ODR, Propensity Scores y Outliers.

Tal como se ha indicado anteriormente, para evaluar las diferentes técnicas de modificación de los valores de los factores, se ha utilizado el conjunto de datos Tarragona (incluido en la librería sdcMicro de R). Con este conjunto de datos y para cada método de modificación, se han simulado las alternativas de selección de factores a modificar (que son las combinaciones de factores existentes) con un bucle de 100 iteraciones con diferentes semillas para poder medir la variabilidad del método empleado en los resultados. En muchos de los métodos de modificación se utilizan valores pseudo-aleatorios, por lo que este bucle se ha realizado para evitar que una ejecución concreta no fuese representativa de ese método. En los siguientes apartados, se han incluido las gráficas obtenidas que ponen de manifiesto la evaluación de “bondad” del algoritmo de

modificación utilizado, teniendo en cuenta tanto la proporción de variabilidad modificada como el número de factores modificados.

Se podría alegar que elegir 7 factores es un sesgo que se ha incluido en el estudio, pero sólo se ha incluido ese filtro por sencillez en la representación de los resultados ya que, tal como se verá para el mecanismo elegido (permutación/swap), la descomposición en más o menos factores es coherente (si se eliminan las descomposiciones de pocos factores que tienen una particularidad). Como curiosidad, en los casos con pocos factores, para que los resultados sean competitivos respecto al resto, es mejor alterar los residuos que los propios factores, como es lógico.

En cada uno de los apartados siguientes se explica la técnica utilizada y se muestran los gráficos resumen de la bondad de cada método. Para cada uno de ellos se representa tanto la mediana de las 100 semillas (en el gráfico de la izquierda) como la amplitud del intervalo de confianza (en el gráfico de la derecha). El intervalo de confianza obtenido es de tipo *percentil bootstrap*, por lo que sus límites inferior y superior vienen dados por los perrcentiles 2.5 y 97.5, respectivamente.

En el eje X se representa el número de factores modificados y, el color, representa la variabilidad modificada, que viene dada por la variabilidad de los factores modificados.

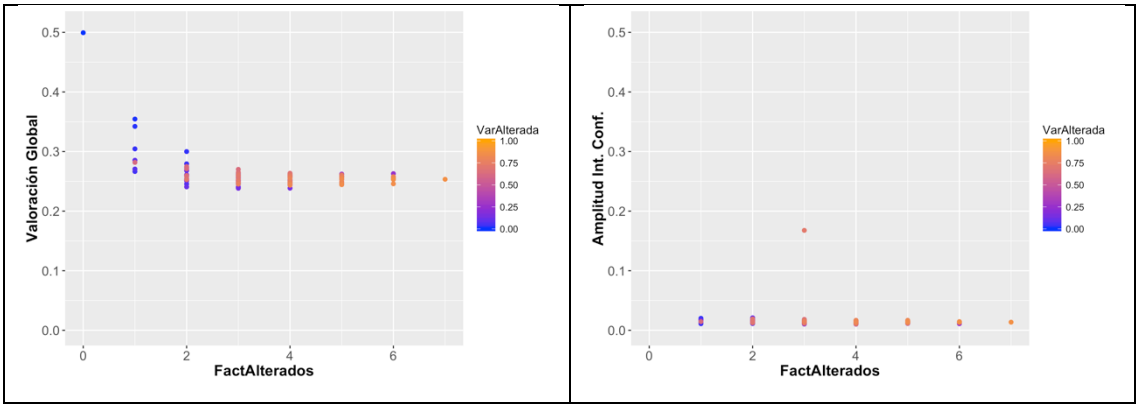


Figura 9. Ejemplo gráfico seguridad/utilidad

En algunos casos también se incluyen los gráficos complementarios donde el eje X representa la variabilidad y el color muestra el número de factores alterados.

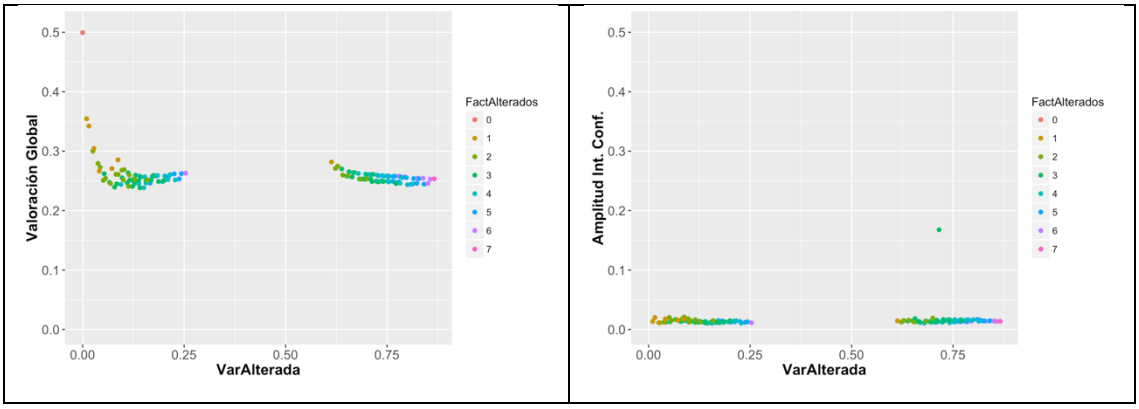


Figura 10. Ejemplo gráfico seguridad/utilidad (II)

En ambos casos se incluye la no modificación como estado inicial que aparece con un valor de 0.5 como resultado global. Como referencia, el no alterar los datos en absoluto (0 varianza y 0 factores alterados) da un resultado con mediana 0.5 y con una amplitud del intervalo de confianza de 0 y, si bien no proporciona seguridad, tiene un 100% de utilidad. En el extremo opuesto se tendría la generación de un conjunto de datos aleatorio del mismo tamaño que aportaría una seguridad del 100% pero una utilidad nula. Por tanto, el 0.5 se puede considerar el peor de los casos (perder totalmente la seguridad o la utilidad) y cualquier modificación que supere este umbral no se considera al ser peor que los dos casos extremos expuestos.

En general la escala para los gráficos globales es de [0-0.5] ya que suelen compensarse los valores de utilidad y seguridad, aunque hay algunos casos donde se ha tenido que alterar la escala ya que los valores eran superiores a 0.5.

También se analiza cada métrica por separado de manera que se podrá comparar los mejores métodos de modificación para obtener mayor seguridad o utilidad.

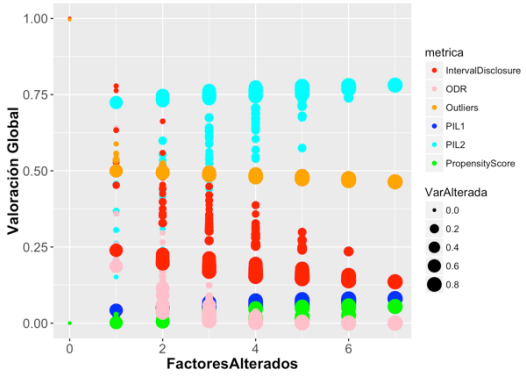


Figure 11. Ejemplo métricas individuales

Para ello, se ha generado un gráfico en el que los colores fríos representan las métricas de utilidad y los colores cálidos las métricas de seguridad.

En el eje de las abcisas se han representado los factores alterados y el diámetro del círculo indica el porcentaje de variabilidad alterado. En algunos casos en que la amplitud del intervalo de confianza era significativa (por ejemplo, por tener gran variedad de valores) se ha incluido también a la derecha de las métricas individuales.

A partir de ese punto inicial se pueden comparar las diferentes combinaciones de alteración de factores y métodos de alteración.

Inclusión de ruido

En este caso, tras aplicar el análisis factorial, se ha incluido ruido univariante siguiendo una distribución normal (aunque se podría haber elegido otra distribución) en cada uno de los factores elegidos.

La cantidad de ruido incorporada se define en función de la varianza de 0.1%, 1%, 10% y un porcentaje dinámico. En este último caso, para proteger mejor los outliers, se ha calculado la distancia entre los 2 outliers más extremos y se ha añadido un ruido máximo,

equivalente a la mitad de la distancia que los separa, para que exista la posibilidad de cambiar los outliers de orden.

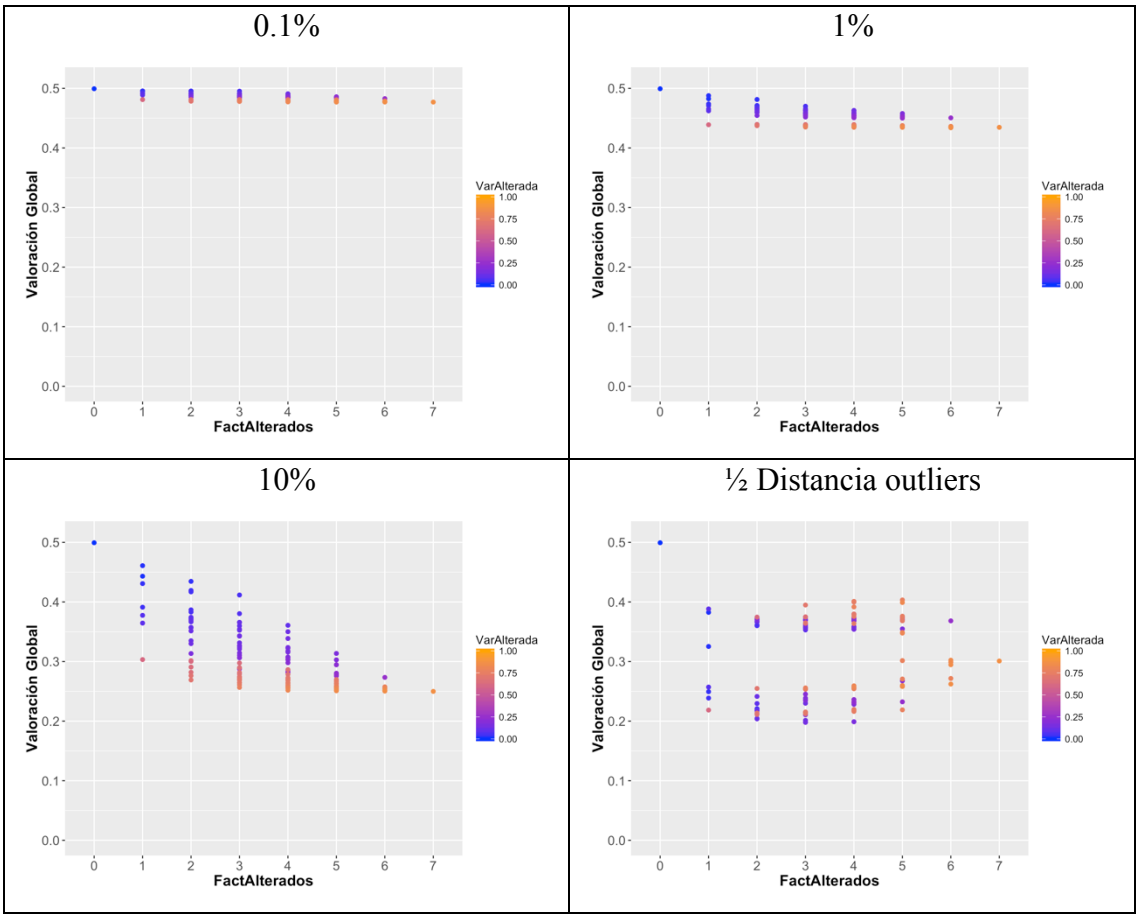
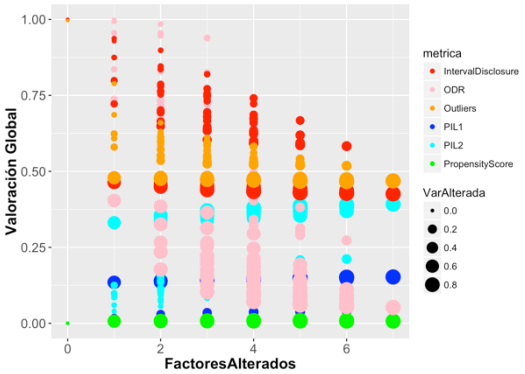


Figura 12. Factorial. Ruido

Los mejores resultados se obtienen para la incorporación de ruido máximo de un 10% de la desviación típica. En los tres primeros 3 casos (0.1, 1 y 10) el comportamiento de la sucesiva alteración de factores es coherente, mejorando conforme se modifican más factores. En cambio, para el caso de incorporar una cantidad de ruido sensiblemente superior, los resultados dejan de ser coherentes y no hay un patrón claro para elegir cuál es la mejor alternativa. El mejor valor global del ruido al 10% es: 0.25 y se obtiene con la modificación de todos los factores.



Para continuar con el análisis se ha elegido la incorporación de ruido del 10% de la desviación típica. En este caso, la inclusión del ruido, tiende a mejorar sensiblemente todas las métricas individuales de seguridad. En cuanto a la utilidad, empeoran con la

inclusión de más modificaciones, pero es menos agresivo que para las métricas de seguridad.

Redondeo

Para probar este método se ha partido de una parametrización dinámica en función de los valores existentes en cada factor. Para ello se ha utilizado la resta del valor máximo con el mínimo y, se ha generado un redondeo correspondiente a dividir esa resta en intervalos del 1% del conjunto. Para el conjunto de Tarragona, cada escalón incluiría (si estuviese uniformemente distribuidos los números) 8 elementos de las 834 observaciones.

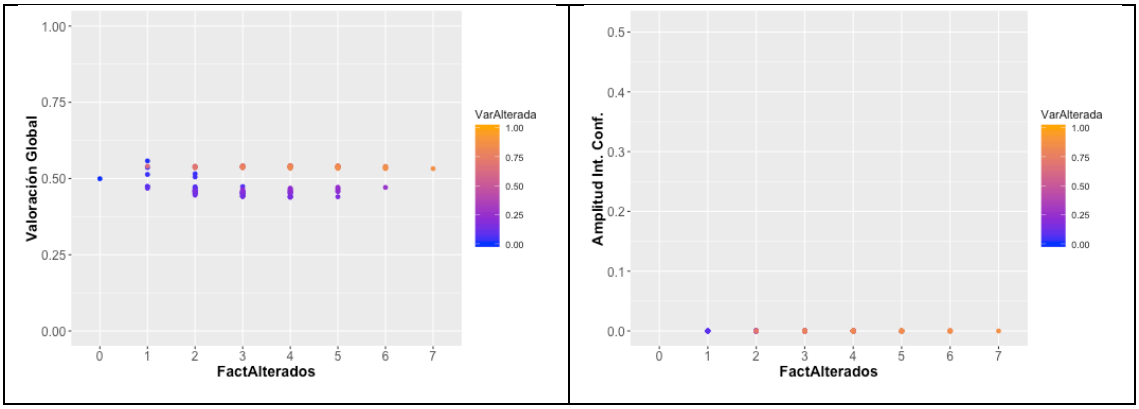


Figura 14. Factorial Redondeo

En este caso se ha tenido que cambiar la escala del eje Y para poder incluir valores superiores a 0.5 ya que se obtienen medidas más cercanas a 1 (peores). El mejor valor de este método con la configuración utilizada es: 0.438635669 con la modificación de los factores: 3, 4, 5 y 7. En cuanto al intervalo de confianza se puede observar (en la figura de la izquierda) que incluye sólo el valor de la mediana; esto ocurre ya que no hay aleatoriedad en la modificación al utilizar la misma aproximación en todos los casos.

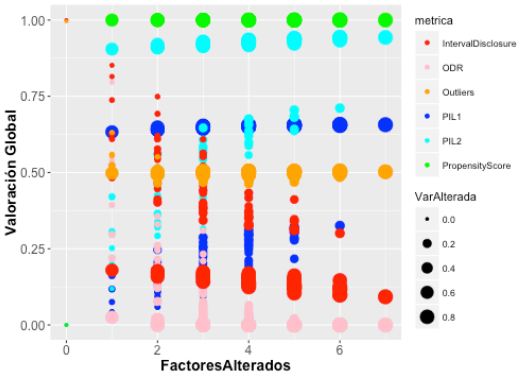


Figura 15. Redondeo. Métricas individuales

Lo que más destaca de esta gráfica es que la métrica Propensity Score es prácticamente 1 en todos los casos, lo que significa que es muy fácil distinguir entre el conjunto original y el modificado. En este caso, el realizar el redondeo sobre los factores en lugar de sobre las variables iniciales no aporta ninguna ventaja, tal como sí ocurre en otros métodos donde esta misma métrica se mantiene en unos niveles muy cercanos a 0. Se puede apreciar que las medidas de seguridad son mucho mejores que las de utilidad ya que,

mediante el redondeo, se está perdiendo mucha de la información contenida en el conjunto inicial.

Micro-agregación

Para esta prueba se ha probado con la agrupación de k=3 ya que en la bibliografía consultada los mejores resultados se conseguían para k pequeños (véase Oganian and Karr (2006)). En el apartado donde se analiza microagregación con ruido se hace un análisis más exhaustivo del parámetro k donde también se concluye que la mejor alternativa para los datos con los que se está trabajando es utilizar k pequeños.

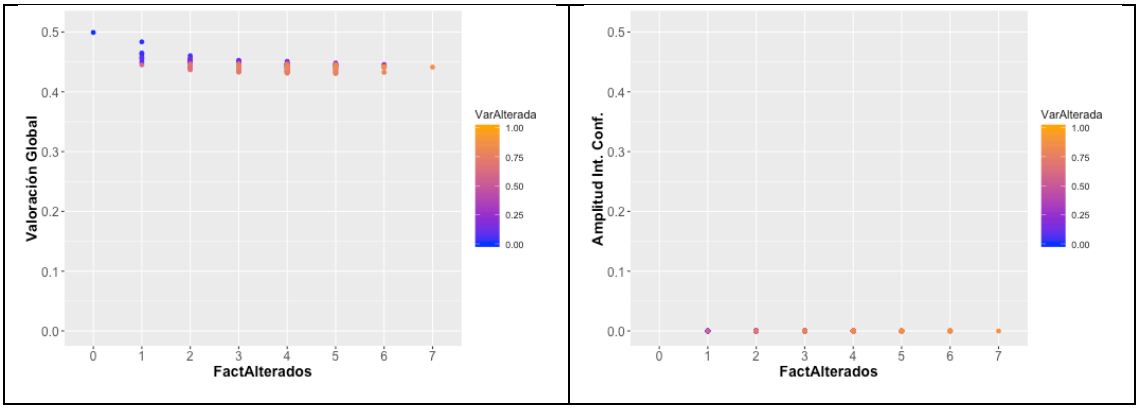


Figura 16.Factorial. Microagregación

Los resultados de este método son bastante mediocres comparados con otros métodos. En ninguno de los casos se baja de 0.4. El mejor valor global de este método es: 0.431 con la modificación de los factores 1, 3, 4 ,6 y 7. Respecto al intervalo de confianza se puede observar que es muy estrecho, aunque al tener valores de la métrica global mayores a 0.4 no aporta un criterio adicional a la propia mediana.

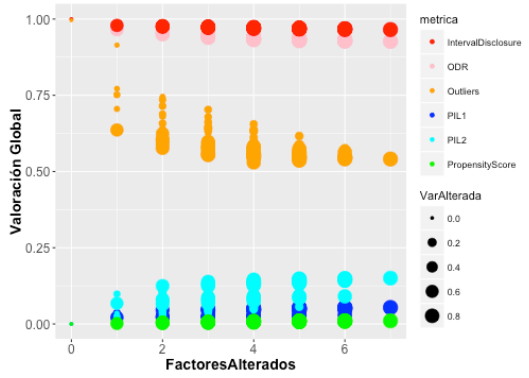


Figura 17.Microagregación. Métricas individuales

Al ser la k tan pequeña, se están agregando conjuntos de valores muy similares por lo que se está premiando la utilidad sobre la seguridad. Los colores fríos quedan cercanos a 0, mientras que los cálidos (seguridad) cercanos a 1.

Además, se puede observar que hay poco cambio tanto con el incremento de factores modificados como con la variabilidad modificada en la mayoría de las métricas.

Permutación

En este método, para cada uno de los factores que se ha decidido modificar, se cambia el orden de los valores. Cabe comentar que cada uno de los factores elegidos se cambia de manera independiente, por lo que no es un cambio de orden de las observaciones; es un cambio que altera los resultados cuando se revierte el análisis factorial de los factores a las variables iniciales.

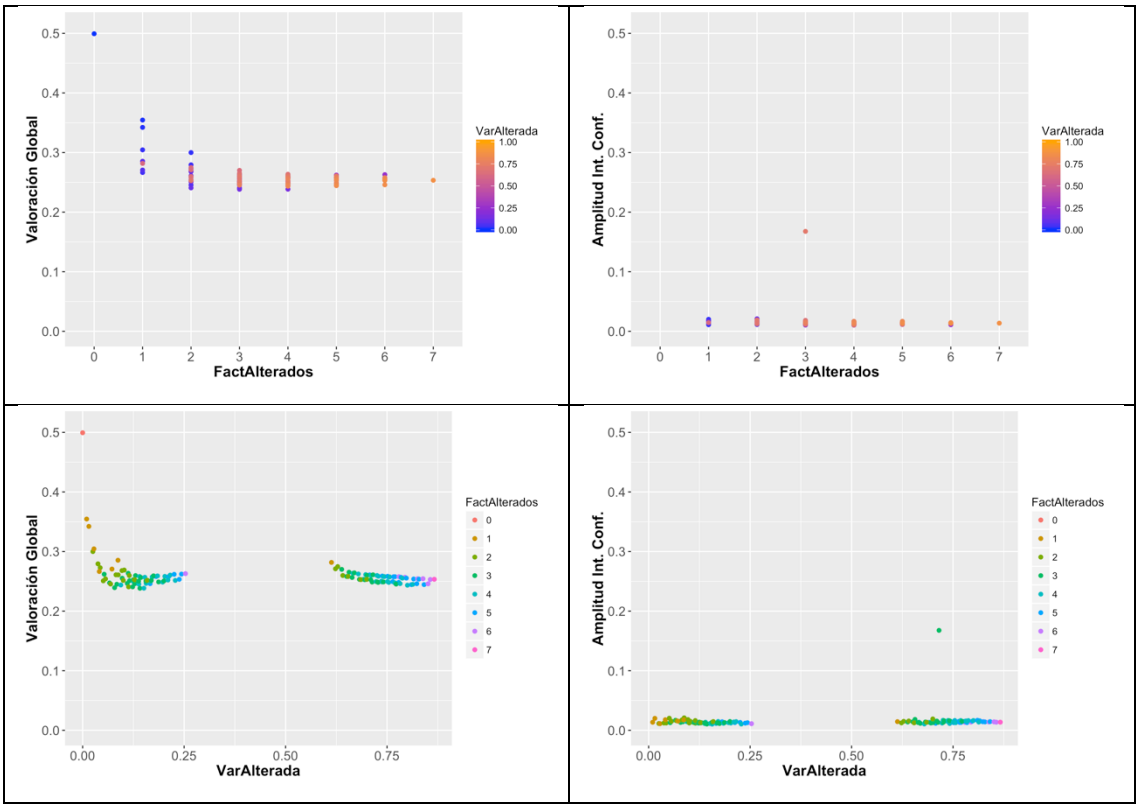
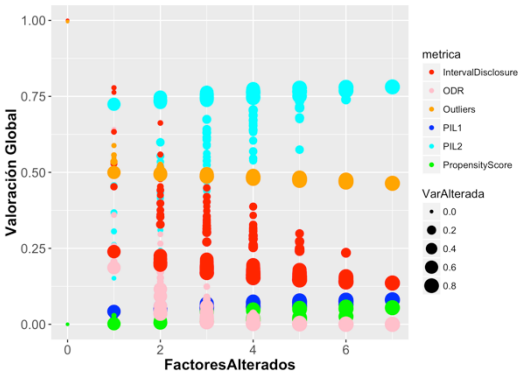


Figura 18. Factorial. Permutación

El mejor valor global de este método es: 0.238, y los factores modificados son los siguientes: 2, 3 y 4. En los gráficos de la derecha se puede observar como hay una configuración (qué factores se modifican) que tiene un intervalo de confianza amplio, cercano a 0.2. Esa configuración, a pesar de tener una mediana similar al resto, se debería eliminar por poder generar resultados muy variables.

La medida de bondad global de este método es bastante constante independientemente de cuántos factores se modifiquen o la variabilidad que representen del conjunto inicial.



Tal como corroboran los gráficos, cuantos más factores se modifiquen, más mejora la seguridad y menos utilidad se consigue.

Lo mismo cabe decir de la variabilidad representada por los factores que se modifiquen.

Figura 19. Permutación. Métricas individuales

Permutación acotada

Este método es una variación de la permutación que tiene en cuenta un parámetro adicional para no cambiar los factores de manera extrema. Para ello, se incluye este nuevo parámetro que indica el máximo rango por el que se puede cambiar un valor. Con esto se calcula un intervalo en el que se puede cambiar un valor dado, lo que genera cambios menos radicales que los que se pueden alcanzar con el método *Permutación*. Para las simulaciones se ha utilizado un parámetro configurado con el valor 20. Este número determina el número máximo de posiciones por lo que el rango máximo tiene una distancia de 40 sobre las 834 observaciones que existentes en el conjunto de datos. Este mecanismo tiene el inconveniente de que hay que ordenar todos los factores que se quieran modificar, por lo que llega a ser mucho más lento que otros algoritmos que requieren menos preprocesamiento.

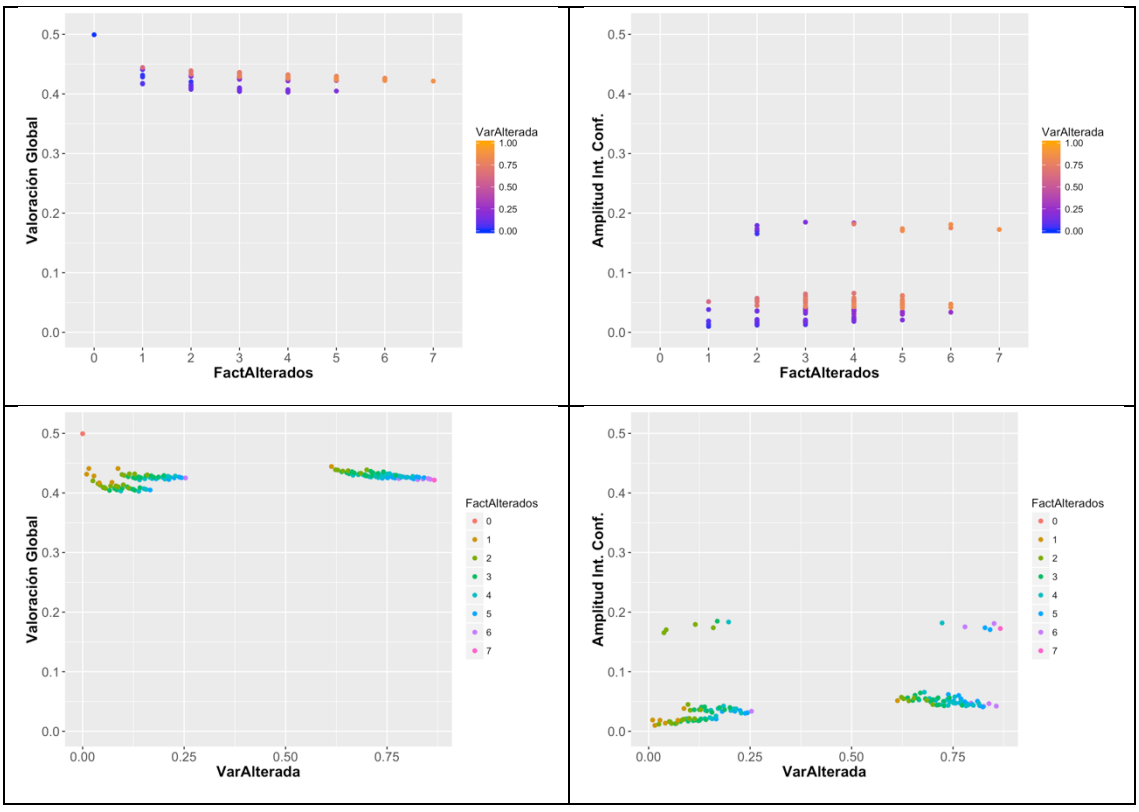


Figura 20. Factorial. Permutación acotada

En este caso se tiene que los valores de la métrica global son superiores en todos los casos a 0.4 y que la amplitud del intervalo de confianza oscila más que en el caso de permutación.

El mejor valor global de este método es: 0.403. Y los factores modificados son los siguientes: 2, 4, 6 y 7.

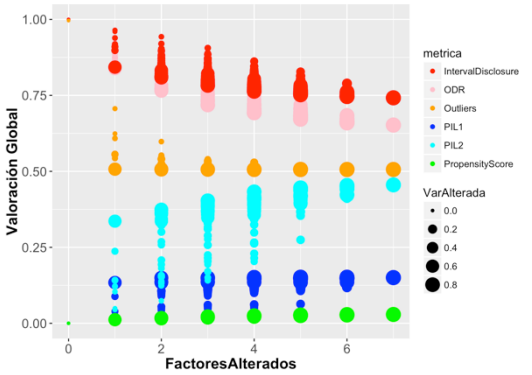


Figura 21. Permutación acotada. Métricas individuales

Este tipo de transformación hace que los cambios sean menos radicales que con la *permutación*, lo que afecta tanto a la seguridad como a la utilidad. Comparativamente con la *permutación*, la utilidad es muy superior (PIL2 que es la peor medida en la *permutación*) y, en general, se mantiene por debajo de 0.5. Por el contrario, todas las medidas de seguridad son peores que en el caso de la *permutación*.

Re-muestreo-bootstrap

En este método se remuestrean los valores de los factores elegidos utilizando selección con repetición. Con este método se conservan las propiedades de la distribución cuando el tamaño de la muestra es lo suficientemente grande, como es el presente caso.

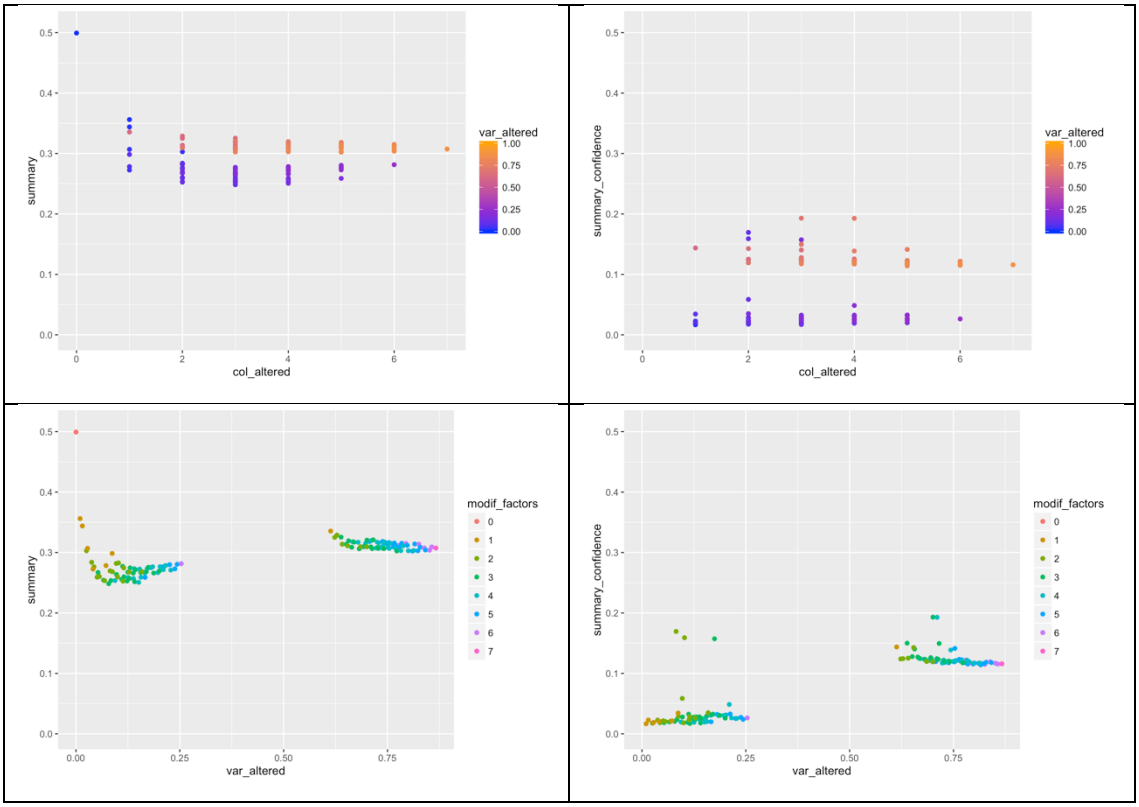


Figura 22. Factorial. Remuestreo

El mejor valor global de este método es: 0.248. Y los factores modificados son los siguientes: 4, 5 y 7.

En este método los resultados parecen buenos en general ya que están desde 0.25 a 0.35, pero el problema que se tiene es que el intervalo de confianza es estrecho para las modificaciones de menos factores/variabilidad. No obstante, para combinaciones de cambios de pocos factores que generan una amplitud elevada. A pesar de tener algunos valores razonables, el no acertar exactamente con la mejor combinación de factores a alterar, puede generar problemas de impredecibilidad de los resultados, que podrían llegar a ser buenos o malos en función de factores, a priori, impredecibles.

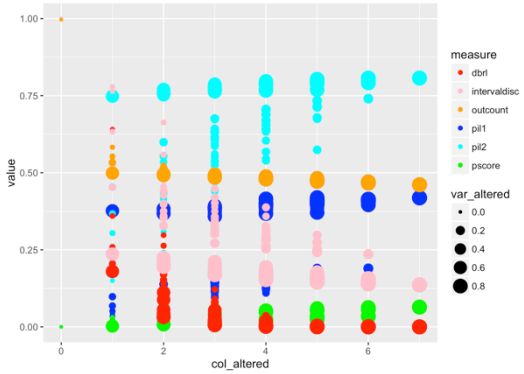
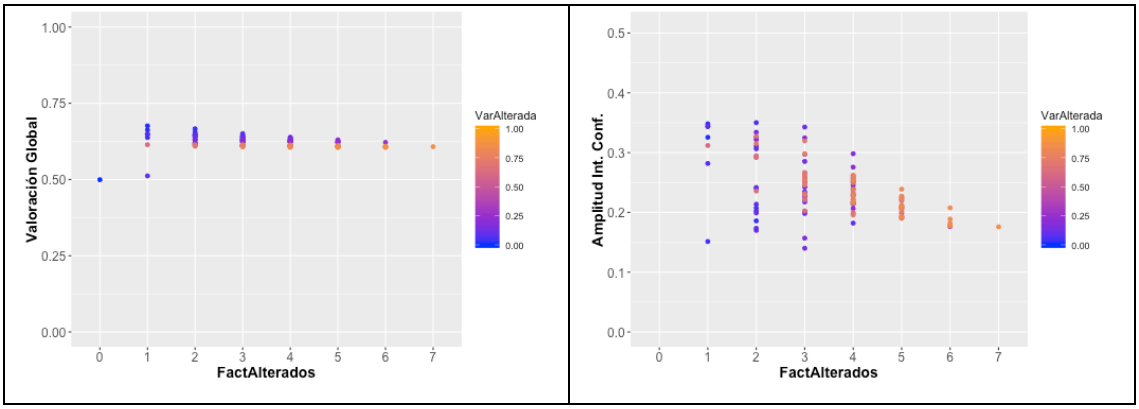


Figura 23. Bootstrap. Métricas individuales

En este método se nota la contraposición de PIL2 e IntervalDisclosure; pero, lo más curioso respecto a los métodos vistos hasta el momento, es que la métrica PIL1 (en azul oscuro) tiene un valor cercano a 0.5 en algunos casos, y también se ve afectada negativamente con el incremento de modificaciones, cosa que no sucede con las modificaciones de *permutación* y la *permutación acotada*.

Sintéticos

En este método se ha utilizado la función de suavizado bw.nrd0 de R que es la que se ha visto en apartados anteriores que mejor se aproximaba a la función de distribución.



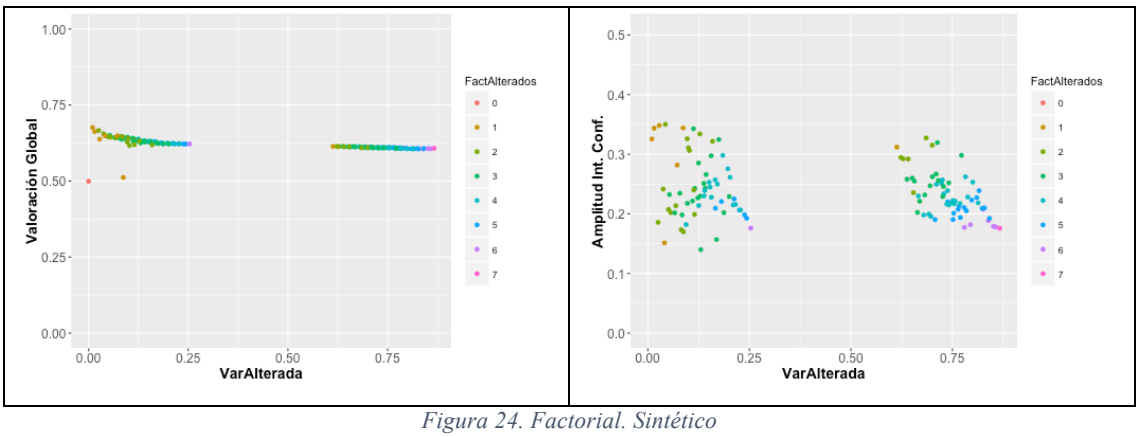


Figura 24. Factorial. Sintético

Se ha tenido que cambiar la escala a [0,1] ya que los valores globales eran peores que en el resto de métodos. Esto, es a causa de que no se compensan las métricas de seguridad con las de utilidad como se ve en el gráfico siguiente.

Por otro lado, la amplitud del intervalo de confianza es muy elevada (con valores cercanos y superiores a 0.2) comparada con otros métodos que tienen esta medida mucho más acotada.

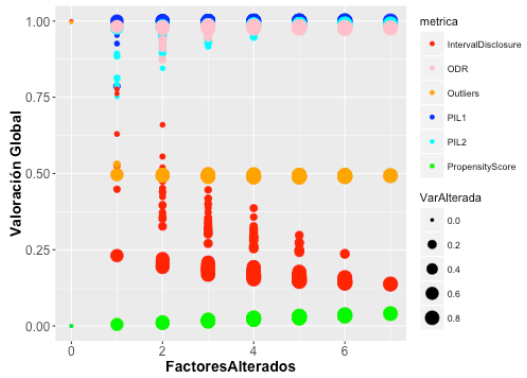


Figura 25. Sintético. Métricas individuales

La bondad general de este método en la presente ejecución es bastante reducida. En las métricas de utilidad tanto PIL1 como PIL2 son pésimas y, en cuanto a seguridad, ODR también es muy mala. Sólo es interesante la métrica de PropensityScores que como se ha indicado anteriormente es buena porque la alteración se realiza sobre los factores de manera que se garantiza, a priori, un buen resultado en la medida PropersityScores.

Para afinar los resultados de este método habría que analizar en profundidad alternativas para generar un conjunto sintético que, además de mantener la relación entre variables, mantenga el resto de propiedades del conjunto inicial.

Residuos

En todos los cambios anteriores, se han modificado cada uno de los factores de manera independiente, es decir, se han cambiado consecutivamente los factores mediante métodos univariantes ya que los factores son ortogonales. Esto no sucede con los residuos que quedan al calcular los factores - que sí están correlados entre sí - de manera que no se pueden utilizar métodos univariantes sino que se obliga a utilizar métodos multivariante con la complejidad que esto aporta.

Las gráficas anteriores, sólo representan las alternativas de modificaciones para los 7 factores (manteniendo los residuos originales) con todas las combinaciones de modificación de factores. No obstante, también se pueden cambiar los residuos de una manera coherente respecto al método utilizado para los factores; de esta manera se tendría para cada uno de los métodos descritos anteriormente:

Alteración de factores	Alteración de residuos
Redondeo	Redondeo multivariante (no se ha implementado)
Microagregación	Microagregación multivariante
Ruido	Inclusión de ruido multivariante normal
Permutación	Permutación de los residuos en bloque (por línea)
Permutación acotada	Permutación acotada de los residuos en bloque (por línea) utilizando la distancia de mahalanobis para ordenar los residuos
Bootstrap	Muestreo utilizando bootstrap
Sintético	Muestreo utilizando bootstrap ya que la única librería multivariante con funciones de densidad no está soportada en versiones recientes de R

Tabla 26. Resumen residuos

El caso particular de la modificación de residuos mediante la permutación se ha analizado en detalle en el siguiente apartado. No obstante, para el resto de modificaciones las conclusiones han sido similares a las obtenidas para la *permutación*.

Conclusiones

La siguiente tabla muestra el resumen de los mejores resultados de cada uno de los métodos analizados (sin alteración de los residuos).

Método	Mejor resultado	Amplitud intervalo confianza	Máscara	Comentarios
Redondeo	0.439	0.000	0011101	No alterando los factores de mayor peso
Microagregación	0.431	0.000	1011011	Alterando los factores de mayor peso
Permutación	**0.238**	**0.012**	**0011100**	**No alterando los factores de mayor peso**
Permutación acotada	0.403	0.021	0011011	No alterando los factores de mayor peso
Ruido	0.250	0.015	1111111	Alterando todos los factores
Muestreo	0.248	0.021	0001101	No alterando los factores de mayor peso
Sintético	0.499	0.000	0000000	

Tabla 27. Comparación factorial

La columna que indica *máscara* se corresponde con los factores que se han modificado. Cada posición corresponde a un factor, donde el 1 representa que se ha modificado ese factor. En la mayoría de los métodos es mejor no cambiar los primeros factores (los que contienen mayor variabilidad).

Los resultados del método de creación de un nuevo conjunto con datos sintéticos no parecen representativos, por lo que habría que analizarlo en detalle para estudiar los parámetros de configuración y mejorar los resultados obtenidos.

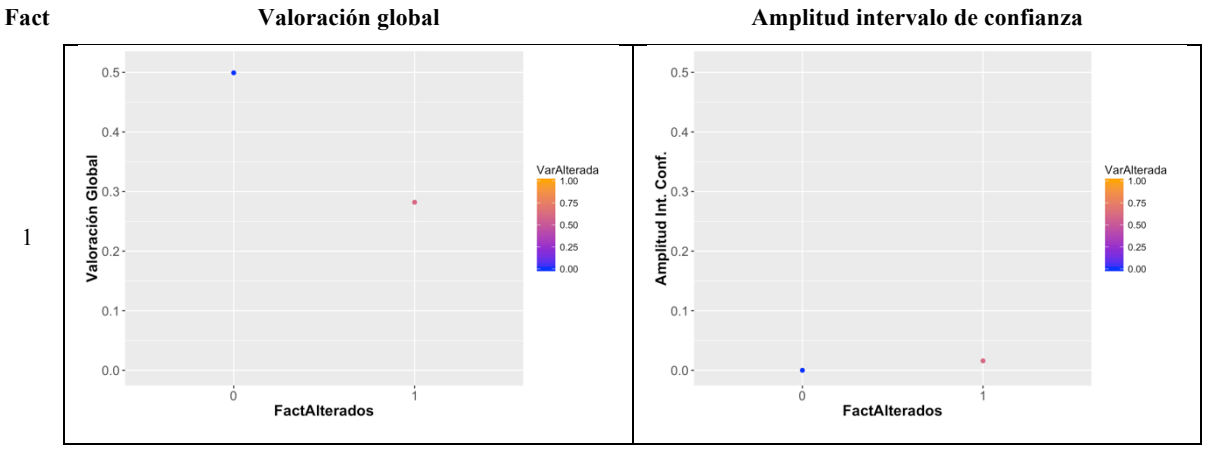
De los resultados obtenidos, se escoge el método de permutación de los factores como mejor alternativa de alteración de los datos; este método se desarrollará en profundidad en el siguiente punto.

9. Parámetros de configuración para permutación

Se ha identificado que el mejor método para la modificación de los factores es la permutación de valores sin acotar el rango máximo. En este apartado se analizará en detalle las opciones a parametrizar en el método de permutación para conseguir los mejores resultados.

Selección de número de factores

En primer lugar se va a analizar en cuantos factores se han de descomponer las variables iniciales. A continuación se muestran los resultados para la descomposición en 1,4, 7, 10 y 13 factores.



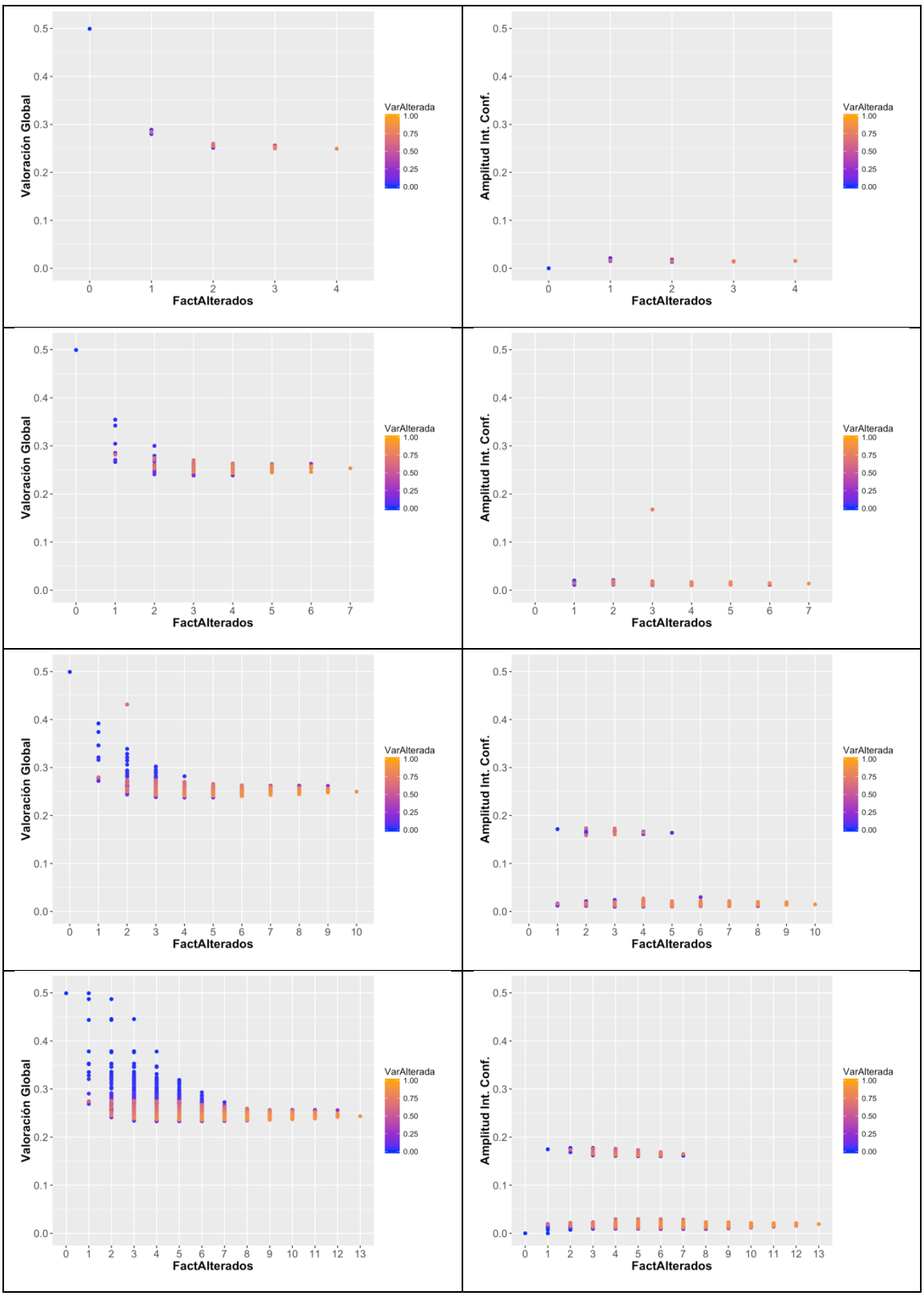


Figura 28. Número de factores

Se puede apreciar que, con la inclusión de más factores, aparecen más posibilidades de configuración, ya que se pueden o no modificar más factores, pero se mantiene la forma del gráfico. Por tanto, se ha considerado interesante utilizar el mayor número de factores y probar todas las posibilidades de modificación.

Para el apartado anterior en el que se han probado todos los métodos, se ha elegido el valor de 7 factores ya que como se puede apreciar, la forma es muy similar y no parece haber grandes diferencias en cuanto a los valores mínimos. Además, los tiempos de computación de las pruebas de todas las alternativas de modificación para 13 factores son muy superiores a las de 7 factores por lo que para la prueba inicial se ha acotado a 7 factores.

Alternativa de modificación

Al generar una simulación para las 8192 posibles combinaciones de modificación de los factores, se puede ver que los mejores valores se dan para un número de factores alterados cercano a la mitad y con una variabilidad relativamente baja.

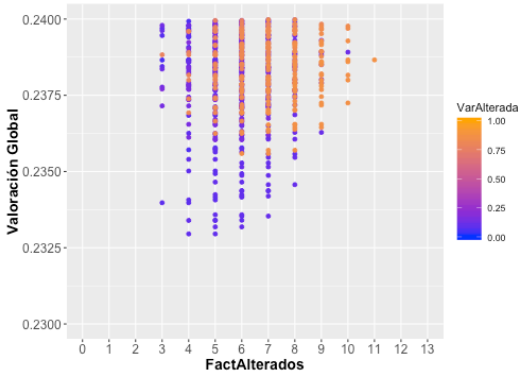


Figure 29. Mejor combinación para modificación

Al elegir los mejores valores (hasta 0.26) y cambiar la escala centrándola en el intervalo [0.23 , 0.24], se puede observar que cuanta menos variabilidad se modifica, los resultados son mejores, independientemente del número de factores que se modifiquen. No se ha incluido el gráfico de la amplitud del intervalo de confianza porque en todos estos casos está por debajo de 0.02.

La siguiente tabla muestra las mejores combinaciones de modificación de factores.

Máscara	Nº factores modificados	Variabilidad de factores modif	Resultado	Int. Confianza
0001110001000	4	0.086074068	0.232953371	0.012699737
0001110001001	5	0.086074068	0.232953371	0.012699737
0001111001000	5	0.097359818	0.23318373	0.015775668
0001111001001	6	0.097359818	0.23318373	0.015775668
0001111000000	4	0.093023606	0.233395646	0.014917419
0001111000001	5	0.093023606	0.233395646	0.014917419
0001110001010	5	0.086421228	0.233402654	0.012393772

Tabla 30. Alternativas modificación

En los mejores valores, se repite el patrón de no modificar los factores de mayor peso, por lo que la variabilidad representada por los factores modificados es muy pequeña.

Alteración de los residuos

Para el caso de que se realice el análisis factorial con pocos factores, los residuos todavía contendrán gran cantidad de la variabilidad del conjunto inicial y de su representatividad, motivo por el cual puede llegar a ser más efectiva la modificación de los residuos que la alteración de los propios factores.

Se representan a continuación los gráficos para las alternativas de modificación con 1, 3 y 7 factores. Los triángulos representan una alteración de los factores y también de los residuos; mientras que los círculos sólo alteración de los factores.

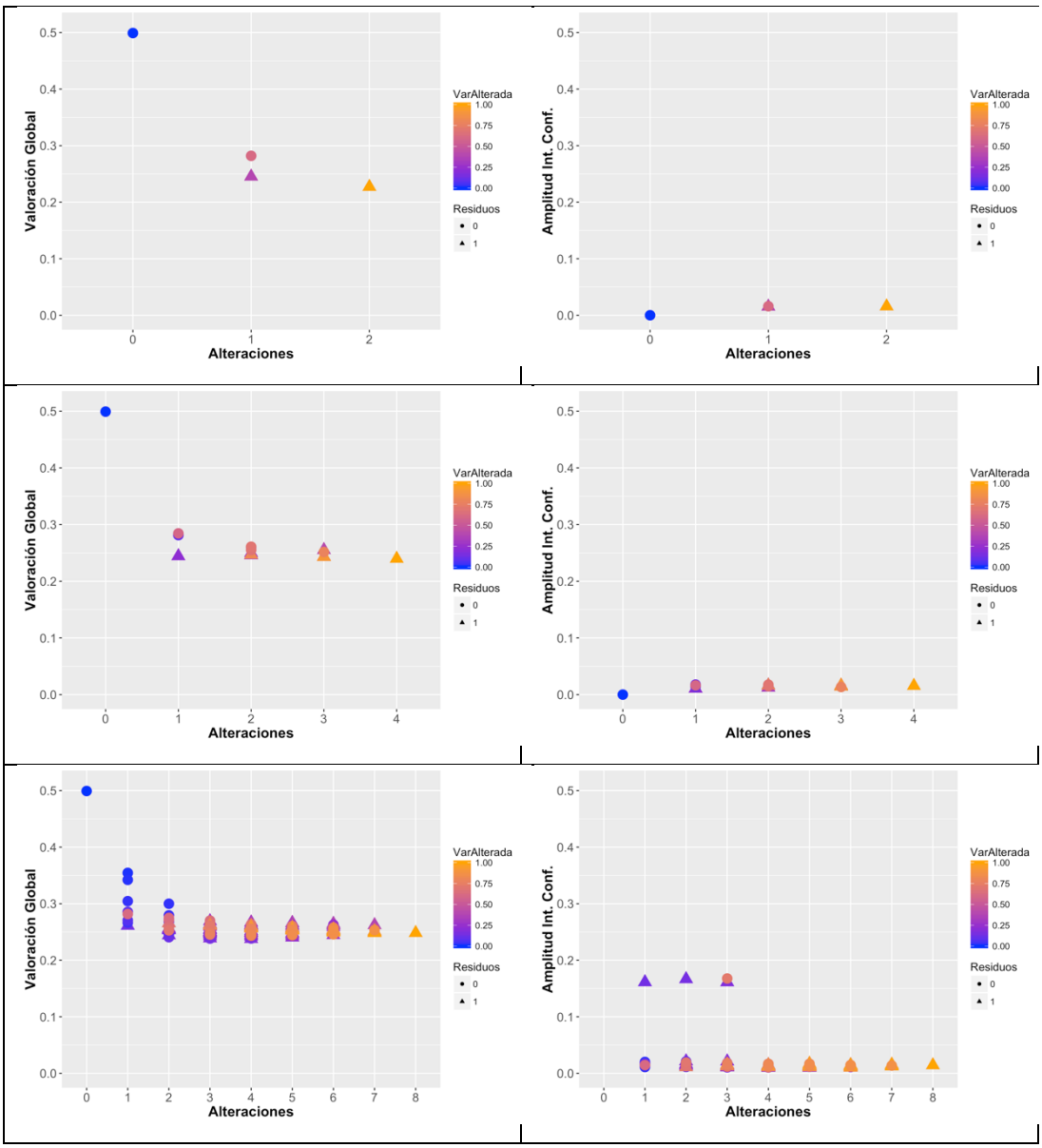


Figura 31. Residuos

En los análisis con pocos factores, sí que la alteración de los residuos supera en resultados a las alternativas que no los modifican, pero este comportamiento deja de ser significativo a partir de cierto número de factores en que ya se está representando la mayoría de la variabilidad del conjunto inicial.

Para el caso más extremo de 1 factor, se obtienen resultados similares de modificar los residuos o modificar el factor. Y modificar el factor y los residuos de manera conjunta es mejor alternativa que las anteriores.

Con el incremento del número de factores se reduce la efectividad de la modificación de los residuos; pero además se ha de considerar que la modificación de un factor es un cambio univariante (ya que son incorrelados) mientras que, para el caso de tener que alterar los residuos, hay que utilizar métodos multivariantes. En el caso de la permutación, la manera que se ha utilizado para modificar los residuos ha sido cambiar directamente el registro en bloque correspondiente a los residuos, pero en otros tipos de modificaciones la complejidad de la modificación aumenta considerablemente.

Por los motivos expuestos, en el análisis de las alternativas de modificación de los factores del apartado anterior, se ha descartado la inclusión de los residuos como un parámetro a estudiar.

Otros conjuntos de datos

Hasta el momento se ha estado trabajando sobre el conjunto de datos Tarragona, pero en este apartado se van a analizar otros dos conjuntos de datos para ver cómo se comporta el análisis factorial con permutación en diferentes escenarios. Todos ellos se describen en el proyecto "CASC Project" (véase Brand et at (2002)).

En primer lugar se va a indicar la variabilidad representada por cada factor, ya que será crucial para ver cómo se comporta el análisis factorial.

Tarragona

Para el conjunto de datos Tarragona se tiene la siguiente tabla:

Factor	Variabilidad representada
1000000000000	0.615937833
0100000000000	0.089932045
0010000000000	0.068564394
0001000000000	0.03768378
0000100000000	0.032004283
0000010000000	0.012049793
0000001000000	0.01128575
0000000100000	0.007939067
0000000010000	0.005311173
0000000001000	0.004336213
0000000000100	0.001571054
0000000000010	0.00034716
0000000000001	*1.00E-30*
Total	**0.886962544**

Tabla 32. Tarragona. Variabilidad por factor

Y los gráficos de las métricas se indican a continuación; a la izquierda se incluye el gráfico de cada métrica y a la derecha la amplitud del intervalo de confianza.

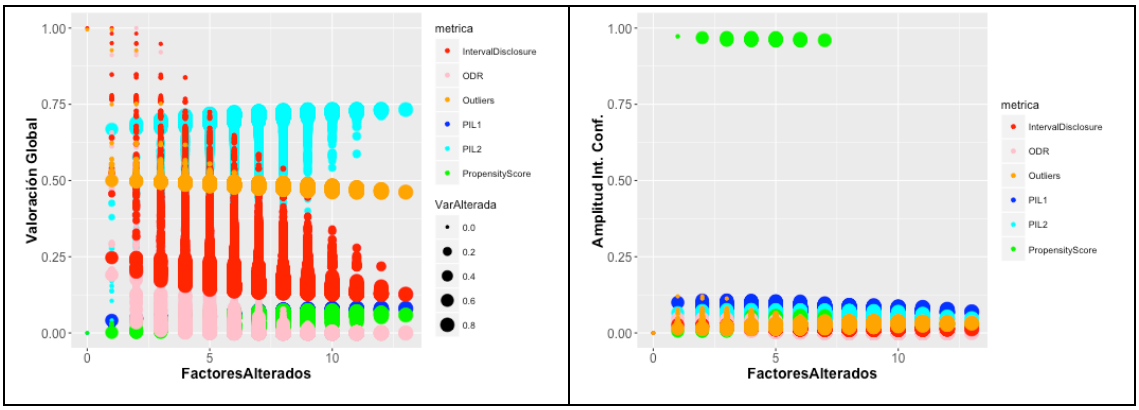


Figura 33. Tarragona. Número de factores

Respecto a la amplitud del intervalo de confianza, cabe comentar que el motivo por el cual había algunas modificaciones que tenían una amplitud mayor es la métrica Propensity Scores que aparece en verde en el gráfico de la derecha. Se observa que hay combinaciones de modificaciones de factores que, si bien mantienen la relación entre variables, se puede encontrar una relación lineal de cada conjunto respecto a sus variables.

EIA

El conjunto de datos EIA (incluido en el paquete sdcMicro) contiene un total de 4092 observaciones y 10 variables continuas.

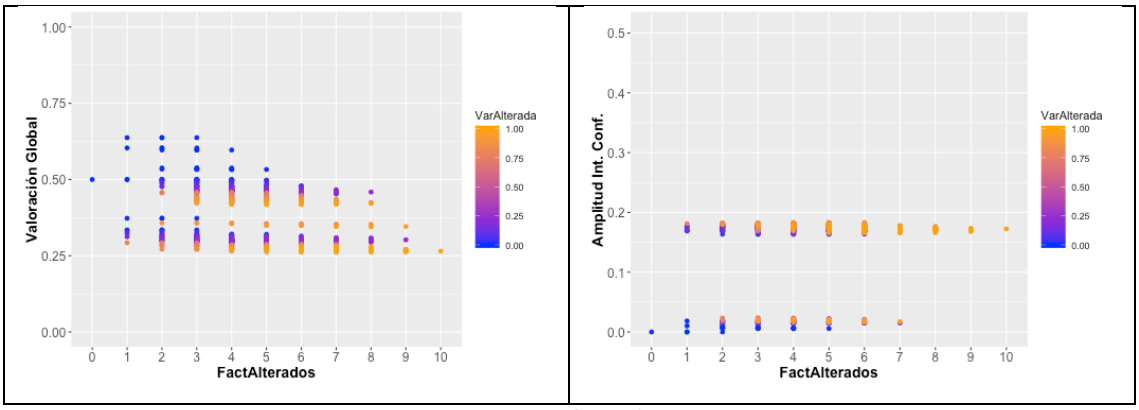


Figura 34. EIA

Se ha tenido que cambiar la escala ya que, en este caso, hay modificaciones que afectan negativamente superando el umbral del 0.5 que se ha establecido de manera general. Este punto podrá verse en el siguiente gráfico donde se representan las métricas de manera individual.

Las modificaciones sobre este conjunto son inestables en la mayoría de los casos en cuanto a la amplitud del intervalo de confianza. Sólo hay 72 combinaciones en las que la amplitud del intervalo de confianza es menor de 0.02; y en los 952 casos restantes la amplitud supera el 0.16.

La variabilidad representada por cada factor se muestra en el siguiente cuadro:

Factor	Variabilidad representada
1000000000	0.760775171
0100000000	0.132651481
0010000000	0.066950435
0001000000	0.022118862
0000100000	0.007505551
0000010000	0.001864102
0000001000	0.000412566
0000000100	0.000201015
0000000010	*1.00E-30*
0000000001	*1.00E-30*
Total	**0.992479183**

Tabla 35. EIA. Variabilidad por factor

En este caso la variabilidad representada por los residuos es prácticamente nula por lo que la alteración de los residuos en este caso será inocua. También se ve que la alteración mediante los factores puede afectar a la práctica totalidad de la información.

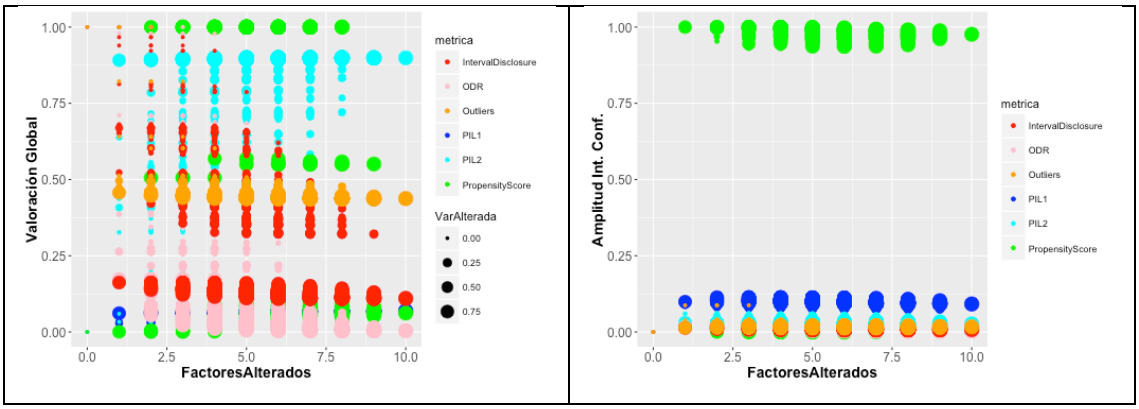


Tabla 36. EIA. Métricas individuales

En el gráfico de las métricas individuales se puede observar que para casi la totalidad de las combinaciones de modificación de factores los valores de Propensity Scores son extremadamente elevados.

El análisis factorial está conservando la relación entre las variables, pero en este caso en concreto, los está alterando de alguna manera que provoca que se pueda discernir la diferencia entre los dos conjuntos (original y modificado). Esto no sucede en los otros dos conjuntos, por lo que podrá ser un aspecto a estudiar en el futuro para identificar las características del conjunto inicial que hacen que el Propensity Scores se mantenga tan cercano a 1.

Los mejores valores para el conjunto EIA son:

Datos	Máscara	Factores modif	Variabilidad modif	Resultado	Int. Confianza
EIA	11111100000	6	0.991865602	**0.262774326**	0.174145892
EIA	01100011000	4	0.200215497	**0.471140072**	0.016286258

Figura 37. EIA. Mejores resultados

El primer resultado con un intervalo de confianza menor de 0.02 tiene un valor de 0.47 que es extremadamente alto comparado con los números que se han estado obteniendo hasta el momento.

CASCrefmicrodata

El conjunto de datos CASCrefmicrodata (incluido en el paquete sdcMicro) contiene un total de 1080 observaciones y 13 variables continuas. Como curiosidad, este conjunto tiene dos variables que son linealmente dependientes, por lo que antes de la aplicación del factorial se ha eliminado la variable que no aportaba información nueva.

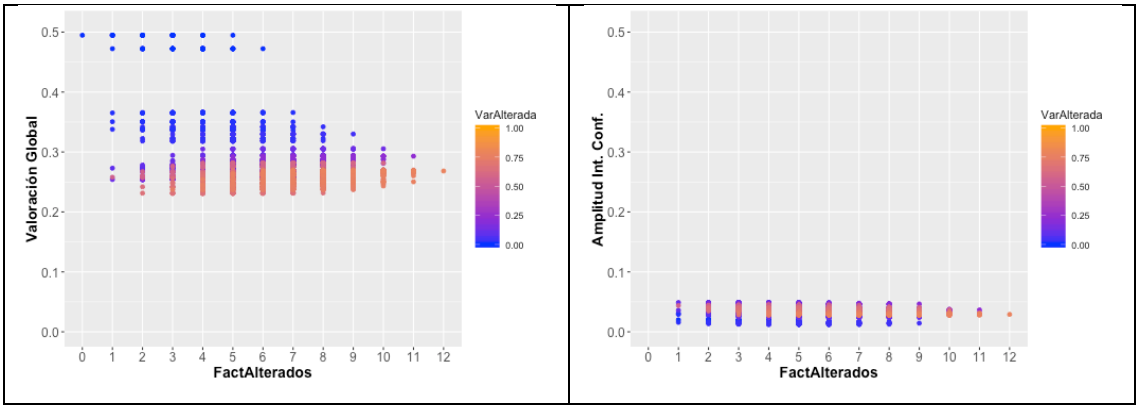


Tabla 38. CASCrefmicrodata

En este conjunto, la amplitud del intervalo de confianza es muy baja, al contrario de lo que sucedía con el conjunto EIA.

La variabilidad representada por cada factor es la que sigue:

Factor	Variabilidad representada
100000000000	0.554673959
010000000000	0.10206025
001000000000	0.082748996
000100000000	0.013855258
000010000000	0.012984091
000001000000	0.010376536
000000100000	0.000486517
000000010000	*1.00E-30*
000000001000	*1.00E-30*
000000000100	*1.00E-30*
000000000010	*1.00E-30*
000000000001	*1.00E-30*
Total	**0.777185607**

Tabla 39. CASC. Variabilidad por factor

Sólo los 7 primeros factores contienen información sobre los datos, el resto se han calculado por homogeneidad ya que es el máximo teórico que podrían existir pero no son útiles en este caso en concreto. Es curioso observar que, aunque quedan residuos con un total del 23% de la variabilidad, a partir del factor 8 ninguno contiene información útil.

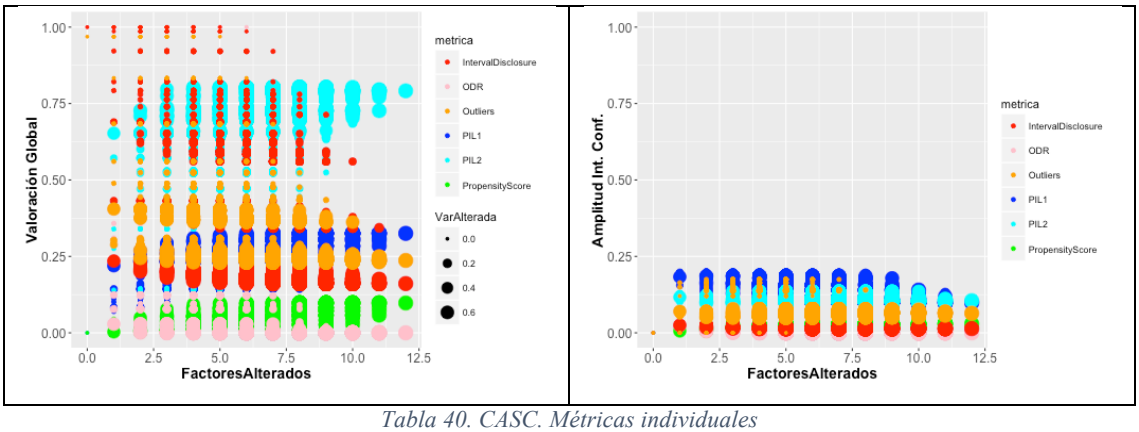


Tabla 40. CASC. Métricas individuales

En este caso, los residuos contienen todavía un 23% de la variabilidad. Además, es el caso en que mejor se comporta el Propensity Scores, ya que en ningún caso tiene un valor cercano a 1. No se ha analizado el motivo de este comportamiento, pero según los tres conjuntos de datos utilizados, parece que, cuanta más información queda en los residuos, más disminuye el Propensity Scores y, por tanto, la amplitud del intervalo de confianza global.

Resumen

Se muestra a continuación el cuadro resumen de las mejores combinaciones de alteración de factores para los conjuntos de datos probados.

Datos	Variabilidad representada	Máscara	Factores modif	Variabilidad modif	Resultado	Int. Confianza
CASC	0.777185607	110000000000	2	0.656734209	**0.23134587**	0.040106326
Tarragona	0.886962544	0001110001000	4	0.086074068	0.232953371	0.012699737
EIA	*0.992479183*	*11111100000*	*6*	*0.991865602*	***0.262774326***	*0.174145892*
EIA	0.992479183	01100011000	4	0.200215497	**0.471140072**	0.016286258

Figura 41. CASC. Número de factores

En los tres conjuntos probados, se puede observar que, cuando hay menos variabilidad hay representada en los factores (y por tanto más en los residuos) la mejor alternativa incluye la alteración de los primeros factores (ej: CASC). Y que cuando hay más variabilidad, la mejor alternativa es mantener los factores de mayor peso inalterados (ej: Tarragona y EIA con intervalo de confianza estrecho).

En el caso de EIA este razonamiento también se cumple para el mejor valor para una amplitud del intervalo de confianza menor a 0.02.

Parece, por tanto, que la variabilidad total del conjunto que representan los factores afectan tanto a los factores a modificar como a la amplitud del intervalo de confianza. No obstante, habría que realizar más pruebas y analizar en detalle si existen otras relaciones entre los factores a modificar y las variables iniciales (por ejemplo si las combinaciones de factores que se pueden modificar no superan un determinado umbral de variabilidad o algún tipo de relación similar). Este punto podría ser una alternativa de estudio a futuro.

10. Comparación con otros métodos

En este apartado se ha comparado el mejor resultado del análisis factorial con permutación con el resto de métodos de anonimización combinados que se han explicado anteriormente. Estos métodos son los que se ha encontrado en la bibliografía como las mejores alternativas en el estado del arte.

Para las siguientes simulaciones se ha utilizado el conjunto de datos Tarragona y el mismo procedimiento con un bucle de 100 iteraciones por cada alternativa de configuración.

Microagregación con ruido

Este método se ha probado empíricamente para todas las variantes existentes en el paquete R, que implementan las modificaciones basadas en:

- Pca: basado en análisis de componentes principales

- Onedims: para cálculos univariantes
- Clustpca: clusterizado con pca
- Rmd: utilizando la distancia de Mahalanobis
- Mdav: utilizando la distancia euclídea
- …

Y, conforme a los resultados de las pruebas realizadas se han elegido los métodos de modificación: mdav y rmd. Estos dos métodos también son los que se referencian en la bibliografía consultada (véase Oganian and Karr (2006)). Además del algoritmo de ordenación y modificación comentado, existe otro parámetro que se ha indicado en apartados anteriores, a saber, k. Este parámetro indica el número de observaciones que se van a microagregar. Los resultados obtenidos se muestran a continuación.

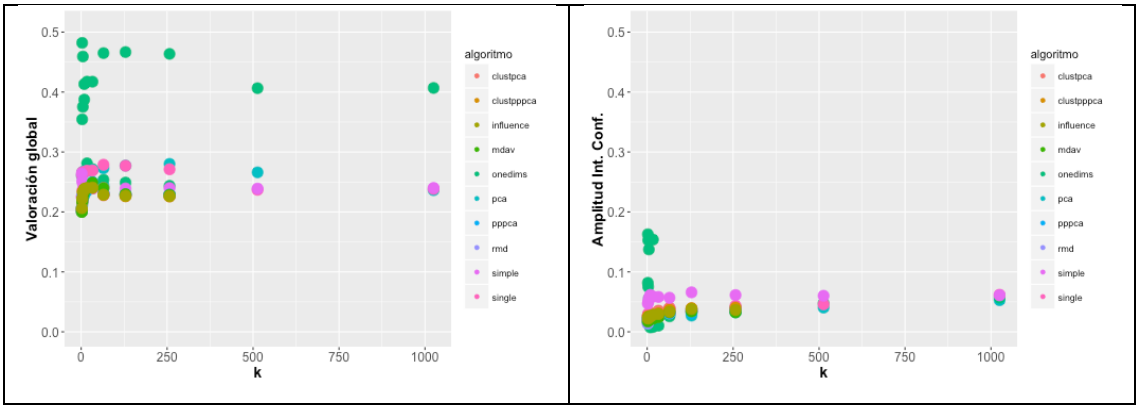


Figura 42. Algoritmos Microagregación con ruido

Tal como se deriva de las gráficas, los mejores resultados se obtienen para valores de k muy pequeños o extremadamente altos y para los métodos rmd y mdav. Los valores de k extremadamente pequeños están primando la utilidad, mientras que los grandes priman la seguridad, como se verá en la siguiente gráfica que representa las métricas de manera individual.

Para la inclusión de ruido, se ha utilizado una distribución normal multivariante con la resta de la variabilidad entre los dos conjuntos (original y microagregado) para restaurar parte de la variabilidad perdida.

En el siguiente gráfico se muestra el desglose de las diferentes medidas de utilidad y seguridad para los métodos elegidos (rmd y mdav) en la columna de la izquierda. En el gráfico de la derecha se muestra la amplitud del intervalo de confianza de las métricas individuales.

Como se puede observar, para valores pequeños de k se mejora la utilidad (colores fríos) y, para valores de k elevados, se mejora la seguridad (colores cálidos).

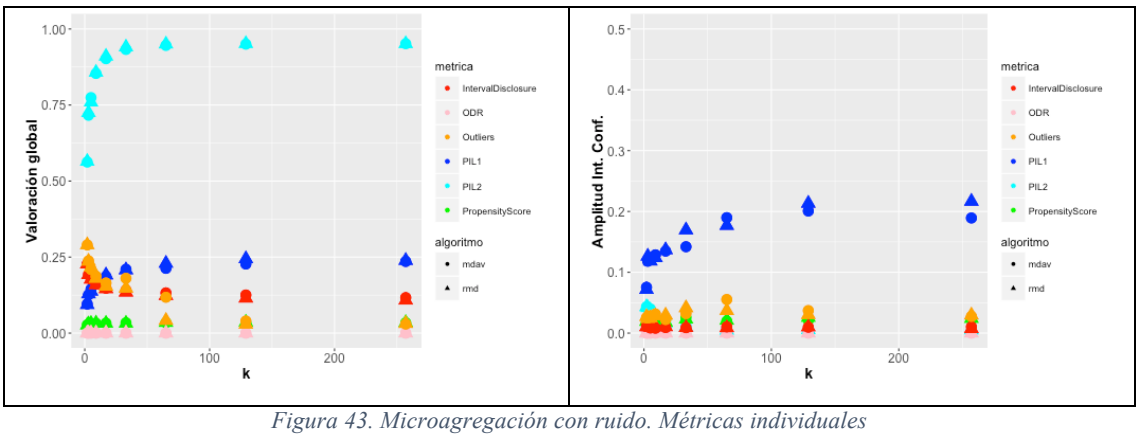


Figura 43. Microagregación con ruido. Métricas individuales

Se puede, además, observar que los resultados de seguridad son, en general, mejores que los de utilidad. Para el caso del PIL2 se tiene el el peor balance ya que empieza elevado y y enseguida llega al peor resultado (cercano a 1) a partir de un valor de k=30 (aprox).

Con el k=3 elegido se tiene un compromiso respecto a todas las medidas, y el PIL2 se mantiene dentro de los mejores valores observados, por lo que se penaliza poco el resultado global.

Análisis de componentes principales (PCA)

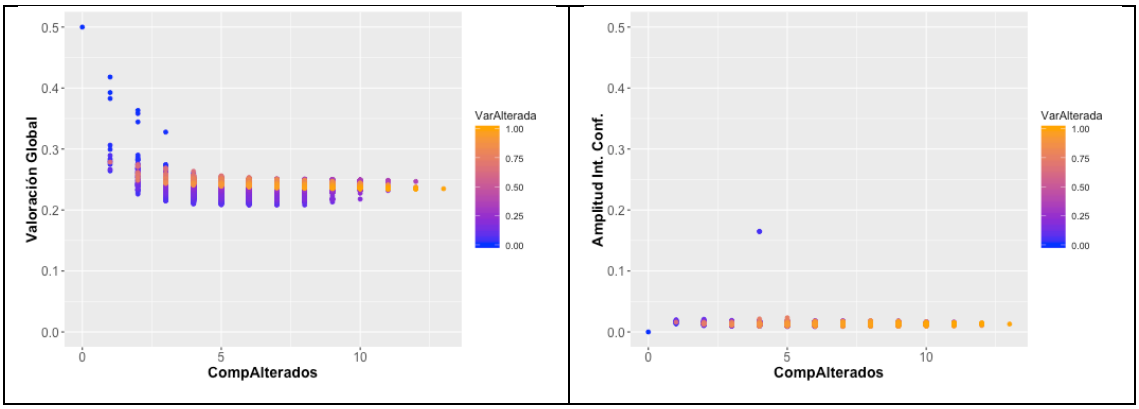


Figura 44. PCA con permutación

La simulación realizada utiliza una aproximación similar al análisis factorial, aunque en este caso en lugar de descomponer en factores se usan los componentes principales. Se puede ver que hay un caso donde la amplitud del intervalo de confianza es superior al resto que se corresponde con un número de componentes modificados inferior a la mitad.

En la gráfica de la izquierda se puede ver que los mejores valores conseguidos por este método, los más bajos, se corresponden a la parte media de componentes alterados y dentro de ese rango aquellos que tienden a tener menos variabilidad alterada.

Estos resultados son muy parecidos a los obtenidos con análisis factorial para el mismo conjunto de datos.

Si se representa la variabilidad en el eje X en lugar del número de componentes alterados, se puede observar los siguientes gráficos.

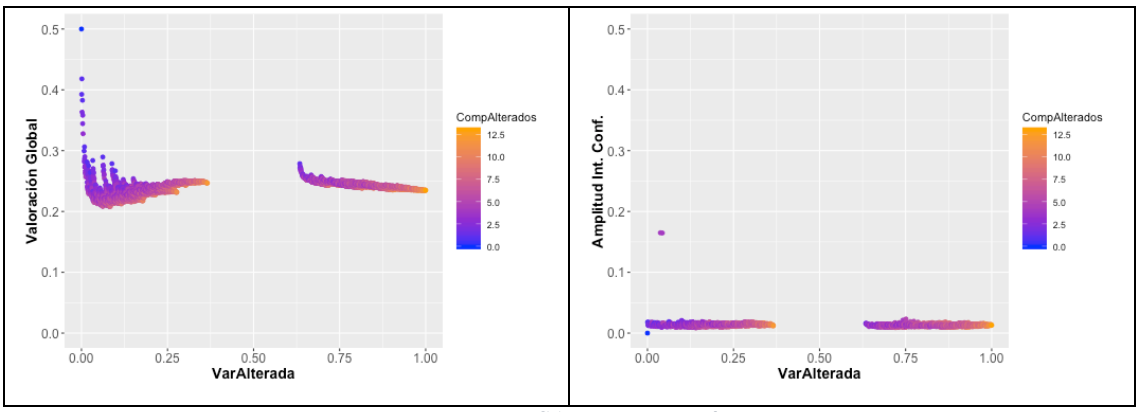


Figura 45. PCA con permutación

Con esta visualización se puede observar que los mejores resultados son aquellos que minimizan la variabilidad a la vez que maximizan los componentes modificados. Esto significa que, en general, será más efectivo modificar muchos de los últimos componentes (que representan menos variabilidad) y pocos de los primeros componentes (que tienen mayor grado de variabilidad representada).

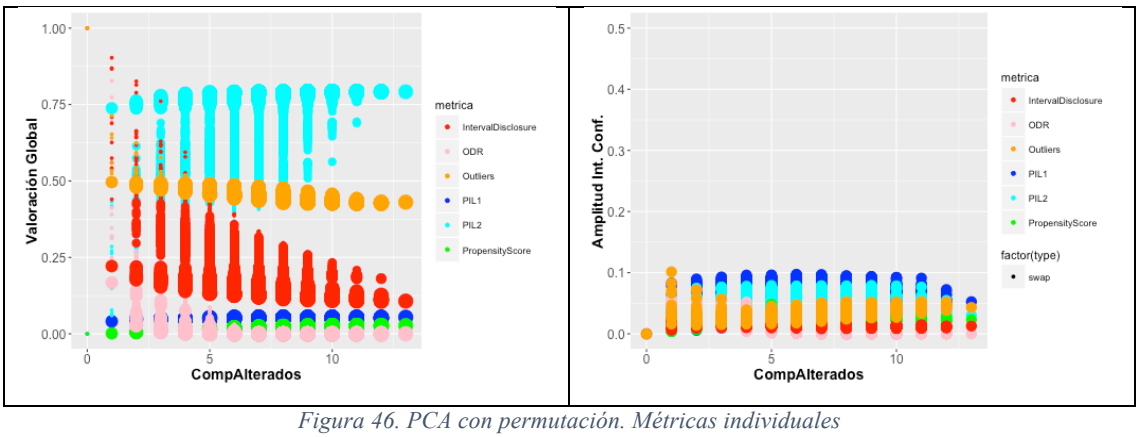


Figura 46. PCA con permutación. Métricas individuales

En cuanto a las métricas individuales, se puede ver cómo el incrementar la variabilidad (representada por el diámetro de las figuras) empeora el PIL2, mejora el Interval Disclosure y también mejora el ODR (aunque en menor medida). Los valores de Outliers, PIL1 y Propensity Scores no parecen estar tan afectados por el incremento de la variabilidad.

Respecto al número de componentes, conforme se incrementa el número de componentes se mejora el ODR, y el Interval Disclosure mientras que se empeora PIL2. El resto de métricas se ven mucho menos afectadas.

En estas gráficas se puede ver porqué los mejores valores globales están en el intervalo medio tanto del número de componentes como de la variabilidad. Esto es porque existe una relación inversa entre PIL2 e Interval Disclosure que se compensa en los valores intermedios. Si se quisiese tener un umbral de seguridad o de utilidad predeterminado se

podrían obtener otros escenarios en que se premiase el cambio de más o menos componentes principales o variabilidad.

Comparación

Se muestra a continuación la tabla de los mejores resultados de las tres alternativas probadas. Se indica en una columna el límite inferior del intervalo de confianza y en la siguiente el límite superior del intervalo de confianza para comparar más fácilmente estos intervalos.

Método	Límite inferior Intervalo de Confianza	Límite superior Intervalo de Confianza	Parámetros
Microagregración con ruido	0.193066874	0.210159149	K=2 Método=mdav
Microagregración con ruido	0.193894322	0.207674913	K=2 Método=rmd
PCA con Permutación	0.20407531	0.215356911	Máscara=0000011111100
Factorial con Permutación	0.227042809	0.239742546	Máscara=0001110001000

Tabla 47. Comparación

Como se puede observar, para el caso de Microagregación/ruido y PCA/permutación los intervalos se solapan, por lo que no se puede concluir que sean diferentes.

En el caso de Factorial/Permutación el resultado es ligeramente peor que en los casos anteriores para este conjunto de datos.

Factorial con Microagregación y ruido

Teniendo en cuenta que la combinación de métodos tiende a generar mejores resultados, se ha probado una triple modificación: factorial y una modificación de los factores mediante microagregación y ruido.

En este caso, tras el cálculo de los factores, se procede a calcular una microagregación de cada uno de los factores elegidos de manera independiente y, a continuación, se incluye un ruido potencialmente equivalente a la variabilidad perdida con la microagregación. La elección de los parámetros se ha basado en la bibliografía (véase Oganian and Karr (2006)) donde aparece k=3 como el mejor de los resultados y en las pruebas realizadas en el apartado anterior (donde se obtenían los mejores resultados para k pequeños y con el método rmd). Para el número de factores se ha elegido descomponer en 7 factores que es lo se ha utilizado para comparar los métodos de modificación.

En este caso, los resultados obtenidos son los siguientes:

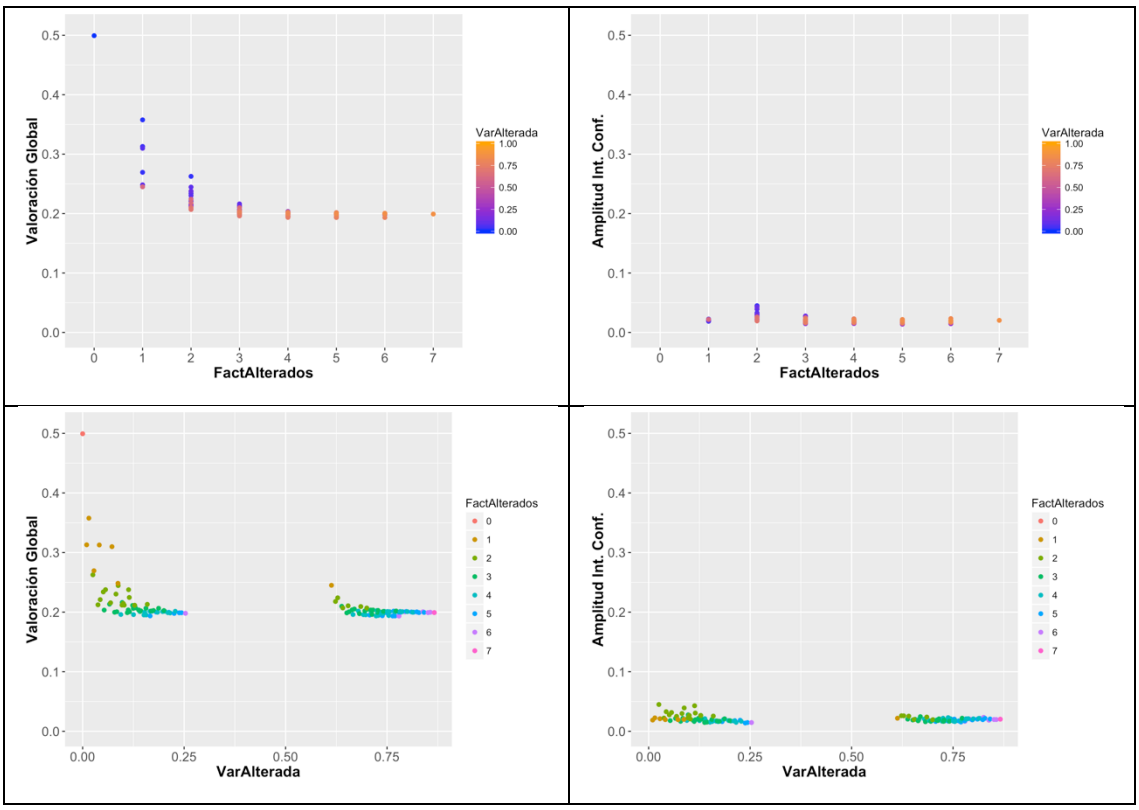


Figura 48. Factorial con microagregación y ruido

Se puede observar que en algunas configuraciones se obtienen valores menores de 0.2 por lo que se estaría mejorando tanto la mejor combinación de factorial con permutación como los métodos combinados de microagregación/ruido y PCA/permutación.

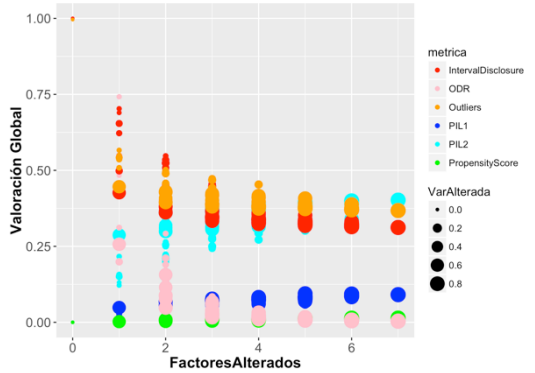


Figura 49. Factorial con microagregación y ruido. Métricas individuales

En este caso, se tiene que ninguna de las métricas individuales supera el valor 0.5 con configuraciones donde se altera mucha variabilidad (es decir, se modifican los primeros factores)

Además, la métrica del Propensity Scores y PIL2 que son las que más sufrían con otros métodos se mantienen por debajo de 0.5.

No obstante, esta triple combinación incrementa los tiempos de computación, aunque tiene la ventaja de utilizar técnicas univariantes para la aplicación de la microagregación.

Método	Q2	Q98	Configuración
Factorial con Microagregración con ruido	0.185272692	0.202023312	Máscara= 0000001011111 K=3 y rmd

Tabla 50. Resumen factorial con microagregación con ruido

El intervalo que se genera con esta combinación queda por debajo del definido tanto por PCA/permutación como factorial/permutación. No obstante, no se puede concluir que sea mejor que microagregación/ruido ya que se solapan los intervalos. En este punto, cabe comentar, que para el factorial/microagregación/ruido no se han probado todas las combinaciones de modificaciones (se ha acotado a K=3 y rmd) por lo que cabría esperar que para otras combinaciones de K o de algoritmo de microagregación se pueda mejorar el resultado.

11. Protección selectiva

Hasta el momento el trabajo se ha centrado en la protección general de todas las variables continuas pero, en la mayoría de los casos, sólo algunas de las variables tienen un riesgo de exposición que se considera necesario proteger. En estos casos, tras determinar el umbral para clasificar las variables sensibles, se puede proceder a proteger sólo las variables elegidas, y no el conjunto completo de todas las variables. Este criterio, a priori, aportará mayor utilidad al no modificar aquellas variables que no se considera necesario proteger.

Criterios de selección

Para poder comparar esta casuística, no basta con tener en cuenta las métricas anteriores, sino que habrá que incluir algún tipo de métrica adicional que determine la precisión de las modificaciones realizadas. Por tanto, se tendrán las siguientes dimensiones:

- Utilidad del conjunto resultante que se puede medir con las mismas métricas que las indicadas para el caso general: PIL1, PIL2 y Propensity Scores.
- Seguridad: que medirá el nivel de protección aunque sólo se valorará la seguridad de las variables elegidas ya que, para el resto de variables, no se ha considerado la necesidad de protegerlas. Se pueden utilizar las mismas métricas, pero sólo se medirán sobre las variables que se deban proteger.
- Especificidad/Precisión: en esta dimensión se medirá cuan acotadas están las modificaciones en las variables sensibles, de manera que sólo se alteren éstas y no se modifiquen las que no sea necesario.

Para la especificidad se ha incluido un algoritmo para calcular cuánto se parecen las variables del conjunto modificado a las iniciales. Para ello se ha utilizado la correlación de Pearson cogiendo sólo la diagonal, que es la relación de una variable consigo misma en los dos conjuntos. Para las variables que se deben proteger debería ser cercana a 0 de manera que no se parezcan. Para las variables no sensibles debería ser cercana a 1. Por

tanto, se suman las distancias hasta el objetivo (a saber 1 para no sensibles, 0 para sensibles) y esa medida informa de cuanta precisión tiene el método. Cuanta mayor precisión (sólo se modifican las variables a proteger) más cercano a 0 será. Además esta métrica tiene en cuenta tanto el modificar las variables sensibles como el no modificar el resto.

Factorial

Una ventaja del análisis factorial es que los factores son ortogonales entre sí, por lo que se pueden rotar. En general la rotación se realiza para conseguir determinadas propiedades como la menor distancia entre factores y variables, pero en este caso se ha aprovechado para mejorar la métrica de precisión que se ha comentado en el apartado anterior.

Se han efectuado tres rotaciones (y la subsiguiente selección de factores):

- Máximo cambio de las variables sensibles
- Mínimo cambio de las variables no sensibles
- Y un valor intermedio entre las dos alternativas

PCA

En el caso de PCA, se ha de elegir qué componentes son los que se modifican para, además de obtener buenos resultados de seguridad y utilidad, éstos sean precisos y focalizados sólo en las variables a proteger.

Para ello, se han ordenado los componentes según la afectación a cada variable sensible y, se han elegido componentes consecutivamente hasta alcanzar el 80% de representación de cada variable a proteger.

Microagregación y ruido

Para la microagregación se ha utilizado la función microaggregation de la librería sdcMicro de R donde se pueden indicar expresamente las variables que se quieren modificar de manera que ,en función del algoritmo, se tiene en cuenta tanto las variables a proteger como las relaciones con el resto.

Para el paso de incorporación del ruido, se ha utilizado una aproximación empírica para incorporar ruido multivariante a todas las variables. Si se hubiese incluido ruido multivariante sólo a las variables modificadas se hubiese podido perder la relación entre las variables. Para evitar esta pérdida de relación pero minimizar el ruido incorporado en las variables que no se habían modificado, se ha utilizado la aproximación de la matriz positiva más próxima (respecto a la matriz de la varianza que se hubiese eliminado con la microagregación).

Resultados

Se han escogido aleatoriamente dos variables a proteger del conjunto de Tarragona, y para ellas se han aplicado los métodos: factorial, pca y microagregación/ruido.

Variables: FIXED.ASSETS CURRENT.ASSETS

Método	Límite inferior IC	Límite superior IC	Parámetros
Microagregración con ruido	**0.199826062**	**0.223786152**	**Rmd** **K=2**
	0.201123464	0.364039211	Rmd K=3
	0.329299997	0.353111061	Rmd K=5
PCA con Permutación	0.270808035	0.290837714	1010000000000
Factorial con Permutación	0.256519454	0.286328619	1100000000000
	0.288000364	0.32633763	1111111110000
	0.305612441	0.322789699	0010000110000

Tabla 51. Protección selectiva. FIXED.ASSETS CURRENT.ASSETS

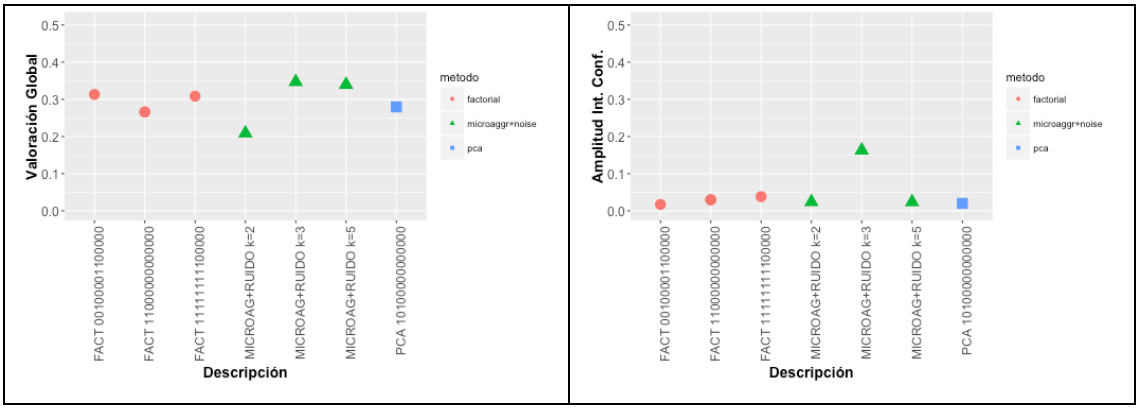


Figura 52. FIXED.ASSETS CURRENT.ASSETS

Las gráficas anteriores representan la media entre las medidas de: seguridad, utilidad y precisión y la amplitud del intervalo de confianza de esas tres medidas. En el resultado general se ha representado el mejor resultado de factorial (círculo rojo), microagregación/ruido (triángulo verde) y PCA (cuadrado azul). De los mejores resultados, la configuración de microagregación/ruido elegidas parece afectar mucho a los resultados y al intervalo. Para factorial y PCA los valores parecen ser más estables.

Para estas variables, los mejores resultados se obtienen con microagregación/ruido con una configuración de k=2 y rmd.

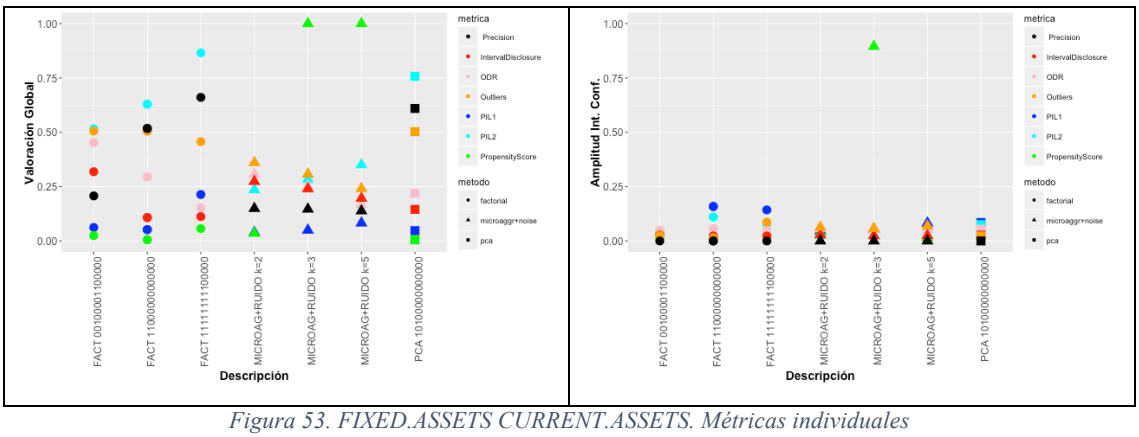


Figura 53. FIXED.ASSETS CURRENT.ASSETS. Métricas individuales

En el gráfico con las métricas individuales se ha incorporado (en negro) la precisión a las métricas que había definidas con anterioridad de seguridad y utilidad.

Se ha detectado que hay un intervalo que se dispara en la gráfica general con respecto a microagregación/ruido. Ahora, en esta gráfica, se observa que corresponde con la métrica de Propensity Scores con una configuración k=3 y rmd. De hecho se ha tenido que cambiar la escala de la amplitud al doble de lo habitual para poder representar ese punto.

En el análisis factorial se puede ver cómo la precisión va empeorando conforme se rotan e incorporan más factores a modificar.

Variables: NET.PROFIT TREASURY

Método	Límite inferior IC	Límite superior IC	Parámetros
Microagregración con ruido	0.380306405	0.390328831	Rmd K=2
	0.379886354	0.391588893	Rmd K=3
	0.35101177	0.362690575	Rmd K=5
PCA con Permutación	**0.248387461**	**0.271193586**	**1110010000000**
Factorial con Permutación	**0.26723787**	**0.294611684**	**1100000000000**
	0.257591191	**0.294864222**	**1111110000000**
	0.364273085	0.373994799	0001000000000

Tabla 54. Protección selectiva. NET.PROFIT TREASURY

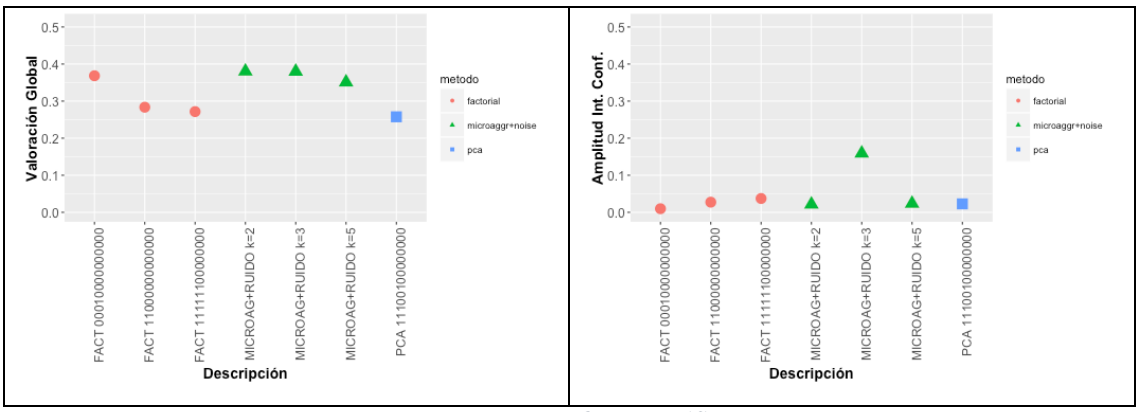


Figura 55. NET.PROFIT TREASURY

En este caso los mejores valores se obtienen con PCA y con factorial. En cambio el resultado de microagregación/ruido es peor con esta selección de variables.

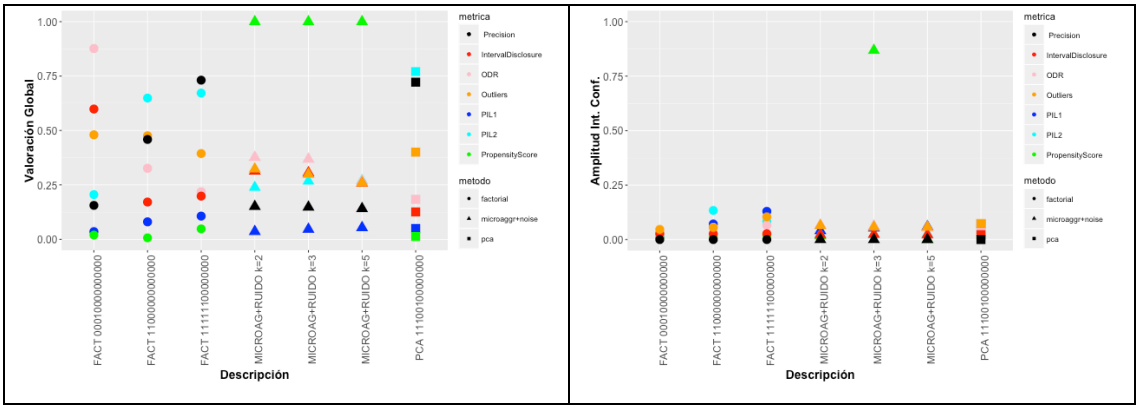


Figura 56. NET.PROFIT TREASURY. Métricas individuales

Como se puede observar, la métrica Propensity Scores (en verde) es la que parece afectar negativamente a los resultados de microagregación/ruido.

12. Conclusiones y trabajo futuro

En este trabajo se ha estudiado el análisis factorial como método estadístico para anonimizar conjunto de datos aprovechándose de las propiedades de ortogonalidad de los factores. Se han comparado las diferentes alternativas de modificación así como las combinaciones de factores que han aportado los mejores resultados. Todo ello se ha podido comparar con otras técnicas utilizadas actualmente como son microagregación con ruido y la técnica PCA con permutación que se ha utilizado como base para el presente estudio, obteniendo unos resultados, en algunos escenarios de la protección selectiva, comparables con PCA y mejores que la microagregación con ruido.

Tal como se dijo al principio del estudio, dependiendo del conjunto de datos funcionan mejor unos métodos que otros, y como se ha visto, el análisis factorial aporta buenos resultados en concreto para la protección selectiva donde, mediante la rotación de

factores, se puede conseguir unos resultados más precisos en cuanto a las variables a proteger.

No obstante, han quedado puntos interesantes para seguir analizando en posteriores trabajos que no se han podido analizar en profundidad en este estudio, como son:

- Correlación de variables y como afectan a la amplitud del intervalo de confianza.
- Influencia de la existencia de residuos y cómo afectan al intervalo de confianza y a las métricas.
- Propiedades de las alteraciones que afectan a la métrica de Propensity Scores.
- Protección selectiva: análisis en profundidad de las posibles rotaciones y la selección de factores.
- Profundizar en métodos de triple modificación: factorial + microagregación + ruido.
- Comparación de complejidad de los algoritmos y sus tiempos.
- Generación de datos sintéticos.

13. Bibliografía/Referencias

2009. "Reglamento (CE) Nº 223/2009 del Parlamento Europeo y del Consejo de 11 de marzo de 2009". Diario Oficial de la Unión Europea

Aida Calviño. 2015. "A simple method for limiting disclosure in continuous microdata based on principal component analysis". Pendiente de publicación

Josep M. Mateo-Sanz, Josep Domingo-Ferrer, Francesc Sebé. 2005. "Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata". Data Mining and Knowledge Discovery

Josep Maria Mateo-Sanz, Francesc Sebé, and Josep Domingo-Ferrer. 2004. "Outlier Protection in Continuous Microdata Masking". Lecture Notes in Computer Science

Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, Peter-Paul de Wolf. 2012. "Statistical Disclosure Control". Ed: Wiley

Ruth Brand, Josep Domingo-Ferrer, Josep M. Mateo-Sanz. 2002. "Reference data sets to test and compare SDC methods for protection of numerial microdata". Computational Aspects of Statistical Confidentiality (CASC) Project

Anna Oganian and Alan F. Karr. 2006. "Combinations of SDC Methods for Microdata Protection". Lecture Notes in Computer Science

Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation". The Journal of Privacy and Confidentiality

Matthias Templ, Bernhard Meindl and Alexander Kowarik . 2015. "Introduction to Statistical Disclosure Control (SDC)"

Matthias Templ, Alexander Kowarik, Bernhard Meindl. 2015. "Package sdcMicro"

Ruth Brand. 2002. "Microdata Protection through Noise Addition". Lecture Notes in Computer Science

14. Índice de Tablas y Figuras