UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

EUROPEAN MASTER OF OFFICIAL STATISTICS

TÉCNICAS DE MINERÍA DE DATOS APLICADAS A LA ESTADÍSTICA DE CONVENIOS COLECTIVOS DE TRABAJO

Por:

Autor: SANTIAGO HERNÁNDEZ GARCÍA

Tutor: CARLOS LAMENCA MARTÍNEZ

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

FACULTAD DE CIENCIAS MATEMÁTICAS.

TRABAJO DE FIN DE MÁSTER

Madrid, 25 de Agosto de 2019

UNIVERSIDAD COMPLUTENSE DE MADRID

Índice

Abstract

Data mining is a developing field inside the statistical analysis. His aim is to discover hidden associations, relations, groupings or classifications which we cannot discover with the naked eye. The objective of this Master Thesis is to treat public data with these techniques in order to generate value added to the excellent information that the public administration has. Motivated by the placement of internships in the Ministry of Labour, Migrations and Social Security , this thesis will use the data produced in the General Subdirectorate of Statistics and socio-labour analysis to apply the techniques of data mining, specifically to the statistical production of collective agreements, in which the author of the thesis has done the internship.

Keywords: *Data mining, public sector, collective agreement, clustering, correspondence analysis.*

Resumen

La minería de datos es un campo en desarrollo dentro del análisis estadístico. Su objetivo es descubrir asociaciones, relaciones, agrupaciones o clasificaciones ocultas que no podemos descubrir a simple vista. El objetivo de este Trabajo de Fin de Máster es tratar los datos públicos con estas técnicas para generar valor añadido a la excelente información que tiene la administración pública. Motivado por la estancia de prácticas en el Ministerio de Trabajo, Migraciones y Seguridad Social, esta tesis utilizará los datos producidos en la Subdirección General de Estadísticas y Análisis Sociolaboral para aplicar las técnicas de minería de datos, específicamente a la producción estadística de convenios colectivos, en los cuales el autor ha participado en su producción durante su estancia en prácticas.

Palabras clave: *Minería de datos, sector público, convenio colectivo, análisis de conglomerados, análisis de correspondencias.*

Introducción.

La Estadística Oficial es una herramienta imprescindible en cualquier sociedad moderna actual. Todos los países tienen órganos encargados en la producción y difusión de estadísticas que, luego servirán como vehículo de transparencia y rendición de cuentas de los políticos. España no puede ser menos y tiene unidades estadísticas en cada ministerio para producir información vinculada a las actividades que se desarrollan en cada uno de ellos. Presente documento, motivado por la estancia de prácticas en la Subdirección General de Estadística y Análisis Sociolaboral del Ministerio de Trabajo, Migración y Seguridad Social, tratará de describir el trabajo que se desarrolla en esta unidad, en concreto con la Estadística de Convenios Colectivos de Trabajo – A partir de aquí CCT-, para permitir al lector tener una mejor composición de la estadística que aquí se produce. Así, posteriormente, seguir con el grueso del estudio que ocupa el Trabajo de Fin de Máster, que consiste en aplicar técnicas de minería de datos con el fin de encontrar patrones y relaciones ocultas a simple vista.

La principal motivación del estudio es describir el proceso de producción estadística que se lleva a cabo en el Ministerio, concretamente en la Subdirección General de Estadística y Análisis sociolaboral, donde se produce la estadística de Convenios Colectivos del Trabajo. Se hará hincapié en reglamentación legal, fuentes y metodología. Con la intención de tener una mejor información sobre el proceso de producción e intentar aportar elementos de mejora a éste. Dicha estadística se encarga de una producción concreta, en el caso de la de convenios colectivos (recogida-depuración-publicación-difusión), por lo que otra vía de desarrollo de mi presente estudio no va a ser solo evaluar la actividad que ya realizan, sino además intentar hacer un tratamiento y explotación de los datos, con la idea de sacar información relevante sobre ella.

La estadística de CCT, es una estadística ampliamente demandada por su utilidad. Reúne información muy útil que podemos aglutinar en cuatro pilares:

- Convenios colectivos de trabajo firmados en un año y trabajadores afectados por ellos.

- Aumentos salariales pactados en dichos convenios, lo que permite hacer una difuminada anticipación de la subida salarial pactada en convenio que habrá en venideros años.

- Inaplicaciones, figura que aparece en el artículo 82.3 del Estatuto de los Trabajadores (Real Decreto Legislativo 2/2015, de 23 de octubre).

"Cuando concurran causas económicas, técnicas, organizativas o de producción, por acuerdo entre la empresa y los representantes de los trabajadores legitimados para negociar un convenio […], se podrá inaplicar en la empresa las condiciones de trabajo previstas en el convenio colectivo aplicable, sea este de sector o de empresa, que afecten a las siguientes materias:

  - ***a)*** *Jornada de trabajo.*
  - ***b)*** *Horario y la distribución del tiempo de trabajo.*
  - ***c)*** *Régimen de trabajo a turnos.*
  - ***d)*** *Sistema de remuneración y cuantía salarial.*
  - ***e)*** *Sistema de trabajo y rendimiento.*
  - ***f)*** *Funciones, cuando excedan de los límites que para la movilidad funcional prevé el artículo 39 de esta Ley.*
  - ***g)*** *Mejoras voluntarias de la acción protectora de la Seguridad Social."* (Real Decreto Legislativo 2/2015, de 23 de octubre).

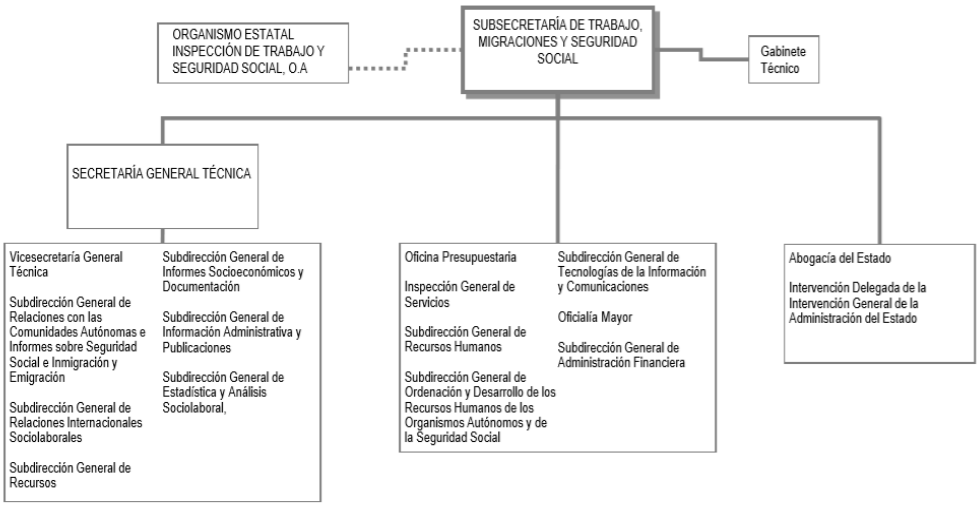
- Batería de variables cualitativas o agrupación como CNAE, provincia, naturaleza del CCT, etc.

Estos cuatro pilares de información convierten la estadística de CCT producida por el MIATRAMIS en un elemento muy demandado por Gobierno, Parlamento, sindicatos, CC.AA o diputaciones provinciales, entre otros organismos públicos y privados. Los datos publicados, que son accesibles para cualquier ciudadano, están presentados en tablas agregadas, lo que complicaría aplicar la minería de datos con una información presentada de tal manera. Por este motivo, se he realizado una petición formal con el objetivo de obtener dicha información en el formato de micro-dato, más accesible al tratamiento y análisis estadístico.

Más adelante se analizará de forma más pormenorizada el contenido de las variables obtenidas a partir de la estancia en la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS para conocer más en profundidad la materia prima con la que se trabajará en presente investigación.

Centro de prácticas y de producción de la estadística de CCT.

El centro de prácticas se ha tratado de la Subdirección General de Estadística del Ministerio de Trabajo, Migración y Seguridad Social. Concretamente, el departamento de producción de estadísticas de convenios colectivos del trabajo.



Fuente: Ministerio de Trabajo, Migracion y Seguridad Social (MITRAMISS)

La Subdirección General de Estadística y Análisis Sociolaboral es un cuerpo dependiente de la Secretaría General Técnica, que ésta a su vez pertenece a la Subsecretaría de Trabajo, Migraciones y Seguridad Social, presidida por Raúl Riesco Roche.

Las **funciones** de la Subdirección General de Estadística y Análisis Sociolaboral según el Real Decreto 903/2018, de 20 de julio, por el que se desarrolla la estructura orgánica básica del Ministerio de Trabajo, Migraciones y Seguridad Social son las siguientes:

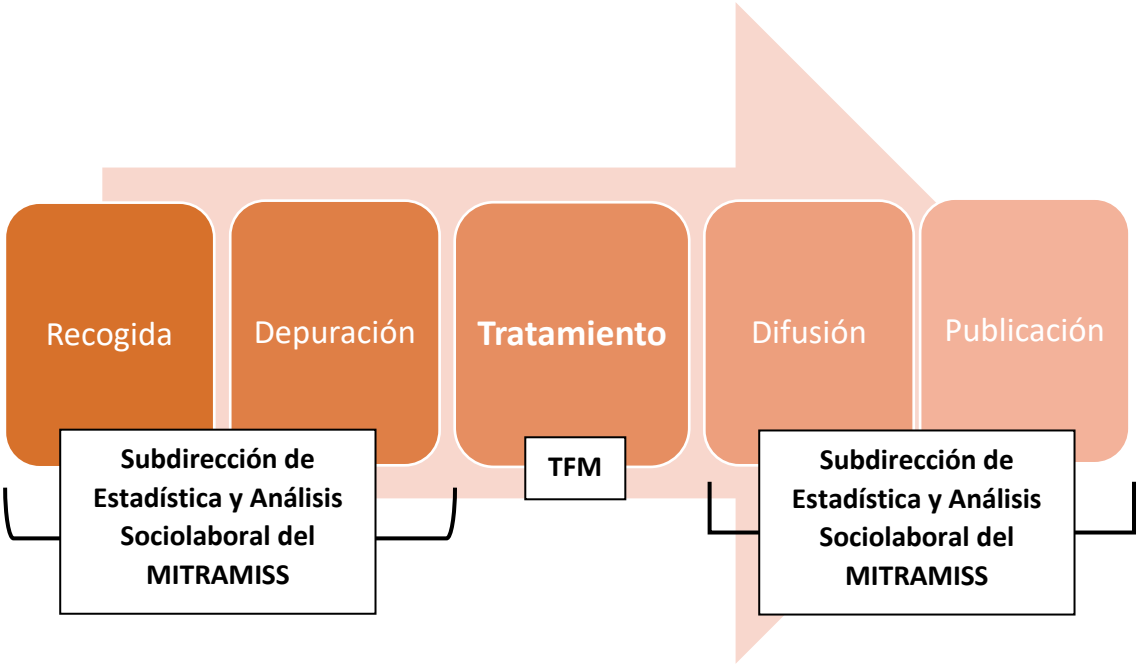
- Desarrollo del Plan Estadístico Nacional y sus correspondientes programas anuales.

- Coordinación a través de la Comisión Ministerial de Estadística y también con el Instituto Nacional de Estadística
- Difusión pública de las estadísticas producidas
- Realización de estudios e informes a partir de los datos estadísticos producidos
- Explotación de la base de datos y desarrollo de indicadores vinculados al mercado de trabajo, la Seguridad Social y los movimientos migratorios.

Proceso de producción de la estadística de CCT.

La idea de presente trabajo es poner en valor la extraordinaria información que tienen en la Subdirección de Estadística y Análisis Sociolaboral, mediante el tratamiento de los micro-datos, con los cuales elaboran los tabulados de la información agregada que posteriormente publican. Mientras la unidad estadística se queda en el proceso de publicación y difusión, este trabajo querrá ir más allá, llegando al proceso de tratamiento de la excelente materia prima que se produce en la unidad estadística del MITRAMISS.

Gráfico 1.



Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

Recogida

Los datos oficiales a veces tienen particularidades con respecto a otra forma de recogida de datos. Las administraciones públicas poseen innumerables registros administrativos que contienen información sobre el ciudadano, los cuales se pueden operacionalizar y

estructurar en una base de datos sobre la que se pueda trabajar. En la producción de estadística de CCT este es el medio de recogida de los datos, el registro administrativo.

Cada vez que las dos partes negociadoras culminan un acuerdo de convenio, este debe ser registrado en REGCON y publicado en el BOE:

"***Primero.*** *Ordenar la inscripción del citado acuerdo en el correspondiente Registro de convenios y acuerdos colectivos de trabajo con funcionamiento a través de medios electrónicos de este Centro Directivo, con notificación a la Comisión Negociadora.*

Segundo*. Disponer su publicación en el «Boletín Oficial del Estado».”* (Real Decreto 713/2010)

Además de registrar el correspondiente texto de negociación y su publicación en el BOE, la Comisión Negociadora debe cumplimentar obligatoriamente una hoja estadística, con el fin de disponer de una tabulación del texto del convenio y poder producir la estadística de CCT. Así lo dispone el Real Decreto 713/2010, de 28 de mayo, sobre registro y depósito de convenios y acuerdos colectivos de trabajo.

“Art.3. Asimismo, se deberán cumplimentar los datos estadísticos recogidos en los modelos oficiales que figuran en el anexo 2 de este real decreto, a efectos de elaboración de la estadística de convenios colectivos.”

Hay dos hojas estadísticas, una relativa a convenios de empresa y otra para convenios de ámbito superior a la empresa (sector,). Además, los anexos de este BOE incluyen la revisión por cláusula de garantía salarial y la revisión salarial anual o plurianual. Esta hoja estadística se registra en una base de datos pública (https://expinterweb.empleo.gob.es/regcon/), en la que se necesitan credenciales para poder registrar o modificar algún tipo de convenio. La visualización de los convenios está abierta a todo tipo de ciudadano a través de su pestaña “Consulta pública”.

Depuración

El proceso de depuración de los datos no es más que una verificación de que la Comisión Negociadora ha cumplimentado bien la hoja estadística y se corresponde con la información pactada en el texto, se desarrolla principalmente bajo dos procedimientos:

- Primeramente, hay personal encargado de encontrar inconsistencias diariamente en los convenios colectivos que se van registrando en la base de datos REGCON. Esto implica un proceso de verificación manual diario de los datos que van entrando a la base de datos. Dependiendo del mes, el volumen de registros diario varía entre los 10 y los 100 registros que se pueden revisar diariamente en la producción de CCT. El técnico contrasta si lo pactado en el texto del convenio corresponde con lo cumplimentado en la hoja estadística que tiene que rellenar la Comisión Negociadora.

- Además de este proceso diario de revisión de los datos, en la Subdirección General de Estadística y Análisis Sociolaboral tienen procesos de detección de errores automático mediante código SAS. Es imprescindible esta etapa del proceso para detectar aquellos errores que no han podido ser identificados mediante la revisión diaria de los datos. El programa desarrolla algoritmos para encontrar incongruencias, inconsistencias o valores atípicos como: que haya más trabajadores que empresas a las que ampara el convenio, convenios con la misma variable identificativa, fechas de vigencia solapadas, vigencias extremadamente cortas, pacto de un número excesivo de días de vacaciones, etc. Este proceso automático se repite varias veces durante el mes con el fin de que cuando haya finalizado se pueda presentar la publicación mensual sin errores detectados en los datos.

Para cumplimentar un dato erróneo los técnicos se vuelven a poner en contacto con las empresas o comisiones negociadoras para que se lo puedan facilitar y así corregirlo. No se realiza ninguna técnica de imputación de los datos. Si finalmente no se ha podido cumplimentar ese dato, esa unidad de observación no entra en la estadística para ese mes concreto.

Publicación

Una vez se ha desarrollado el arduo proceso de depuración-revisión, que es en lo que más tiempo ocupan los técnicos, se tabulan y agregan los datos para maquetar la información en un formato asequible y accesible al público. Este proceso de tabulación y maquetación se hace mediante programa SAS, sobre el cual se han diseñado unas plantillas que tienen la misma estructura todos los meses. Así, en la unidad se garantizan la consistencia y

comparabilidad de los datos a lo largo del tiempo. Se realizan publicaciones mensuales y anuales, estructuradas en cuatro pilares básicos:

1. Datos sobre los **convenios colectivos registrados** en un lapso determinado (mes o año), además de las empresas (si es de ámbito superior a la empresa) y trabajadores que acoge.
2. Datos relativos a los **efectos económicos** que se pactan en el convenio, como las subidas salariales pactadas, ajustes salariales, tipo de remuneraciones, pluses salariales, etc.
3. Inaplicaciones, antiguamente llamadas descuelgues, que no es más que alguna cláusula del convenio que no se va a ejecutar. Se recogen las inaplicaciones registradas, tipo de inaplicación, etc.
4. También se produce una publicación en serie temporal relativa a los datos mencionados en los puntos anteriores. Se recogen datos para los doce meses del año y años anteriores.

Los datos, presentados en hojas Excel, van acompañados de Powers Points que resumen la información presentada en tablas. Este es el enlace web donde se suben las tablas producidas por la unidad (http://www.mitramiss.gob.es/estadisticas/cct/welcome.htm)

Difusión

Los datos publicados, que contienen información muy valiosa de pactos salariales, pactos vacacionales, mejoras laborales, etc. Tienen gran importancia para algunos organismos. La unidad estadística recibe una serie de peticiones (de sindicatos, provincias, CCAA, gobierno, parlamento, etc.) en las que se solicita información concreta de la estadística de CCT. Para estas peticiones la unidad elabora un tabulado especial maquetando así la información que les requiere cada organismo. Como ejemplo, el parlamento hace peticiones expresas a la unidad para que le proporcione datos concretos sobre la estadística CCT. Estos datos se maquetan y se envían puesto que es información no accesible mediante la publicación mensual ordinaria.

También se desarrollan peticiones más concretas de cesión de datos, como en el caso de este Trabajo de Fin de Master, que para poder trabajar con los micro-datos que en el MITRAMISS se producen se ha de desarrollar una petición formal. Si es con un motivo justificado (académico, en el caso del documento investigativo), la unidad podrá realizar

una cesión bajo unas condiciones previas que son formalizadas en un contrato de cesión de datos.

Plan Estadístico Nacional y la estadística de CCT.

Como muchas de las estadísticas elaboradas oficialmente por los organismos estatales españoles, la estadística de CCT, está incluida en el Plan Estadístico nacional (2017-2020). Según la naturaleza del Real Decreto 1043/2017 el Plan Estadístico Nacional pretende:

"La actividad estadística para fines estatales, que es la incluida en el Plan Estadístico Nacional, está orientada a satisfacer las necesidades de los usuarios ofreciendo productos estadísticos de calidad en un entorno en que son cada vez más esenciales para la toma de decisiones. Han de contemplarse, en este sentido, tanto las necesidades de los usuarios institucionales (Administraciones Públicas y Unión Europea, principalmente), con sus demandas estadísticas para la determinación, seguimiento y evaluación de sus políticas públicas, como las necesidades de los agentes sociales, organizaciones, empresas, investigadores, analistas, prensa, etc." (Real Decreto 1043/2017, de 22 de diciembre, por el que se aprueba el Programa anual 2018 del Plan Estadístico Nacional 2017-2020.)

La estadística de CCT tiene la identificación 7407 en el Plan Estadístico Nacional y la esencia del documento establece que esta es promovida para convertir en estadística los registros de convenio guardados en la aplicación REGCON con el fin de difundir la información (Real Decreto 1043/2017). Mencionado Plan Estadístico Nacional acoge así la tarea que debe realizar el MESS (Ministerio de Empleo y Seguridad Social), actualmente MITRAMISS:

"***Dirección, elaboración y coordinación***.

Tratamiento de la información estadística contenida en la de la aplicación electrónica REGCON […], cuyos formularios estadísticos son inscritos por las comisiones negociadoras de convenios colectivos de trabajo por medios electrónicos previa autorización de la autoridad laboral competente (D. G. de Empleo y consejerías de las comunidades autónomas).

Requerimiento a las comisiones negociadoras para que inscriban en REGCON las revisiones salariales de los convenios plurianuales, pactadas, conocidas y cuantificables.

Lectura de los textos pactados para su posible incorporación a la Base de datos de Estadística, si las revisiones salariales no han sido inscritas en REGCON por las comisiones negociadoras.

Difusión mensual de los datos avance acumulados correspondientes a diciembre de 2017 y difusión de los datos definitivos del año 2016.

Las consejerías/departamentos con competencia en la materia de todas las comunidades autónomas una vez que hayan comprobado que los convenios colectivos no vulneran la legalidad vigente, dictan resolución ordenando su registro y publicación en el BOE." (Real Decreto 1043/2017, de 22 de diciembre, por el que se aprueba el Programa anual 2018 del Plan Estadístico Nacional 2017-2020.)

No hay unas directrices europeas a la hora de desarrollar estadística de CCT a nivel supranacional, la mayor de las razones es porque no todos los países de la Unión Europea tienen un tipo legal homogéneo en torno al convenio colectivo. De esta manera la comparabilidad de los datos de convenios entre países se vuelve un sinsentido. La estadística de CCT tiene una naturaleza puramente nacional y así se verá reflejado en este documento investigativo.

La minería de datos y el sector público.

Con la llegada de la sociedad de la información, los grandes volúmenes de datos y el desarrollo computacional, se hace indispensable buscar herramientas que permitan clasificar, buscar relaciones o patrones dentro de esta basta información (J.F.Seiffert, 2006). La minería de datos es una disciplina encargada de ocuparse de esta tarea.

Según Alessandro Zanasi , tres recientes avances en materias conectadas hacen posible la creciente evolución y perfeccionamiento de la minería de datos; el primero, desarrollo de nuevos y eficientes técnicas, modelos y algoritmos estadísticos que nos ayudan a tomar decisiones a partir de la información que manejemos; el segundo, desarrollo de potentes bases de datos que permitan almacenar la información sobre la que vamos a trabajar; y finalmente, el desarrollo de ordenadores y sistemas que nos permitan analizar y aplicar las técnicas a la masiva información que tenemos almacenada (A. Zanasi , 1998) . La minería de datos y la informática actualmente guardan esta estrecha relación, ya que las técnicas podían ser difícilmente aplicadas sin máquinas capaces de desarrollar cálculos en base a miles o incluso millones de casos y observaciones. Aunque no es menos cierto que varias de las técnicas que se aplican en la minería de datos existen antes de que la computación si quiera existiese en nuestras vidas, como es el caso del análisis de correspondencias (M. Greenacre, 2008) o el análisis clúster (O.F. Santana, 1991).

Actualmente, tanto el sector privado, como el sector público están incorporando mecanismos y herramientas que les permitan tratar la masiva información que poseen. Banca, aseguradoras y múltiples grandes empresas usan la minería de datos como una ventaja comparativa que usar para ser más eficientes y minimizar así las pérdidas económicas o posibles fallos en el desarrollo de sus actividades. El sector público, el cual también cuenta con grandes cantidades de información, puede estructurar ésta y convertirla en una materia prima sobre la que poder aplicar técnicas de minería de datos. Uno de los primeros usos que se dieron en el sector público a las técnicas era para preservar la seguridad de la ciudadanía por medio de los medios de inteligencia. Modelos de decisión o clasificatorios permitían identificar y hacer saltar la alarma cuando alguna célula yihadista pretendía cometer un atentado (J.W.Seiffert , 2006). Las agencias de inteligencia, las cuales, en contacto con todo tipo de empresas y cuerpos públicos, tenían

información suficiente – estructurada en una base de datos - como para desarrollar algoritmos que les permitiesen adelantarse al cometimiento de un atentado. Hay países que, en la línea de la seguridad nacional, desarrollan modelos de puntuaciones *(scores)*, para conocer la peligrosidad y el riesgo que supone la entrada de cierto individuo al país (J.W. Seiffert).

Aunque la minería de datos se ha usado esencialmente para garantizar la seguridad nacional de un país, hay cada vez más países que la están introduciendo en ámbitos como el trabajo o la salud (R. Mosquera et al., 2016). Por ejemplo, el ministerio de trabajo colombiano encargó varios trabajos para garantizar la seguridad laboral y psicosocial de los docentes de primaria y secundaria, con la intención de encontrar aquellos profesionales más vulnerables y potencialmente enfermos psicosociales. Para ello se desarrolló un modelo de decisión en el cual en base a unas variables discriminantes se podía valorar el riesgo que tenía esa persona de padecer una enfermedad psicosocial a causa de su actividad como docente en las escuelas colombianas (R. Mosquera et al, 2016).

El Trabajo de Fin de Máster no pretende analizar información que preserve la seguridad nacional, pero si aplicar técnicas de minería de datos a información perteneciente al ámbito público, como en este caso al MITRAMISS. Desde una iniciativa pública, esta información no se trata, por lo que, en actual investigación, se intentará generar ese valor añadido con la estadística e información que se obtiene de los CCT.

Base de datos, variables y paquetes estadísticos empleados para el análisis

Para conocer mejor que información se produce en el MITRAMISS, es necesario hacer una documentación de las variables, información y volumen que tiene las bases de datos que se van a usar en presente documento de investigación. Además, especificar los softwares que se han empleado y las tareas que se han realizado con ellos

Base de datos

Primero, cabe recalcar que los datos que se tratarán en este documento no son de acceso público, aunque sí lo son los agregados de estos, como ya se ha especificado en el capítulo de introducción. Para acceder a esta valiosa información que produce la Subdirección General de Estadística y Análisis Sociolaboral el autor del trabajo ha tenido que cumplimentar y acordar un contrato de cesión de datos. En él hay un compromiso explícito por el cual el individuo al que se le ceden los datos los va a usar para un objetivo puramente académico, que en todo momento se va a citar la fuente en la elaboración de tablas o gráficos y que no va a haber una manipulación de estos. Además, el investigador se compromete a preservar el secreto estadístico recogido en la ley 12/1989 de 9 de mayo que regula la Función Estadística Pública (LFEP), que obliga a los servicios estadísticos a no difundir ninguna información de una persona que sea identificada o identificable a través de los datos.

Contenido de los ficheros

El MITRAMISS proporciona un archivo Excel para cada año (2011-2018). Cada Excel contiene hojas con la siguiente información:

1. **Textos Nuevos y Nuevos Acuerdos de Empresa.**
2. **Textos Nuevos y Nuevos Acuerdos de Sector.**
3. Acuerdos de Inaplicación.
4. Promoción de Negociaciones.
5. Denuncias y Denuncias y Promociones.

6. **Tablas Salariales.**
7. Prórrogas.
8. Modificaciones.
9. Acuerdos Parciales.
10. Acuerdos de Ampliación de la Ultraactividad.
11. Acuerdos Derivados de Convenio.
12. Acuerdos de Comisión Paritaria.
13. Pronunciamiento de los Tribunales.
14. Recursos en Vía Administrativa.
15. Calendarios Laborales.
16. Comunicaciones de Oficio.

Para la investigación y aplicación de las técnicas de minería de datos únicamente se van a usar las hojas estadísticas que se han resalado en negrita (convenios de empresa, sector y sus tablas salariales), siendo las hojas con variables de mayor enjundia estadística.

Tablas Maestras

Las hojas estadísticas seleccionadas contienen esta información, de las cuales se usarán las resaltadas en negrita:

1. **Naturaleza del convenio (sector o empresa).**
2. **Autoridades Laborales.**
3. Ámbitos Funcionales.
4. Ámbitos Funcionales de los acuerdos de inaplicación.
5. Estados de los trámites.
6. **CNAE.**
7. Ámbitos Territoriales.
8. Ámbitos Personales.
9. Representación de los trabajadores.
10. Organizaciones Sindicales.
11. **Variación salarial**
12. Provincias.
13. Municipios.
14. Acuerdos o Procedimientos de Inaplicación.
15. Tipos de Trámite

16. Medidas Salariales de Inaplicación

Variables

CNAE (Código Nacional de Actividad Económica): 4 dígitos que identifican cuál es la principal actividad económica que ocupa el convenio (categórica).

Sector de Actividad: Variable creada a partir de una recodificación de la CNAE (categórica).

1 = Sector primario

2 = Actividad extractiva

3 = Industria

4 = Suministro de energía y agua

5 = Gestión de residuos y descontaminación

6 = Construcción

7 = Servicios

Autoridad laboral: órgano provincial encargado, autonómico o estatal, de activar el convenio en dicha región en la que tenga competencia (categórica).

Comunidad autónoma: A partir de la autoridad laboral se ha recodificado una variable para conocer en qué comunidad autónoma se activa el convenio colectivo del trabajo (también existe la categoría estatal, puesto que dicho convenio se aplicará en todo el territorio español). Por lo tanto, las categorías que contiene esta variable son las 17 CC.AA , las dos ciudades autónomas (Ceuta y Melilla) y el tipo estatal (20 en total).

Zona geográfica: Agrupación a partir de la comunidad autónoma donde se aplica el convenio.

1 = Norte

2 = Sur

3 = Islas

4 = Estatal

Variación salarial pactada: Variable en porcentaje de subida salarial que pacta el convenio colectivo (cuantitativa)

Total de trabajadores: Número total de trabajadores que acoge el convenio (cuantitativa).

Paquetes estadísticos

Para la realización de presente análisis estadístico se han usado dos paquetes estadísticos SAS y SPSS.

SAS

Con el paquete SAS se ha trabajado principalmente para unificar todos los ficheros Excel en una única base de datos apta para el tratamiento. Se han realizado tareas como:

- Unificar todos los Excel (uno por año) en un único fichero Excel.
- Ordenar la variable por la "ID" para poder ejecutar la sentencia *merge.*
- Coger las hojas de (Sector, empresa, acuerdos salariales) y unificarlas mediante la "ID" con la sentencia *merge.*
- Eliminar aquellas variables que no son de interés para la minería de datos.
- Eliminar registros perdidos.

En el anexo a presente documento de investigación se adjuntará el código usado para ejecutar dichas tareas.

SPSS

Statistical Programme of Social Science (SPSS) ha sido utilizado principalmente por el dinamismo del programa, lo que permite repetir varios análisis de una forma rápida, sin necesidad de modificaciones de código. Ha sido el programa ocupado de aplicar las técnicas, previa transformación de alguna de las variables de entrada para las técnicas. De forma más concreta, las tareas realizadas por el paquete estadístico son:

- Creación de la variable **CC.AA** a partir de la transformación de la variable **autoridad laboral**
- Creación de la variable **Zona geográfica** a partir de la transformación de la variable **CC.AA**
- Creación de la variable **Sector de actividad** a partir de la variable **CNAE**

- Tipificación de las variables cuantitativas usadas (**variación salarial pactada y número total de trabajadores**) en el análisis de conglomerados bietápico.
- Transformaciones logarítmicas para intentar la normalización de las variables cuantitativas anteriores
- Aplicación del análisis de correspondencias
- Aplicación del análisis de conglomerados bietápico
- Caracterización de la agrupación del análisis de conglomerados mediante estadística descriptiva.

Las técnicas de minería de datos

Para concretar qué herramientas estadísticas y técnicas de análisis tiene la minería de datos voy a pasar a documentar aquellas que se van a emplear en el tratamiento de los micro-datos proporcionados por el MITRAMISS.

Análisis de correspondencias

El análisis de correspondencias – a partir de aquí AC – es una técnica estadística que nos permite trabajar con variables categóricas e intentar buscar asociaciones gráficamente y visualmente identificables (M.Greenacre , 2008). Nos facilita una fácil comprensión de los datos y poder barajar la agrupación de dos o más categorías que pudieran tener una alta asociación. El análisis de correspondencias, aunque es una técnica en auge, tiene ya varias décadas de existencia, donde la computación todavía dejaba mucho que desear. J.P Benzécri, un estadístico francés, le quiso dar la importancia que se merecía a ésta humilde técnica, debido a que se podían representar visualmente variables categóricas, lo que hacía más fácil la interpretación, sobre todo a gente no experta en la materia (Benzècri, 1977).

La idea del AC era que, mediante la distancia χ^2, se pudiese graficar en un plano bidimensional las categorías para poder ver el nivel de asociación o diferencia entre ellas. Esta técnica está vinculada a la reducción de dimensiones por su utilidad para poder pasar información presentada en *n* dimensiones a un plano de dos ejes (Benzècri, 1977). El número de categorías que pueden usarse siempre ha sido un objeto sometido a debate, siendo lo más sensato que el técnico sea el que establezca una cantidad determinada para que la nube de puntos sea lo suficientemente interpretable. Por ejemplo, si manejamos una variable categórica que sean las provincias de España, puede ser útil para la visualización convertir ésta en una agrupación por comunidades autónomas. En un primer momento tendríamos una variable con 50 categorías, que pasadas al plano se hace más complicado interpretar las posibles asociaciones. Sin embargo, con la comunidad autónoma tendríamos solo una nube de 17 puntos que nos facilitaría el análisis.

La técnica parte de una prueba χ^2 para contrastar la independencia de dos variables categóricas, con la diferencia de que en el AC se usarán las frecuencias relativas para poder interpretar estas en un plano de n dimensiones (siendo las dimensiones el número

mínimo de categorías que tengan las variables – 1). Lo primero de todo es construir una tabla de distancias χ^2, que lo haremos con la siguiente fórmula.

$$X^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Como la pretensión del AC no es contrastar únicamente la independencia de las variables categóricas, construimos una distancia χ^2 transformada para poder representar así tanto los perfiles fila, como los perfiles columna.

$$X^2 = \sum \frac{\left(\frac{f_0}{Total_{fil/col}} - \frac{f_e}{Total_{fil/col}}\right)^2}{\frac{f_e}{\left(Total_{fil/col}\right)^2}} = \sum \left(Total_{fil/col}\right) \cdot \frac{\left(\frac{f_0}{Total_{\frac{fil}{col}}} - \frac{f_e}{Total_{\frac{fil}{col}}}\right)^2}{\frac{f_e}{\left(Total_{\frac{fil}{col}}\right)}} = X^2$$

Ahora tenemos perfectamente representados tanto los perfiles fila, como los perfiles columna. Lo único que falta aplicarle a la ecuación es tener en cuenta el tamaño total de la muestra, para así tener el producto de los perfiles fila/columna con la masa, quedando de la siguiente manera:

$$X^2 = \sum \underbrace{\frac{\left(Total_{\frac{fil}{col}}\right)}{n}}_{\text{Masa del perfil}} \cdot \underbrace{\frac{\left(\frac{f_0}{Total_{\frac{fil}{col}}} - \frac{f_e}{Total_{\frac{fil}{col}}}\right)^2}{\frac{f_e}{\left(Total_{\frac{fil}{col}}\right)}}}_{\text{Dist. perfil fil/col}}$$

De este modo tenemos por un lado la masa de esa fila/columna y por el otro el valor de las distancias de los perfiles fila/columna, lo que nos resultará la distancia de los perfiles fila/columnas ponderados por la masa que tenga determinada fila/columna.

Finalmente tenemos otro valor muy importante en el AC, que es la **inercia**, la cual la podemos representar como χ^2/n. Por lo que no es más que un valor de χ^2 relativizado por la frecuencia total. Podemos sacar así un valor de inercia para cada perfil de fila y columna con la intención de obtener un valor que nos cuantifique cuánto se desvía ese perfil fila/ columna del perfil fila/columna esperado.

Una vez tenemos las distancias y las inercias podemos representar nuestros perfiles fila en un gráfico. El problema que surge en esta técnica es que la mayoría de ocasiones las dimensiones son mayores a 3 lo que se hace imposible la representación gráfica. Lo que se debe hacer es una reducción de *n* dimensiones a 2 dimensiones con la intención de perder la menor parte de la información que contienen las variables analizadas.

Gráfico 2.

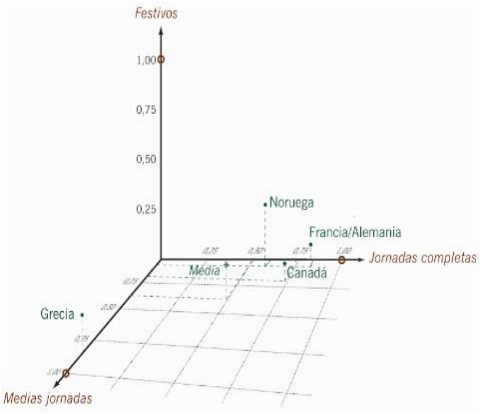


Gráfico 3.

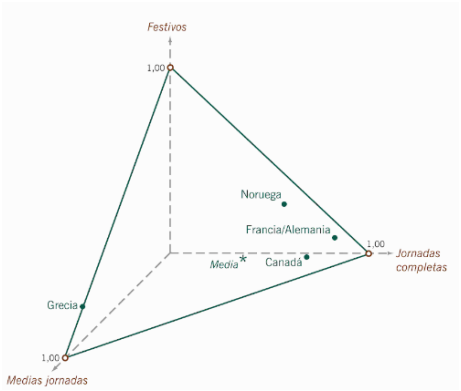
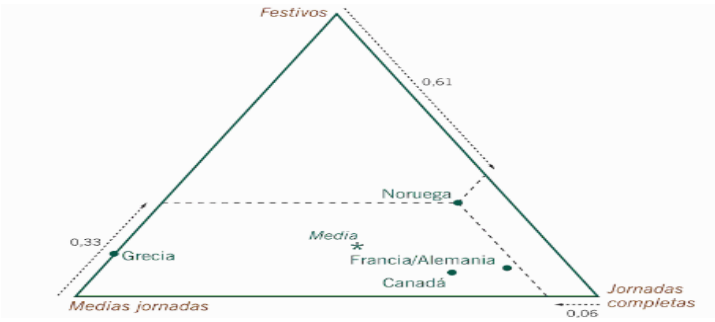


Gráfico 4.



Fuente: Greenacre, M. (2008). La práctica del análisis de correspondencias. Bilbao: Fundación BBVA.

Para llegar a un espacio bidimensional desde uno de *n* dimensiones es esencial usar la computación para que haga esta ardua tarea por nosotros, lo que facilita el trabajo a los técnicos estadísticos. Una vez representados los perfiles en un espacio bidimensional la interpretabilidad de los datos será mucho más fácil para cualquier persona que se preste a analizarlos, aunque no sea experta en el análisis estadístico.

Análisis clúster

El análisis clúster es una técnica de agrupación de observaciones en base a *n* variables, usando la distancia entre observaciones para valorar la formación de determinados grupos. La finalidad del análisis es conseguir grupos cuyas observaciones sean muy homogéneas dentro de él y muy heterogéneas con respecto a otros grupos.

Hay varios métodos de análisis clúster, presente Trabajo de Fin de Máster va a usar el análisis de conglomerados bietápico, por ser la técnica más apropiada para analizar bases de datos con gran número de casos. La agrupación de dos etapas nace con la idea de algoritmos como BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), los cuales funcionan en las siguientes dos etapas:

- **En el paso de pre-agrupación,** SPSS escanea la base de datos con la intencionalidad de crear un árbol parecido al dendograma (cluster feature tree) que agrupa los casos en clústeres y sub-clústeres.
- En el segundo paso se usa **un método jerárquico aglomerativo** (más útil para procesos automáticos de determinación de los grupos). Éste funciona utilizando como entrada el *cluster feature tree* creado en el paso previo. Parte de que hay tantos clúster o nodos, como hayan sido calculados en el paso número uno. A partir de ahí va aglomerando clúster o sub-clúster atendiendo a la medida de distancia seleccionada en el análisis (Zhang, Ramakrishnan, & Livny, 1996).

Si escogemos un criterio automático de selección de conglomerados, el programa se apoya en métricas bayesianas de validación de modelos como pueda ser BIC (se detallará posteriormente) o AIC (C. Fraley ; A.E . Raftery,1998)

Alguna de las características del clúster bietápico son:

- Se pueden incluir tanto variables cuantitativas como variables categóricas (en presente trabajo solamente se incluirán cuantitativas, compatibles con el algoritmo BIRCH.

- Se necesita independencia de las variables a usar.
- Es necesario que las variables cuantitativas cumplan normalidad.
- Útil para grandes bases de datos.
- Determina el número óptimo de conglomerados.

La distancia log-verosimilitud, que permite funcionar tanto con variables categóricas como cuantitativas, realiza supuestos de normalidad (cuantitativas), multinomialidad (cualitativas) e independencia. La técnica puede funcionar aun no cumpliendo los supuestos debido a la gran robustez de ésta.

$$D_{euc} = \sqrt{(x_{ij} - x_{ik})^2}$$

Al igual que podemos usar la distancia log-verosimilitud para cuantificar la homogeneidad entre dos clústeres, podemos usar la distancia euclídea usando el centroide de cada clúster para cuantificar la distancia entre ellos.

Para determinar el número de conglomerados automáticamente, como ya se ha mencionado, se usa el criterio de información Bayesiana (BIC) de Swartz, el cual es útil para la selección de un modelo entre un número de modelos finito.

$$BIC = k \cdot \ln(n) - 2 \cdot \ln(L)$$

k= número de parámetros

L = Es el valor máximo de verosimilitud del clustering.

Como vemos el BIC penaliza el uso de muchos parámetros, incorporando así el principio de parsimonia en el modelo, por el cual menos parámetros y más simplismo es positivo en el criterio BIC para la selección del número de clústeres.

Para determinar la calidad de la agrupación, en el análisis clúster suele ir acompañado de medidas de cohesión dentro de los clúster y separación entre ellos. L. Kaufman y P.J. Rosseeuw encuentran una métrica ideal para valorar la calidad de agrupación, representada de la siguiente manera:

$$VCS = \frac{\sum \left(\frac{(D_a - D_b)}{max(D_a, D_b)} \right)}{n}$$

Da = Distancia entre un caso y el centroide del clúster no propio más cercano

Db = Distancia entre un caso y el centroide del clúster propio

Tendremos así un valor comprendido entre [1,-1] que nos ayude a cuantificar la calidad de determinado *clustering*. El valor 1 es que todos los casos están ubicados en el centroide del clúster propio, siendo -1 el valor que representa que todos los casos están ubicados en el centroide de un clúster no propio *(*L. Kaufman y P.J. Rosseeuw,1990).

Análisis: aplicación de las técnicas de minería de datos

Una vez contextualizado el Trabajo de Fin de Máster, se van a aplicar las dos técnicas de minería a la información perteneciente al MITRAMISS. Es lo que realmente nos permitirá generar ese valor añadido a la información tan valiosa que tiene bajo dominio la administración sobre el mercado laboral y en concreto los CCT. Correspondientes análisis pueden ser útiles para tener una mejor composición de cómo funciona el mercado laboral en España y más en concreto la negociación colectiva. Permitiendo al ejecutivo y a los distintos cuerpos políticos del estado tomar decisiones fundamentadas en análisis provenientes de la minería de datos en el ámbito público, práctica estadística que se quiere fomentar con este documento investigativo.

Análisis de correspondencias con la información de CCT.

La base de datos proporcionada por el MITRAMISS contiene diversas variables categóricas que pueden ser de amplia utilidad para su tratamiento. El AC nos permite graficar esta información en un espacio bidimensional que ayuda a interpretar los resultados provenientes de la técnica. Pretendiendo de esta manera buscar asociaciones entre categorías de una manera simple y accesible para públicos no expertos, lo que permite divulgar los resultados a otros técnicos del MITRAMISS que no sea estadísticos.

Las variables que se van a usar en presente análisis van a ser tres, como ya se ha especificado en el capítulo 7 (comunidad autónoma donde se activa el convenio, zona geográfica y sector de actividad que ocupan los trabajadores a los que acoge el convenio). De esta manera podremos valorar la dependencia que hay entre la zona de actuación del convenio y el sector de actividad que ocupa.

Tabla 1.

Tabla de frecuencias

Sector de actividad	Zona geográfica				
	NORTE	SUR	ISLAS	ESTATAL	Margen activo
Sector Primario	107	116	8	14	245
Actividad extractiva	46	36	3	8	93
Industria	2315	1044	84	243	3686
Suministro de Energía y agua	130	159	20	41	350
Gestion de residuos y descontaminación	454	376	63	11	904
Construcción	386	367	21	58	832
Servicios	5351	4028	620	1486	11485
Margen activo	8789	6126	819	1861	17595

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

La tabla de frecuencias nos da una información preliminar útil para tener una idea de cómo están distribuidos los perfiles fila, columna y sus totales. A recalcar que tanto la industria como los servicios ocupan la mayor parte de CCT que se firman en España, con una gran diferencia con otros sectores de actividad, por lo que se presume serán categorías de importancia en el análisis de correspondencia por la masa que contienen. Relativo a los perfiles columna, lo único destacable es que el norte firma más convenios colectivos que el sur y que la categoría estatal no tiene tanto importancia como su carácter regional (Norte, sur o islas).

Tabla 2.

Resumen

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,145	,021			,676	,676	,007	-,037
2	,091	,008			,267	,942	,007	
3	,042	,002			,058	1,000		
Total		,031	549,995	,000[a]	1,000	1,000		

a. 18 grados de liberta

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS

El espacio inicial es de 3 dimensiones (n-1), teniendo una inercia total de 0,31 y un valor chi-cuadrado de 549,995, que atendiendo a su p- valor asociado nos permite rechazar independencia entre las dos variables. Con estas primeras evidencias que tenemos a partir de los resultados podemos adelantar que la zona geográfica en la que se activa convenio tiene cierta asociación con la actividad económica que ocupa dicho convenio.

Tabla 3. **.Perfiles de fila generales**

SECTOR DE ACTIVIDAD	Masa	Puntuación en dimensión		Inercia	Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
		1	2		1	2	1	2	Total
Sector Primario	,014	-,016	,853	,001	,000	,111	,000	,836	,836
Actividad extractiva	,005	,097	,259	,000	,000	,004	,110	,489	,599
Industria	,209	,692	-,169	,015	,691	,065	,964	,036	1,000
Suministro de Energía y agua	,020	-,548	,493	,001	,041	,053	,635	,322	,957
Gestion de residuos y descontaminación	,051	,267	,954	,006	,025	,512	,097	,773	,869
Construcción	,047	,064	,596	,002	,001	,184	,012	,656	,668
Servicios	,653	-,232	-,099	,006	,241	,071	,892	,103	,995
Total activo	1,000			,031	1,000	1,000			

a. Normalización simétrica

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

Si atendemos a los perfiles de fila, podemos ver la inercia que aporta cada categoría, destacando que la industria aporta casi la mitad de la inercia total, lo que la convierte en un perfil fila claramente diferenciado del resto. Las contribuciones totales de la mayoría de perfiles son altas, por lo que la reducción de dimensiones no ha supuesto una gran pérdida de información, excepto con la actividad extractiva y la construcción, que muestran contribuciones totales más bajas.

Tabla 4.

Perfiles de columna generales

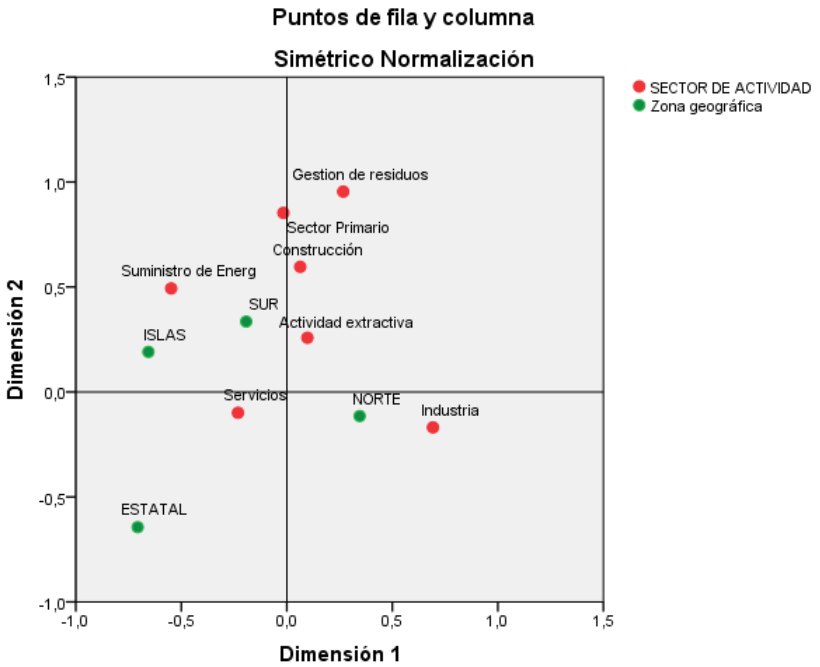
Zona geográfica	Masa	Puntuación en dimensión 1	Puntuación en dimensión 2	Inercia	Contribución: Del punto en la inercia de dimensión 1	Contribución: Del punto en la inercia de dimensión 2	Contribución: De la dimensión en la inercia del punto 1	Contribución: De la dimensión en la inercia del punto 2	Contribución: De la dimensión en la inercia del punto Total
NORTE	,500	,345	-,115	,009	,410	,072	,931	,065	,996
SUR	,348	-,193	,335	,006	,089	,428	,331	,627	,958
ISLAS	,047	-,656	,191	,004	,138	,019	,647	,034	,681
ESTATAL	,106	-,707	-,644	,012	,364	,481	,652	,340	,992
Total activo	1,000			,031	1,000	1,000			

a. Normalización simétrica.

Fuente: Elaboración propia a partir de la información de la Subidrección General de Estadística y Análisis Sociolaboral del MITRAMISS.

Con los perfiles columna cabe subrayar que una gran parte de la inercia la aporta el convenio estatal, lo que le diferencia claramente del resto de perfiles. Cabe recalcar que esta categoría, por su naturaleza, es claramente diferente al resto y, estadísticamente, esto tampoco ha pasado desapercibido para el AC. Todos los perfiles fila en su reducción a dos dimensiones siguen teniendo una contribución alta, excepto las islas, lo que no supone una gran pérdida de información.

Gráfico 5.



Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

Teniendo los puntos ya representados en un espacio bidimensional podemos, de una forma más visual, encontrar a simple vista las asociaciones y diferencias entre categorías. Sin olvidar que esta técnica no lleva un contraste de hipótesis asociado para ver si hay una asociación significativa estadísticamente ente las categorías. Será el técnico el que realice conclusiones propias a partir de la representación gráfica de las categorías.

La relación, a priori, más llamativa es la del norte con la industria, lo que manifiesta que la zona septentrional de España tiene una negociación colectiva más asentada con sectores relacionados a esta actividad. Mientras que el resto de actividades están asociadas, aunque no fuertemente, con los convenios que tienen una posición geográfica meridional e insular. Podemos confirmar una vez graficados los perfiles que la categoría estatal está claramente diferenciada del resto de categorías y no se le puede encontrar asociación ninguna, confirmando la particular naturaleza de ésta.

Aunque los resultados que se han obtenido con una variable geográfica altamente agregada son bastante reveladores (en parte, por su simpleza visual), para tener una mejor composición de las asociaciones entre la zona geográfica y el sector de actividad, se desagregará la variable geográfica por las 17 comunidades autónomas (añadidas Ceuta, Melilla y la categoría estatal). De esta forma tendremos resultados más detallados, aunque visualmente más complejos, dificultando la interpretabilidad de los resultados.

Tabla 5. **Resumen**

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,207	,043			,534	,534	,008	,022
2	,123	,015			,189	,723	,007	
3	,098	,010			,119	,842		
4	,089	,008			,099	,941		
5	,059	,003			,043	,983		
6	,036	,001			,017	1,000		

Total		,080	1415,994	,000[a]	1,000	1,000		

a. 114 grados de libertad

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

El p-valor asociado al contraste de independencia nos permite rechazar la hipótesis nula del contraste de independencia. Por lo que hay evidencia estadística para poder confirmar una dependencia entre la comunidad autónoma donde se firma el convenio y el tipo de actividad que ocupa. Las dos primeras dimensiones acumulan el 72,3% de la inercia, por lo que en su reducción a dos dimensiones acarreará una pérdida de información más importante que en el caso anterior de AC.

Tabla 6. **Perfiles de fila generales**

SECTOR DE ACTIVIDAD	Masa	Puntuación en dimensión		Inercia	Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
		1	2		1	2	1	2	Total
Sector Primario	,014	,120	-,488	,007	,001	,027	,006	,056	,062
Actividad extractiva	,005	-,205	-1,130	,003	,001	,055	,014	,246	,260
Industria	,209	,873	-,063	,033	,769	,007	,995	,003	,998
Suministro de Energía y agua	,020	-,474	-,012	,005	,022	,000	,175	,000	,175
Gestion de residuos y descontaminación	,051	-,340	-,919	,012	,029	,352	,106	,458	,564
Construcción	,047	-,426	-,995	,011	,041	,379	,165	,534	,699
Servicios	,653	-,209	,184	,009	,137	,180	,672	,312	,984
Total activo	1,000			,080	1,000	1,000			

a. Normalización simétrica

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

De la misma manera que con el primer AC, la industria sigue suponiendo el perfil fila que más aporta a la inercia total, manteniéndolo en esa posición de heterogeneidad con el resto de categorías de la variable. En esta ocasión, la reducción de dimensiones ha supuesto una considerable pérdida de información en la nueva conformación de los ejes, sobre todo para la categoría sector primario, actividad extractiva y suministro de energía y agua.

Tabla 6. **Puntos de columna generales[a]**

Comunidades Autónomas	Masa	Puntuación en dimensión		Inercia	Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
		1	2		1	2	1	2	Total
MAD	,074	-,256	,144	,003	,023	,012	,354	,066	,420
CYL	,081	,470	-,192	,004	,087	,024	,876	,087	,963
CLM	,041	,041	,252	,001	,000	,021	,011	,250	,261
CAT	,083	-,113	-,113	,002	,005	,009	,093	,055	,148
EUS	,062	1,190	,195	,019	,427	,019	,964	,015	,980
AND	,189	-,385	-,259	,008	,136	,103	,737	,198	,935
EXT	,028	-,574	-,878	,007	,045	,178	,273	,379	,652
CANT	,034	,263	-,194	,001	,011	,010	,388	,125	,513
AST	,035	,359	-,205	,003	,022	,012	,297	,058	,356
RIOJA	,008	,002	-,139	,001	,000	,001	,000	,016	,016
ARA	,024	,253	,160	,001	,007	,005	,468	,111	,579
BAL	,011	-,551	,399	,001	,016	,014	,558	,174	,732
VAL	,088	,158	-,105	,003	,011	,008	,161	,043	,203
MUR	,011	,358	,207	,003	,007	,004	,093	,018	,111
NAV	,024	,599	,394	,002	,042	,030	,733	,188	,922
CANAR	,036	-,567	,173	,003	,055	,009	,700	,039	,739
GAL	,061	,219	-,282	,001	,014	,040	,499	,492	,990
CEU	,001	-,968	,548	,001	,006	,003	,519	,099	,617
MEL	,002	-1,202	-,041	,003	,013	,000	,183	,000	,183
ESTATAL	,106	-,379	,761	,012	,073	,497	,267	,642	,909
Total activo	1,000			,080	1,000	1,000			

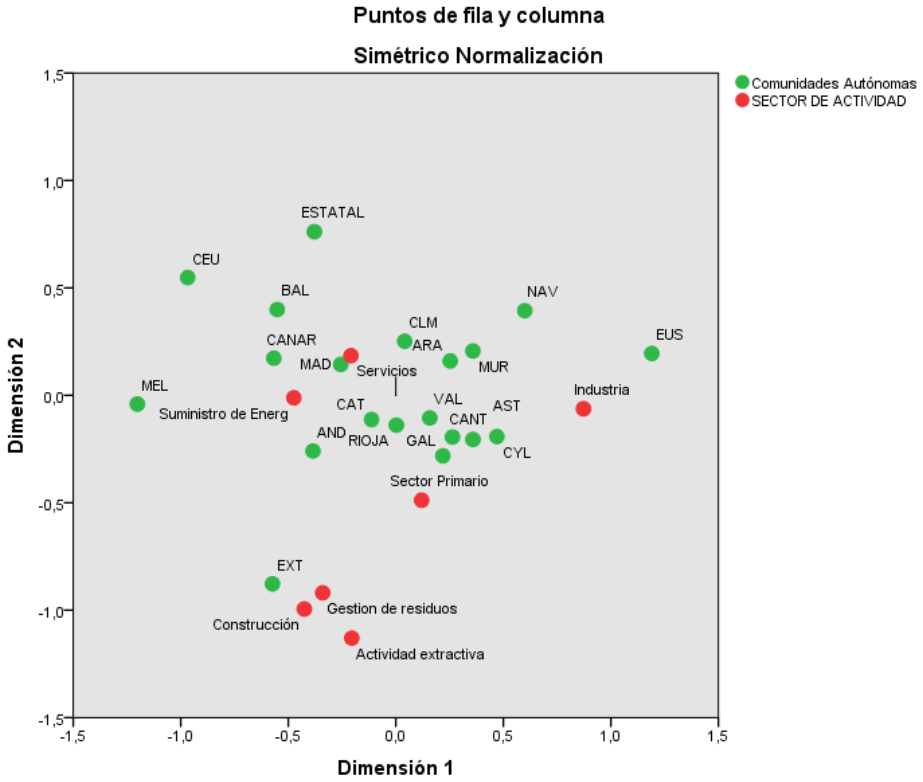
a. Normalización simétrica

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

El perfil columna que más aporta a la inercia total es País Vasco, seguido de Andalucía y Extremadura. Las contribuciones de los perfiles fila a la configuración de los ejes son altas en la mayor parte de las categorías, exceptuando las bajas contribuciones de La rioja, Murcia, Melilla, Cataluña y Asturias.

Se confirma la gran diferenciación de Extremadura (gráfico 6), la cual tiene una alta asociación con la construcción y la gestión de residuo, poniendo de manifiesto la buena situación de negociación que hay en estos sectores de actividad. Aunque no podemos asociar directamente la región con estos sectores de actividad, si podemos aseverar que existe una cultura asentada de negociación dentro de ellos. Otra clara asociación entre perfiles fila y perfiles columna es Madrid y los servicios. Se visualiza en el AC la predominancia de este sector en la Comunidad Autónoma, lo que se refleja posteriormente en la negociación colectiva de la región.

Gráfico 6.



Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

En el primer AC hecho con la variable de región agregada se visualizó una gran asociación entre el norte y la industria, siguiéndose manteniendo en este segundo análisis. Con la diferencia de poder apreciarse cuales son las CC.AA del norte que más tradición de negociación tienen para con la industria. Euskadi mayoritariamente y, en menor medida, Navarra y Castilla y León las regiones con mayor asociación a la negociación colectiva dentro de la industria.

El AC también permite identificar homogeneidades entre regiones a la hora de firmar convenios colectivos de un determinado sector de actividad. Destacable es que el AC reconozca agrupaciones geográficas naturales por su homogeneidad en la negociación colectiva, como se aprecia en el gráfico 6.

- Agrupación insular (Canarias y Baleares) por su proximidad en el plano
- Agrupación de las ciudades autónomas Ceuta y Melilla
- Cierta homogeneidad y posible agrupación entre Cantabria, Asturias, Galicia, Comunidad Valenciana y Castilla y León.

- Clara diferenciación del tipo de convenio colectivo de aplicación estatal, lo que tiene sentido por su diferente naturaleza como categoría.

Curioso, cuanto menos, es como el análisis reconoce estas agrupaciones geográficas también en la negociación colectiva. Por lo que se puede aseverar que hay patrones geográficos y económicos que hacen que algunas categorías se agrupen en el plano. No es casual que las islas, por su naturaleza económica insular, aparezcan con gran proximidad en el gráfico. Al igual que aquellas CC.AA de la zona septentrional de España (Cantabria , Galicia y Asturias) , las cuales comparten claros rasgos económicos e históricos que se ven traducidos en la cierta homogeneidad que hay en su negociación colectiva por sectores de actividad.

Gracias al AC, con un sencillo repaso visual, se han podido identificar asociaciones, diferenciaciones y posibles agrupaciones de categorías. Permitiendo de esta manera a los técnicos de la administración y el MITRAMISS tener una excelente composición de la realidad de la negociación colectiva por sectores en el estado español. Lo importante es que los resultados pueden ser fácilmente interpretados por individuos del MITRAMISS que no tengan una familiarización con la estadística, por lo que la divulgación de la información es bastante cómoda. Luego, con un mejor conocimiento del mapa de la negociación, los actores políticos tienen mejores herramientas al alcance para tomar decisiones sobre la negociación colectiva en España. La minería de datos y en este caso el AC pueden ayudar a la toma de decisiones y mejor asignación de recursos públicos, generando ese valor añadido que se pretende promocionar en presente trabajo de investigación.

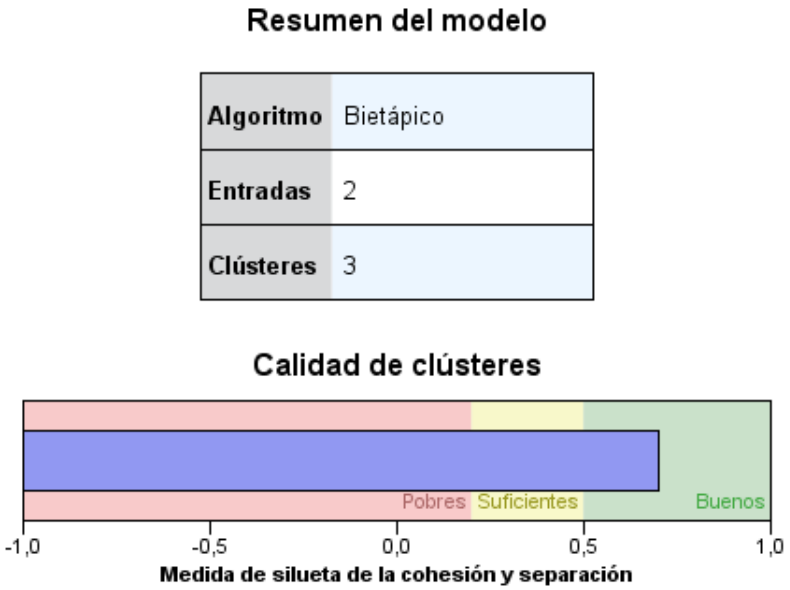
Análisis de conglomerados con la información de CCT.

El AC ha demostrado ser una excelente herramienta para tratar y graficar variables categóricas, tanto, como el análisis de conglomerados puede ser para tratar variables cuantitativas (aunque no únicamente). El AC también ha demostrado ser una útil técnica para valorar la agrupación de categorías, de la misma manera que el análisis de conglomerados será ampliamente útil para agrupar observaciones de una base de datos. Por ello, en actual subcapítulo se empleará dicha técnica para tratar datos relativos a los CCT.

El análisis de conglomerados usará la variación salarial pactada en el CCT (en porcentaje) y el número de trabajadores para realizar las correspondientes agrupaciones halladas.

Ambas son de tipología cuantitativa, aunque, como hemos visto en capítulo 6, el clúster bietápico permita utilizar variables cualitativas. Se usará la distancia euclídea entre clústeres y para que no haya perturbaciones de escala se han tipificado las variables de entrada.

Gráfico 7.



Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

El análisis nos propone con la selección automática una agrupación en tres conglomerados, atendiendo a las dos variables de entrada seleccionadas. La métrica de L. Kaufman y P.J. Rosseeuw sobre la cohesión y separación de los clústeres permite confirmar que la calidad del clúster es lo suficientemente adecuada como para darle el visto bueno a la agrupación realizada por la técnica de conglomeración bietápica. Las observaciones están lo suficientemente cerca del centroide del clúster propio y lo suficientemente lejos del centroide del clúster ajeno como para confirmar la calidad de los resultados obtenidos.

Tabla 7.

Agrupación en clúster automática

Número de clústeres	Criterio bayesiano de Schwarz (BIC)	Cambio BIC[a]	Razón de cambios BIC[b]	Razón de medidas de distancia[c]
1	7867,508			
2	4572,481	-3295,027	1,000	1,774
3	2730,564	-1841,917	,559	3,674
4	2254,374	-476,190	,145	1,236

5	1875,546	-378,828	,115	1,296
6	1591,252	-284,293	,086	1,252
7	1371,080	-220,172	,067	1,552

a. Los cambios son del número anterior de clústeres de la tabla.

b. Las razones de los cambios son relativas al cambio para la solución de dos clústeres.

c. Las razones de medidas de distancia se basan en el número actual de clústeres respecto al número anterior de clústeres.

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

El criterio de información bayesiano de Shwarz , medida de referencia para la selección automática de conglomerados, vemos como va disminuyendo a medida que el número de conglomerados aumenta. Si el único criterio que tuviésemos es el BIC, nos quedaríamos con el mayor número de conglomerados, dado que tiene un mejor valor BIC. Pero el criterio de selección automática no toma únicamente el valor en absoluto, sino el cambio relativo que hay al añadir una agrupación más. El cambio del valor BIC de tres a cuatro conglomerados no es lo suficientemente grande como para que el criterio de selección valore positivamente la creación de un conglomerado más. Recordemos que este criterio se base también en criterios de parsimonia, por lo que, a más parámetros, mayor valor BIC resultará y lo que se trata es de encontrar valores bajos.

Planteado de otra manera, el criterio de selección automática se detiene en tres conglomerados porque la razón de distancia (aumento relativo de la distancia entre clústeres) tiene el valor más alto cuando se pasan de dos a tres conglomerados y un valor considerablemente más bajo cuando se pasan de tres a cuatro conglomerados. Resultando de esta manera un modelo de agrupación más simple (cumpliendo principios de parsimonia) y con conglomerados lo suficientemente cohesionados (intra) y diferenciados (entre).

Tabla 8.

Descriptivos

		N			% de combinado	Media total_trabajadores_afectados	Media var_sal_pac
Clúster	1	220			3,9%	22665,33	,906591
	2	1305			23,1%	691,82	2,764552
	3	4126			73,0%	440,77	,560201

Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

Después de realizar las agrupaciones se quiere, de alguna manera, caracterizar los grupos y ver si hay algún patrón detrás de las agrupaciones. La idea es encontrar algunas similitudes en las observaciones en cada grupo basándonos en una variable externa o incluso desconocida.

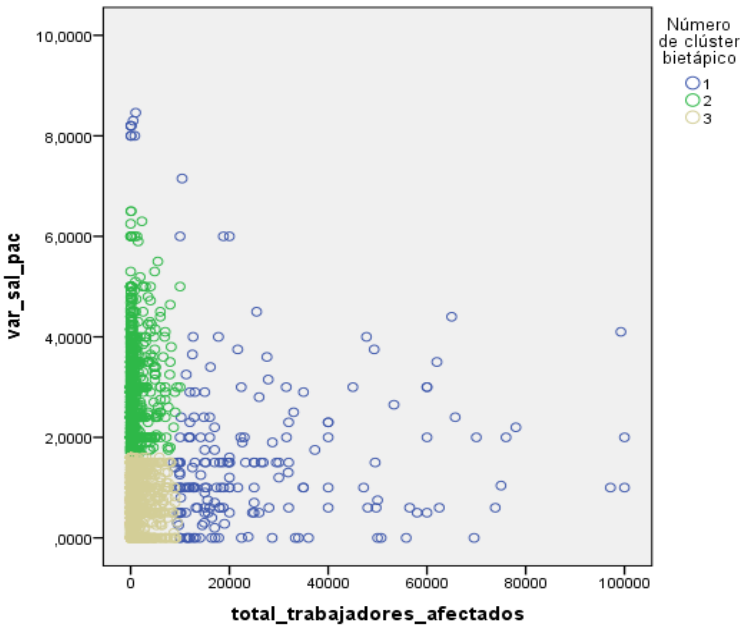
Es preciso comenzar la caracterización de los grupos con algunos estadísticos descriptivos:

Clúster 1. Es el menos numeroso, únicamente un 3,9% de los casos totales se hallan aquí. Es un grupo de CCT que acoge a un número trabajadores bastante alto y que tienen variaciones salariales moderadas.

Clúster 2. Casi un cuarto de las observaciones se halla dentro de este conglomerado de CCT, que se caracteriza por no acoger a un gran número de trabajadores, pero tener altas variaciones salariales pactadas.

Clúster 3. Como muestra la tabla, en esta agrupación se encuentran el mayor número de CCT (73%). Son convenios que acogen a muy pocos trabajadores y en los que se pactan variaciones salariales bajas, inclusive negativas.

Gráfico 8.



Fuente: Elaboración propia a partir de información de la Subdirección General de Estadística y Análisis Sociolaboral del MITRAMISS.

El grupo 1 experimenta gran dispersión, de hecho, es el grupo con casos más heterogéneos a pesar de ser el conglomerado menos numeroso. Lo interesante de la agrupación es que principalmente agrupa a convenios que tienen un número de trabajadores altos, sin importarle demasiado la variación salarial pactada en dicho convenio. Casualmente, o mejor dicho, causalmente, este grupo está compuesto mayoritariamente por convenios de sector, en contraposición a la otra tipología, convenios de empresa. Estos convenios se caracterizan principalmente por abarcar un sector de actividad al completo (ej.: construcción, hostelería, juego, TIC, telemarketing, etc.), por lo que, en general, abarcarán a un gran número de trabajadores. Sin embargo, el grupo 2 y 3 está compuesto mayoritariamente por convenios de empresa y convenios de sectores minoritarios en la economía española, los cuales abarcan considerablemente menos trabajadores. La única diferencia que hay entre el número 2 y 3 es que el primero contiene subidas salariales considerablemente altas, comúnmente por encima del 2% (y por encima del IPC), mientras que el otro tiene subidas bastante bajas, inclusive negativas (por debajo del IPC).

El grupo 3, en contraposición al 1, es el grupo más homogéneo y donde se encuentran la mayor parte de CCT. Las observaciones, tanto para la variable variación salarial, como para el número de trabajadores se agrupan mayoritariamente en torno al 0, siendo comprensible que este grupo, que tiene los valores medios más bajos para ambas variables, tenga el mayor número de observaciones. Suponiendo que la mayor parte de convenios que se firma en el estado español son de empresa o sectores minoritarios y con una variación salarial pactada bastante moderada.

Finalmente, el grupo 2, con bastante homogeneidad también se caracteriza principalmente porque, aunque la mayor parte de los convenios son de empresa o sectores minoritarios, tienen variaciones salariales pactadas bastante altas. Atendiendo al nombre de cada CCT, en este grupo podemos identificar una gran mayoría de convenios relacionados con sectores en auge (juegos de azar), estratégicos (alimentación, industria) o subcontratas públicas (servicio de ambulancias, de recogida de residuos). Al ser un grupo que acogen a pocos trabajadores, mayoritariamente son convenios de empresa y convenios de sectores minoritarios, como ocurre en el grupo 3.

Por un lado, que el convenio sea de sector o de empresa supone un factor importante para que las observaciones se agrupen:

Convenios de sector – grupo 1

Convenios de empresa – grupo 2 y 3

Luego, el IPC parece suponer también un factor diferenciador del grupo 2 y 3. Lo que cobra sentido dado que las partes negociadoras a la hora de pactar una variación salarial siempre tienen muy en cuenta el valor del IPC del año anterior al que pactan la subida. Si la empresa aplica una variación salarial por encima del IPC, quiere decirse que sus trabajadores no van a perder poder adquisitivo, mientras que, si la empresa pacta una variación salarial por debajo del IPC, sus trabajadores verán afectado su poder adquisitivo negativamente. El grupo 2 tiene casos mayoritariamente en el que los trabajadores que acoge verán afectado positivamente su poder adquisitivo, mientras que, en el 3, sin embargo, si verán afectado su poder adquisitivo negativamente.

Conclusiones

La minería de datos en el ámbito público es una disciplina que está claramente en auge. La idea de este trabajo es poner en valor la importancia que tienen las técnicas para tratar información proveniente de la administración. Por ello, se han aplicado técnicas de minería de datos a la información proveniente del MITRAMISS con la intención de demostrar la gran utilidad que tiene aplicar técnicas para tener una mejor composición de la realidad, en este caso de la negociación colectiva. Agilizar los procesos, promover la eficiencia, desentramar asociaciones, agrupaciones, etc. Son algunas de las múltiples ventajas que ofrece la minería de datos en el tratamiento de datos de dominio público.

Hubiese sido mucho más revelador aplicar varias técnicas de minería de datos, pero por limitaciones de espacio solo ha sido posible aplicar dos técnicas, de las cuales se han sacado resultados bastantes reveladores en ambas.

Con respecto al AC, cabe recalcar:

- La clara asociación del norte con la firma de convenios colectivos relacionados con la industria, lo que demuestra que en este sector de actividad hay una cultura asentada con la negociación colectiva. Al desagregar por CC.AA se sacan las mismas conclusiones con respecto al norte, siendo el País Vasco la Comunidad Autónoma más asociada a la negociación colectiva de industria.

- El AC ha identificado la categoría *Estatal* como una claramente diferenciada del resto, lo que tiene sentido puesto que es la única categoría que tiene un ámbito de aplicación en todo el territorio español, mientras que el resto solo tienen aplicación en el ámbito autonómico. Esta clara diferenciación ha sido identificada por el AC, lo que me parece recalcable en presente capítulo de conclusiones.

- Otro resultado a subrayar del AC son las agrupaciones y diferenciaciones que se han encontrado. Clara homogeneidad y asociación entre CC.AA como Asturias, Cantabria, Galicia , Castilla y León y Comunidad Valenciana. De la misma manera entre las categorías insulares Canarias y Baleares. Gran diferenciación del resto de Ceuta y Melilla, lo que cobra sentido puesto que la estructura económica

de estas ciudades autónomas tiene sus particularidades con respecto al resto de CC.AA.

Dos AC nos ha permitido tener una composición por regiones y comunidades autónomas de la realidad de la negociación colectiva de forma muy visual. Esta es la gran utilidad del AC, que, aunque tiene detrás procesos estadísticos y matemáticos relativamente complejos, acompañados de reducción de dimensiones, es posible presentarlo de una forma simple para conocer la realidad de un vistazo. Este es el gran valor del AC, permitiendo a los técnicos públicos tratar variables cualitativas y representarlas en un plano de dos ejes para su fácil interpretación.

Por otro lado, con respecto al análisis de conglomerados bietápico se pueden subrayar los siguientes resultados obtenidos:

- La decisión automática del número de agrupaciones ha sido tres, con buenos valores de cohesión intra-grupos y diferenciación entre-grupos (>0,5).

- El grupo 1 tiene pocas observaciones (3,9%) y se caracteriza por ser convenios con un gran número de trabajadores y variaciones salariales moderadas. Sobre todo, este grupo está formado por convenios colectivos de sector mayoritariamente, que usualmente por su naturaleza suelen acoger a gran número de trabajadores.

- El grupo 2 tiene un 23% y se caracteriza por ser convenios que acogen a pocos trabajadores pero con unas altas variaciones salariales pactadas. Este grupo lo forman en su mayoría convenios colectivos de empresa que han experimentado un desarrollo económico en su actividad desde 2011.

- El grupo 3 tiene un 73% de las observaciones, siendo el grupo claramente mayoritario. Se caracteriza por ser convenios que acogen también a pocos trabajadores, pero que en este caso tienen variaciones salariales pactadas bajas, inclusive negativas. En el conglomerado predominan convenios colectivos de empresa que, en su mayoría, afectadas por la crisis, han tenido que pactar subidas salariales bastante bajas o inclusive bajadas salariales.

- El IPC es un valor clave para diferenciar grupos. Por lo general, las empresas en reunión con los representantes sindicales y los comités de empresa, pactan subidas salariales siempre teniendo en cuenta el IPC del ejercicio anterior. Por ello, el análisis de conglomerados ha identificado el IPC como un claro elemento diferenciador. Encontramos el grupo que pacta subidas salariales por encima del IPC (grupo 2) y las que lo suelen pactar por debajo (grupo 1 y 3).

- También la naturaleza del convenio es un gran elemento diferenciador para conformar los grupos. El grupo 2 está conformado por un gra número de convenios de sector, puesto que suelen ser convenios que acogen a más trabajadores que aquellos de empresa (como lo son el grupo 1 y 3).

El análisis de conglomerados, de esta manera, nos ha permitido identificar los grupos de convenios colectivos en el territorio español. A posteriori, después de identificar los grupos, se han podido caracterizar para, de esta forma, saber por qué la técnica de conglomerados ha hecho presentes agrupaciones. Permitiendo encontrar parámetros ocultos, si los hay, que hayan propiciado estos conglomerados. De esta manera, podemos tener una mejor composición de los convenios colectivos en España y cuales son los parámetros que diferencian o homogenizan las observaciones.

Queda demostrado que las técnicas de minería de datos pueden servir de gran ventaja y utilidad a la administración pública para conocer, como en este caso, como es la realidad de la negociación colectiva en el estado español. Por lo que, desde la humildad y modestia de presente trabajo de fin de máster se quiere incentivar el uso de la minería de datos dentro del ámbito público, puesto que puede servir de gran ayuda tanto para la eficiencia de los entes estatales, como para la toma de decisiones de los entes políticos. Vivimos en la sociedad de la información y el estado cada vez maneja mayores e ingentes cantidades de datos que, desde la minería de datos, pueden ser procesados para sacar conclusiones sobre ellos. En última instancia son los ciudadanos los que pueden verse beneficiados por las virtudes de la estadística y en particular de la minería de datos, la cual debe ser un pilar fundamental dentro de los organismos públicos y estatales.

Bibliografía

Benzècri, J.-P. (1977). Les cahiers de l'analyse des données. *Numdam*, 125-144.

C, F., & Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 578-588.

Gang-Hoon Kim, S. T.-H. (2014). Big-Data applications in the goverment sector. *Communications of the ACM*, 78-85 .

Greenacre, M. (2008). *La práctica del análisis de correspondencias.* Bilbao: Fundación BBVA.

Maria Belen Castañéda, A. F. (2010). *Procesamiento de datos y análisis estadístico utilizando SPSS.* Porto Alegre: EdiPUCRS.

Rodolfo Mosquera, L. P.-O. (2016). *Metodología para la Predicción del Grado de Riesgo Psicosocial en Docentes de Colegios Colombianos utilizando Técnicas de Minería de Datos.* Bogota: Universidad Nacional de Colombia.

Rousseeuw., L. K. (1990). *Finding groups in data : An introduction to cluster analysis.* New York: Wiley.

Santana, O. F. (1991). El análisis clúster: Aplicación , interpretación Validación. *"Papers" Revista de Sociologia*, 65-76.

Seiffert, J. W. (2006). Data mining : An overview. En D. D. Pegarkov, *National Security Issues* (págs. 201-217 (chapter 11)). New York: Nova Science Publishers.

SPSS. (2001). *The SPSS TwoStep:A scalable component enabling more efficient customer segmentation.* New York.

Zanasi, A. (1998). Competitive Intelligence through data mining public sources. *Competitive Intelligence Review*, Vol. 9(1) 44–54.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record*, 103-114.

Bibliografía jurídica

Real Decreto Legislativo 2/2015, por el que se aprueba el texto refundido de la Ley del Estatuto de los Trabajadores. Madrid. España a 23 de octubre de 2015.

Ley 12/1989, de la Función Estadística Pública. Madrid. España a 9 de mayo de 1989.

Real Decreto 903/2018 por el que se desarrolla la estructura orgánica básica del Ministerio de Trabajo, Migraciones y Seguridad Social. Madrid. España a 20 de julio de 2018.

Real Decreto 1043/2017, por el que se aprueba el Programa anual 2018 del Plan Estadístico Nacional 2017-2020.Madrid . España a 22 de diciembre de 2017.

Anexos

Código SAS

```
*/ GENERO LIBRERÍA*/;

LIBNAME DAT 'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP';
RUN;

*/ IMPORTO DATOS DEL EXCEL (CONVENIOS REGISTRADOS)*/;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2011.xlsx
'
 out  =  total2011
 dbms  =  xlsx
 replace
 ;
run;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2012.xlsx
'
 out  =  total2012
 dbms  =  xlsx
 replace
 ;
run;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2013.xlsx
'
 out  =  total2013
 dbms  =  xlsx
 replace
 ;
run;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2014.xlsx
'
 out  =  total2014
 dbms  =  xlsx
 replace
 ;
run;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2015.xlsx
'
 out  =  total2015
 dbms  =  xlsx
 replace
 ;
run;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2016.xlsx
'
```

```
 out  =  total2016
 dbms  =  xlsx
 replace
 ;
run;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2017.xlsx
'
 out  =  total2017
 dbms  =  xlsx
 replace
 ;
run;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2018.xlsx
'
 out  =  total2018
 dbms  =  xlsx
 replace
 ;
run;

*/ IMPORTO DATOS DEL EXCEL (AJUSTES Y REVISIONES SALARIALES)*/;

proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2011.xlsx
'
 out  =  sal2011
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2012.xlsx
'
 out  =  sal2012
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2013.xlsx
'
 out  =  sal2013
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2014.xlsx
'
 out  =  sal2014
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;
```

```
 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2015.xlsx
'
 out  =  sal2015
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2016.xlsx
'
 out  =  sal2016
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2017.xlsx
'
 out  =  sal2017
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total2018.xlsx
'
 out  =  sal2018
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total.xlsx'
 out  =  total
 dbms = xlsx
 replace;
 run;

 proc import datafile  =
'C:\Users\jacob\Desktop\Santiago\Máster\TFM\DATOS_MANIP\total.xlsx'
 out  =  sal_total
 dbms = xlsx
 replace;
 sheet = 'GARANTÍAS Y REVISIONES SALARIAL';
 run;

 */ Ahora vamos con los SORT para poder hacer MERGE*/;

 proc sort data=total2011;
by C_digo_de_Acuerdo;
run;

 proc sort data=total2012;
by C_digo_de_Acuerdo;
run;
```

```
 proc sort data=total2013;
by C_digo_de_Acuerdo;
run;

 proc sort data=total2014;
by C_digo_de_Acuerdo;
run;

 proc sort data=total2015;
by C_digo_de_Acuerdo;
run;

 proc sort data=total2016;
by C_digo_de_Acuerdo;
run;

 proc sort data=total2017;
by C_digo_de_Acuerdo;
run;

 proc sort data=total2018;
by C_digo_de_Acuerdo;
run;

 proc sort data=sal2011;
by C_digo_de_Acuerdo;
run;

 proc sort data=sal2012;
by C_digo_de_Acuerdo;
run;

proc sort data=sal2013;
by C_digo_de_Acuerdo;
run;

proc sort data=sal2014;
by C_digo_de_Acuerdo;
run;

proc sort data=sal2015;
by C_digo_de_Acuerdo;
run;

proc sort data=sal2016;
by C_digo_de_Acuerdo;
run;

proc sort data=sal2017;
by C_digo_de_Acuerdo;
run;

proc sort data=sal2018;
by C_digo_de_Acuerdo;
run;


 */ Ahora vamos con la sentencia MERGE */;

data merge_total2011;
```

```
merge total2011 sal_total2011;
by C_digo_de_Acuerdo;
run;

data merge_total2012;
merge total2012 sal_total2012;
by C_digo_de_Acuerdo;
run;

data merge_total2013;
merge total2013 sal_total2013;
by C_digo_de_Acuerdo;
run;

data merge_total2014;
merge total2014 sal_total2014;
by C_digo_de_Acuerdo;
run;

data merge_total2015;
merge total2015 sal_total2015;
by C_digo_de_Acuerdo;
run;

data merge_total2016;
merge total2016 sal_total2016;
by C_digo_de_Acuerdo;
run;

data merge_total2017;
merge total2017 sal_total2017;
by C_digo_de_Acuerdo;
run;

data merge_total2018;
merge total2018 sal_total2018;
by C_digo_de_Acuerdo;
run;
```