UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Análisis Estadístico Multivariante de Licitaciones de Compra del Gobierno de Chile

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Velasquez Pizarro, Alejandro

Tutor/a: Zarzo Castelló, Manuel

CURSO ACADÉMICO: 2021/2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

MÁSTER UNIVERSITARIO EN INGENIERÍA DE ANÁLISIS DE DATOS, MEJORA DE PROCESOS Y TOMA DE DECISIONES

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESPAÑA

TRABAJO FIN DE MÁSTER:

"ANÁLISIS ESTADÍSTICO MULTIVARIANTE DE LICITACIONES DE COMPRA DEL GOBIERNO DE CHILE"

Valencia, julio 2022

Alumno: Velásquez Pizarro, Alejandro Iván

Profesor: Dr. Zarzo Castelló, Manuel

Dedicatoria

A Mahra, Dalma, Claudia y Vicenta, por ese… !te quiero familia! que se siente todos los días, por buscar ese algo más que no se sabe que es y debiera estar por ahí, por ir a por los anhelos cuando están allá lejos, de plantar cara cuando hay tiempos difíciles y de creer en que podemos vencer juntos.

A Rodrigo, te admiro hermano mío por tu resiliencia y dedicación a la familia.

Resumen

En la Unión Europea, el Estado ejecuta una parte importante de su presupuesto mediante contratos públicos con terceros, que se asignan por la vía de licitaciones. El gasto promedio anual en Europa es el 12% del PIB y en América Latina llega al 5%. España, en 2019, publicó más de 100 mil licitaciones por valor de 72.500 millones de euros aproximadamente. Chile, en 2020, publicó más de 79 mil licitaciones por 10.500 millones de euros.

Para implementar y gestionar este ejercicio presupuestario, la Unión Europea ha desarrollado un marco institucional mediante políticas OCDE y marcos metodológicos Eurostat, entre otros, los cuales han establecido procesos transparentes de contratación pública, también adoptados por numerosos países de América Latina. El mecanismo de licitaciones de los países OCDE favorece la promoción y publicidad de cada acto administrativo, disminuyendo las barreras de entrada a los proveedores. Las tecnologías desarrolladas para gestionar estos procesos generan información relevante para los “actores del ecosistema de licitaciones”, entre los que identificamos los proveedores o postores (proveen los servicios), los mandantes (instituciones públicas), las empresas del sector financiero (aportan capital de trabajo) y, por último, las entidades del Estado (licitan, fiscalizan, regulan y monitorean).

En este escenario con cientos de miles de licitaciones que representan un volumen de negocios importante, con miles de proveedores que participan sin información de su competencia, resulta natural analizar lo que plantea la teoría de subastas estudiada por Milgrom, Wilson, Klemperer y otros autores, respecto a los incentivos y equilibrios que se producen cuando el postor puja por ganar sin conocer a priori el precio del activo al que ofertarán los demás postores. Al respecto, la teoría de subastas de Milgrom y Wilson enfatiza la importancia de contar con información oportuna y segmentada de valores estimados de precios de los activos de otros postores. Esto sugiere que la posibilidad de segmentar licitaciones es un paso en este sentido, cuyo resultado reduce la incertidumbre del postor y aumenta el beneficio esperado en cada licitación (Wilson, 1960 y 1970).

Con esta motivación, el presente Trabajo Final de Máster (TFM) se plantea como objetivo elaborar una herramienta de clasificación de licitaciones que entregue a los postores información segmentada, basada en técnicas estadísticas de análisis multivariante. El método aplicado cumple con ser una herramienta genérica, sencilla de poner en práctica y cuyo resultado, fácil de interpretar, permite resolver eficientemente los inconvenientes de manipulación de datos. Al aportar valor a todos los agentes, favorece la competencia, reduce la colusión y el abuso de la posición dominante, elementos que se discuten en el debate del comité de competencia de la OCDE, realizado en octubre de 2006, donde plantea la importancia del diseño de subastas.

En este estudio los datos utilizados son las 330.000 licitaciones del Estado de Chile registradas en 36 meses consecutivos desde enero 2018 a diciembre 2020. Todos los productos y servicios de una licitación se encuentran codificados a 8 dígitos según el UNSPSC (Código Estándar de Productos y Servicios de la ONU). Para la aplicación del método se considerará un código agregado a 2 dígitos. Así, la canasta de cada licitación se puede representar como un vector de 55 variables categóricas, donde el descriptor tomará el valor 1 cuando tenga asociado uno o más productos y 0 en caso contrario. Considerando, en este caso, la canasta como “el conjunto de productos reunidos en una sola compra (licitación), en un instante de tiempo”.

Tomando en cuenta la enorme cantidad de licitaciones, se ha seleccionado una muestra aleatoria equilibrada representativa de 10.000 observaciones, a partir de la cual serán necesarios tres pasos para modelar el clasificador.

En el primer paso, dado que la variable original es categórica, se propone el análisis de correspondencia múltiple (MCA) basado en variables latentes, para un análisis de canasta exploratorio. En el segundo paso, con una función de segmentación jerárquica se obtienen 50 segmentos de licitaciones. Finalmente, en el tercer paso, la función SVM (*Support Vector Machine*) permite modelar un clasificador que asigna, con alta precisión, la clase que más se le aproxima. Con este clasificador de licitaciones ha sido posible procesar el 100% de las licitaciones que se llevan a cabo en la plataforma electrónica de contratación pública del gobierno de Chile (*e-procurement*) con una eficiencia promedio del 93,8% de aciertos en la clasificación.

El potencial práctico del clasificador de licitaciones propuesto en este TFM radica en que la herramienta ha sido construida a partir de datos públicos, proporcionando a los agentes información valiosa del segmento al que pertenece una licitación, que servirá posteriormente para estimar precios de los activos basados en precios históricos.

El ámbito de aplicación del clasificador de licitaciones no se limita sólo al caso chileno, se extiende también a los países que siguen las recomendaciones y buenas prácticas OCDE en términos de contratos públicos basados en licitaciones.

Se trata de un caso de clasificación de licitaciones con variables categóricas, ya que los descriptores del vector asociado a una canasta toman el valor “0” o “1”, y por tanto es adecuado para ser analizado con MCA. Sin embargo, nos ha parecido interesante analizar los resultados y confiabilidad de la prueba si lo consideramos un caso de variable binaria, mediante la aplicación de análisis de componentes principales (PCA) y reglas de asociación (AR).

Retomando estos resultados y la confiabilidad del análisis de canasta de licitaciones (Paso 1), se realizó una comparación de tres técnicas, MCA, AR y PCA, obteniéndose que, bajo ciertas condiciones, no se puede descartar ninguna, pues todas son potencialmente útiles y efectivas para el Paso 3, que modela el clasificador.

Considerando que se trata de una herramienta sencilla de poner en práctica, cabe mencionar que el proceso completo de modelización del clasificador a partir de una muestra de 10.000 licitaciones requiere un tiempo computacional de 5 minutos, lo que permite ser rediseñado con facilidad. Por otro lado, se sabe que se publican unas 500 licitaciones diarias, este conjunto de datos nuevos puede clasificarse en menos de 10 segundos.

Palabras Clave: **Contratación Pública**, **Subasta**, Licitaciones, Postores, Puja, Mercado de ofertas, Análisis de Correspondencia Múltiple (**MCA)**, Análisis de Componentes Principales (**PCA)**, Análisis de Canasta de Mercado (**MBA)**, Reglas de Asociación (**AR**), Organización para la Cooperación y el Desarrollo Económicos (OCDE), Registros Administrativos.

Abstract

At the European Union, the State executes an important part of its budget through third parties public procurement, assigned through public tenders. The annual mean expense in Europe is 12% of the GDP and in Latin America it reaches 5%. Spain, in 2019, issued more than 100.000 public tenders worth 72.500 million euros approximately. Chile, in 2020, issued more than 79.000 public tenders worth 10.500 million euros.

To implement and manage this budget exercise, the European Union has developed an institutional framework through OECD policies and methodological frameworks, Eurostat among others, which have established transparent public procurement processes, also adopted by a number of countries in Latin America. The mechanism of public tenders of the OECD countries encourages the promotion and advertisement of each administrative act, alleviating the entry barriers to suppliers. Technologies developed to manage these processes produce information relevant for the “public tenders ecosystem actors”, among which we identify suppliers or bidders (providing services), public buyers (public institutions), companies from the financial sector (contributing labor capital) and, lastly, State entities that bid, audit, regulate and monitor the performance of a bidding process.

In this scenario with hundreds of thousands of public tenders representing an important business volume, with thousands of providers involved without information about their competition, it becomes natural to analyze what was proposed by the auction theory studied by Milgrom, Wilson, Klemperer and others, regarding the incentives and stability produced when the bidder bid to win without knowing in advance the price of the asset being offered by the other bidders. In this regard, Milgrom and Wilson’s auction theory emphasizes the importance of having opportune and segmented information on the estimated price values of the other bidder’s assets. This suggests that the possibility to segment public tenders would be a step forward, which outcome reduces the bidder’s uncertainty and rises the expected benefit on each public tender (Wilson, 1960 and 1970).

With this incentive, the present Final Master Project (TFM) poses the objective of elaborating a public tenders classification tool that provides segmented information to bidders, based on statistical techniques of multivariate analysis. The applied method complies with being a generic tool, simple to implement and which outcome, easy to interpret, allows to overcome efficiently the inconvenients of data manipulation. By providing value to the agents, it promotes the competition, reduces collusion and abuse of the dominant position, elements discussed in the competition committee debate of the OECD, undertaken in October 2006, where the importance of auction design is raised.

In this study, the data set comprised all 330.000 public tenders of the Chilean State registered in 36 consecutive months from January 2018 to December 2020. All products and services of a public tender are codified in 8 figures according to UNSPSC codes (United Nations Standard Products and Services Code). To implement the method, they will be considered as a two-digit code. Thus, each tender’s basket can be represented as a vector of 55 categorical variables, where the descriptor’s value is 1 when it is associated to one or more products and 0 otherwise. In this case, taking into consideration the basket as the “items set together in a single purchase at the same point of time”.

Considering the big number of public tenders, a sample was obtained by random selection, balanced and representative of 10.000 observations, where three steps will be necessary to model the classifier.

In the first step, since the original variable is categorical, a Multiple Correspondence Analysis (MCA) is suggested based on the latent variables, for an exploratory Basket Market Analysis (MBA). In the second step, with a hierarchical segmentation function, 50 tender segments were obtained. Finally, in the third step, the function SVM (Support Vector Machine) allows to model a classifier that assigns, with high accuracy, the closest class. This public tender classifier has made possible to process 100% of the public tenders published in the government's electronic platform for public procurement (*e-procurement*) with an average efficiency of 93,6% of successful classification.

The practical potential of the tender classifier proposed in this TFM lies in the tool developed from public data, providing valuable information to the agents about the segment to which the public tender belongs, which can be used later to estimate prices of the assets based on historical prices.

The scope of the tender classifier is not limited to the Chilean case, it extends to countries following the OECD recommendations and good practices in terms of public procurement based on public tender.

It is a tender classification with categorical variables, since the vector descriptors associated to a basket are valued "0" or "1", and therefore it is adequate to be analyzed with MCA. However, it is interesting to analyze the results considering a binary variable case, by applying principal components analysis (PCA) and association rules (AR).

Taking back the results from the tender basket analysis in the first step, a comparison was performed through three techniques (i.e., MCA, AR and PCA). Results lead us to conclude that, under certain conditions, none can be discarded, since they all are potentially useful and effective to execute third step, to model the classifier.

Considering it as a rather simple tool to implement, it should be mentioned that the entire process of modeling the classifier from a sample of 10.000 public tenders requires a total computational time of 5 minutes, which allows an easy redesign. Besides, about 500 public tenders are issued every day, and this new data can be classified in less than 10 seconds.

Key Words: **Public Procurement**, **Auction**, Tender, Bidders, Bid, Bidding markets, Winner's Curse, Multiple Correspondence Analysis (MCA), Principal Component Analysis (PCA), Basket Market Analysis (MBA), Association Rule Mining, Organization for Economic Cooperation and Development (OECD), Business Registers.

Resum

A la Unió Europea, l'Estat executa una part important del seu pressupost mitjançant contractes públics amb tercers, que s'assignen per la via de licitacions. La despesa mitjana anual a Europa és el 12% del PIB i a Amèrica Llatina arriba al 5%. Espanya, en 2019, va publicar més de 100 mil licitacions per valor de 72.500 milions d'euros aproximadament. Xile, en 2020, va publicar més de 79 mil licitacions per 10.500 milions d'euros.

Per a implementar i gestionar aquest exercici pressupostari, la Unió Europea ha desenvolupat un marc institucional mitjançant polítiques OCDE i marcs metodològics Eurostat, entre altres, els quals han establit processos transparents de contractació pública, també adoptats per nombrosos països d'Amèrica Llatina. El mecanisme de licitacions dels països OCDE afavoreix la promoció i publicitat de cada acte administratiu, disminuint les barreres d'entrada als proveïdors. Les tecnologies desenvolupades per a gestionar aquests processos generen informació rellevant per als "actors de l'ecosistema de licitacions", entre els quals identifiquem els proveïdors o postors (proveeixen els serveis), els mandants (institucions públiques), les empreses del sector financer (aporten capital de treball) i, finalment, les entitats de l'Estat (liciten, fiscalitzen, regulen i monitoren).

En aquest escenari amb centenars de milers de licitacions que representen un volum de negocis important, amb milers de proveïdors que participen sense informació de la seua competència, resulta natural analitzar el que planteja la teoria de subhastes estudiada per *Milgrom, Wilson, *Klemperer i altres autors, respecte als incentius i equilibris que es produeixen quan el postor licita per guanyar sense conéixer a priori el preu de l'actiu al qual oferiran els altres postors. Sobre aquest tema, la teoria de subhastes de Milgrom i Wilson emfatitza la importància de comptar amb informació oportuna i segmentada de valors estimats de preus dels actius d'altres postors. Això suggereix que la possibilitat de segmentar licitacions és un pas en aquest sentit, el resultat del qual redueix la incertesa del postor i augmenta el benefici esperat en cada licitació (Wilson, 1960 i 1970).

Amb aquesta motivació, el present Treball Final de Màster (TFM) es planteja com a objectiu elaborar una eina de classificació de licitacions que entregue als postors informació segmentada, basada en tècniques estadístiques d'anàlisis multivariant. El mètode aplicat compleix amb ser una eina genèrica, senzilla de posar en pràctica i el resultat de la qual, fàcil d'interpretar, permet resoldre eficientment els inconvenients de manipulació de dades. En aportar valor a tots els agents, afavoreix la competència, redueix la col·lusió i l'abús de la posició dominant, elements que es discuteixen en el debat del comité de competència de l'OCDE, realitzat a l'octubre de 2006, on planteja la importància del disseny de subhastes.

En aquest estudi, les dades utilitzades són les 330.000 licitacions de l'Estat de Xile registrades en 36 mesos consecutius des de gener 2018 a desembre 2020. Tots els productes i serveis d'una licitació es troben codificats a 8 dígits segons el UNSPSC (Codi Estàndard de Productes i Serveis de l'ONU). Per a l'aplicació del mètode es consideraran amb un codi agregat a 2 dígits. Així, la canastra de cada licitació es pot representar com un vector de 55 variables categòriques, on el descriptor prendrà el valor 1 quan tinga associat un o més productes i 0 en cas contrari. Considerant, en aquest cas, la canastra com "el conjunt de productes reunits en una sola compra (licitació), en un instant de temps".

Tenint en compte l'enorme quantitat de licitacions, s'ha seleccionat una mostra aleatòria equilibrada representativa de 10.000 observacions, a partir de la qual seran necessaris tres passos per a modelar el classificador.

En el primer pas, atés que la variable original és categòrica, es proposa l'anàlisi de correspondència múltiple (MCA) basat en variables latents, per a una anàlisi de canastra exploratori. En el segon pas, amb una funció de segmentació jeràrquica s'obtenen 50 segments de licitacions. Finalment, en el tercer pas, la funció SVM (*Support Vector Machine*) permet modelar un classificador que assigna, amb alta precisió, la classe que més se li aproxima. Amb aquest classificador de licitacions ha sigut possible processar el 100% de les licitacions que es duen a terme en la plataforma electrònica de contractació pública del govern de Xile (e-procurement) amb una eficiència mitjana del 93,6% d'encerts en la classificació.

El potencial pràctic del classificador de licitacions proposat en aquest TFM radica a ser una eina construïda a partir de dades públiques, proporcionant als agents informació valuosa del segment al qual pertany una licitació, que servirà posteriorment per a estimar preus dels actius basats en preus històrics.

L'àmbit d'aplicació del classificador de licitacions no es limita només al cas xilé, s'estén també als països que segueixen les recomanacions i bones pràctiques OCDE en termes de contractes públics basats en licitacions.

Es tracta d'un cas de classificació de licitacions amb variables categòriques, ja que els descriptors del vector associat a una canastra prenen en valor “0” o “1”, i per tant és adequat per a ser analitzat amb MCA. No obstant això, ens ha semblat interessant analitzar els resultats i confiabilitat de la prova si ho considerem un cas de variable binària, mitjançant l'aplicació d'anàlisi de components principals (PCA) i regles d'associació (AR).

Reprenent aquests resultats i la confiabilitat de l'anàlisi de canastra de licitacions del Pas 1, es va realitzar una comparació de tres tècniques, MCA, AR i PCA, obtenint-se que, sota unes certes condicions, no es pot descartar cap, perquè totes són potencialment útils i efectives per al Pas 3, que modela el classificador.

Considerant que es tracta d'una eina senzilla de posar en pràctica, cal esmentar que el procés complet de modelització del classificador a partir d'una mostra de 10.000 licitacions requereix un temps computacional de 5 minuts, la qual cosa permet ser redissenyat amb facilitat. D'altra banda, se sap que es publiquen unes 500 licitacions diàries, aqueixa dada nova pot classificar-se en menys de 10 segons.

Paraules Clau: Contractació Pública, Subhasta, Licitacions, Postors, Licitació, Mercat d'ofertes, Anàlisis de Correspondència Múltiple (MCA), Anàlisi de Components Principals (PCA), Anàlisis de Canastra de Mercat (MBA), Regles d'Associació (AR), Organització per a la Cooperació i el Desenvolupament Econòmics (OCDE), Registres Administratius.

Tabla de contenido

1 INTRODUCCIÓN Y ANTECEDENTES

1.1 Contratos públicos

1.1.1 OCDE: Gobernanza en contratos públicos

La Organización para la Cooperación Económica Europea (OEEC) fue creada en 1948 para implementar el Plan Marshall tras la segunda Guerra Mundial. Posteriormente, el 14 de diciembre de 1960 en el Chateau de la Muette en París, se firmó el convenio que transformó la OEEC en La Organización para la Cooperación y el Desarrollo Económicos (OCDE).

Este convenio de 1960, que entró en vigor el 30 de septiembre de 1961, contaba en su inicio con 19 países, incluida España, hasta llegar a fines de 2021 con 38 países, entre los cuales se encuentran Chile y otros países de América Latina. Cabe señalar que desde un inicio han participado países que no forman parte de la Unión Europea.

Tabla 1: Países miembros OCDE 2021
Fuente: web oficial www.oecd.org

Alemania	Eslovenia	Israel	Polonia
Australia	España	Italia	Portugal
Austria	Estados Unidos	Japón	Reino Unido
Bélgica	Estonia	Letonia	República Checa
Canadá	Finlandia	Lituania	República Eslovaca
Chile	Francia	Luxemburgo	Suecia
Colombia	Grecia	México	Suiza
Corea del Sur	Hungría	Noruega	Turquía
Costa Rica	Irlanda	Nueva Zelanda	
Dinamarca	Islandia	Países Bajos	

Actualmente, la OCDE tiene como misión "diseñar mejores políticas" y "promover políticas que favorezcan la prosperidad, la igualdad, las oportunidades y el bienestar para todas las personas". Colabora con más de 100 países no miembros, muchos de los cuales incluso participan en sus más de 200 comités y grupos de trabajo de expertos.

Adicionalmente, Brasil, India, Indonesia, la República Popular China y Sudáfrica son socios estratégicos y participan en programas de trabajo específicos. También mantiene relaciones con otros organismos, entre los cuales se destaca la Organización Internacional del Trabajo (OIT), la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), el Fondo Monetario Internacional (FMI), el Banco Mundial, la ONU y el Organismo Internacional de Energía Atómica, entre otras.

En su Marco de políticas para una buena **gobernanza pública**, la OCDE define este concepto como "*la formulación, ejecución y evaluación de reglas, procesos e interacciones formales e informales entre las instituciones y los agentes que componen el Estado, y entre el Estado y los ciudadanos*."

La división de gobernanza pública de la OCDE, bajo el concepto de **Gobernanza**[1] (G) busca fortalecer el ejercicio de la autoridad política, económica y administrativa. Es la entidad encargada de crear comisiones que permitan fortalecer, elaborar, actualizar, recomendar, difundir, promover y evaluar las buenas prácticas entre los 38 países miembros en materia de

[1] *Glosario de términos estadísticos de la OCDE*. www.oecd-ilibrary.org

gobierno digital, políticas anticorrupción, comercio ilícito, infraestructura, innovación y, muy en particular, la contratación pública para mejorar y transparentar el gasto público.

La **Contratación Pública** (CP) definida en el comité de gobernanza pública de la OCDE de 2015, se refiere a la compra por parte de los gobiernos y las empresas estatales de bienes, servicios y obras. Dado que la contratación pública es el 12% del PIB en los países OCDE, se espera que los gobiernos la lleven a cabo de manera eficiente y con altos estándares de conducta a fin de garantizar la alta calidad de la prestación del servicio y salvaguardar el interés público.

De lo anterior, se desprenden las recomendaciones OCDE que más influyen en dar eficiencia y transparencia a la contratación pública, las cuales deben complementarse con el diseño de licitaciones, definido en la "teoría de subastas" de Milgrom y Wilson (2004), y con una plataforma que gestione todo el proceso licitatorio de contratación electrónica.

1.1.2 Eurostat: Oficina Estadística de la Unión Europea

Eurostat se encarga de publicar estadísticas e indicadores de alta calidad a escala europea, que permitan hacer comparaciones entre países y regiones (https://ec.europa.eu).

Entre sus responsabilidades se encuentran:

– desarrollar definiciones, clasificaciones y metodologías armonizadas para la elaboración de las estadísticas oficiales europeas, en colaboración con las autoridades estadísticas nacionales.
– calcular los datos agregados para la Unión Europea y la zona euro a partir de la información recopilada por las autoridades estadísticas nacionales según normas armonizadas.
– poner estadísticas europeas a libre disposición de los responsables y los ciudadanos a través de la web de Eurostat y otros canales.

Algunas definiciones que rigen las recomendaciones de Eurostat están también alineadas con la OCDE para dar coherencia, pero en este caso aportan a la estandarización y validación de las estadísticas europeas y de los países miembros, que se basan en registros administrativos. Algunas de sus definiciones derivadas son las siguientes:

Registros administrativos (RA: *Register Based Administrative Data*): es la serie de datos sobre un hecho, evento, suceso o acción sujeta a regulación o control, recabados por una oficina del sector público como parte de su función.

El libro de Wallgren & Wallgren (2014) contiene una presentación exhaustiva de los registros administrativos RA que alimentan cuatros registros básicos: de población, de empresas, registro inmobiliario y de actividades. Estos han sido generados por las diferentes unidades del Estado responsables del control y seguimiento de las acciones que los generan. Además, se revisan las experiencias aplicadas en el documento metodológico de la CEPAL para el aprovechamiento estadístico de registros administrativos económicos (CEPAL, 2021)

Los términos BR y SBR que se definen a continuación están documentados en el "Manual de Recomendaciones para Registros Administrativos" de Eurostat (2003):

Registros Administrativos Económicos (BR: *Business Register*): el término se refiere a listas de empresas y otras unidades relacionadas que hayan sido registradas, ya sea de forma voluntaria o por reglamentación, y cuyas actividades contribuyen al Producto Interior Bruto

(PIB) del Estado Miembro. Estas unidades pueden definirse como aquellas que ejercen control sobre el uso de recursos (incluidos la tierra, el trabajo, el capital, los bienes y servicios) y tienen el objetivo de producir bienes y servicios para su propio consumo o para el consumo de otras unidades.

Registros Administrativos económicos de carácter estadístico (SBR *Statistical Business Registers*): son repositorios referidos a las unidades de existencia legal o estadística, para utilizarlos en el desarrollo de estadísticas de negocios y macroeconómicas. Incluyen información de la población activa de Empresas o grupos de empresas con actividad económica, que aportan al PIB; unidades legales de esas empresas; unidades productivas locales u otros tipos de unidades activas. Se dice que estos registros son la columna vertebral en la producción de estadísticas económicas pues proveen la infraestructura central que asegura la consistencia de los datos entre diferentes fuentes.

Directorio Central de Empresas (DIRCE), es un registro administrativo conjuntamente del tipo empresa y actividad, que es de carácter estadístico (SBR). El DIRCE reúne, en un sistema de información único, a todas las empresas españolas y a sus unidades locales ubicadas en el territorio nacional. Su objetivo básico es hacer posible la realización de encuestas económicas por muestreo. Se actualiza una vez al año, generándose un nuevo sistema de información a 1 de enero de cada período.

Según el Instituto Nacional de Estadísticas de España (INE), la explotación estadística del directorio central de empresas contiene información agregada de las empresas y unidades locales que operan en el territorio nacional. El 1° de enero de cada año, se publica una explotación estadística de los resultados para empresas y unidades locales, desglosados por comunidades autónomas, condición jurídica, actividad económica o estrato de asalariados asignado.

1.1.3 Escenario actual en Chile

En 2009 Chile ingresó como país invitado a la OCDE, para lo cual presentó previamente su postulación, en 2003, con un plan de trabajo que consiste en mejorar la calidad de todos los Registros Administrativos (RA), estandarizar su uso e implementar un Directorio Económico-Estadístico (SBR: *Statistical business registers*) que permita transparentar el gasto y sistematizar la gestión de las instituciones públicas.

Con el propósito de habilitar un registro administrativo (RA) específico para la Contratación Pública (CP), ese mismo año 2003 se crea la institución *ChileCompra* dependiente del Ministerio de Hacienda, con la misión de **"generar eficiencia en la contratación pública con altos estándares de probidad y transparencia"**. A partir de 2004 se implementa una plataforma de licitaciones (registro administrativo de licitaciones) para la subasta del contrato público de productos, servicios y obras, que está disponible en el portal web www.mercadopublico.cl para consultas *online* (de uno en uno). A este nivel de detalle se identifica cada día, de manera actualizada minuto a minuto, quién compra, quién vende, cuánto, cuándo, qué vende y en qué estado administrativo se encuentra el proceso licitatorio.

En 2020 esta plataforma registró más de 100 mil licitaciones (contratos públicos) provenientes de 911 instituciones públicas (mandantes) y más de 25.000 proveedores vigentes en el directorio económico de empresas y establecimientos (*business register* BR). Las compras públicas a través de la plataforma de licitaciones de ChileCompra, transaron 10.483 millones de euros, equivalentes al 4.73% del PIB.

Otras instituciones como el Instituto Nacional de Estadísticas (INE), Ministerio de Hacienda, Banco Central de Chile (BCCH), Servicio de Impuestos Internos (SII), Aduanas, Tesorería General de la República (TGR), entre muchos otros estamentos gubernamentales, también tienen sus propios registros administrativos que permiten acceder a datos complementarios, que enriquecen la información del proceso licitatorio de ChileCompra y que aportan características adicionales de las empresas proveedoras del Estado.

En este mismo ámbito, la Cámara de Comercio de Santiago (CCS) procesa otro registro administrativo de **pago** (no público) que aporta diariamente un detalle de la morosidad y protestos que las empresas proveedoras mantienen con el sector financiero y el comercio formal, consolidando información de todas las transacciones impagas que éstas mantienen a nivel nacional. El servicio de entregar el detalle de esta información de empresas se comercializa a través de empresas especializadas del sector privado.

Toda la información del Estado que se considera registro administrativo es pública y sin costo para quien la solicite. En ocasiones, las organizaciones e instituciones del Estado no entregan oportuna y adecuadamente esta información, por tanto, es necesario solicitarla formalmente al Consejo para la Transparencia (CPLT), que es una corporación autónoma de derecho público, creada por la Ley de Transparencia de la Función Pública y de Acceso a la Información de la Administración del Estado (Ley 20285 del 11 de agosto de 2008).

En algunos casos, se puede acceder a registros administrativos desde aplicaciones web previstas para consultas *online* de manera individual (uno en uno) un dato a la vez. También hay disponibles, al menos, dos mecanismos de acceso masivo a datos de registros administrativos. Uno de ellos es por la vía de ***interfaz de programación de aplicaciones***, conocida también por la sigla API, que ofrece una biblioteca de rutinas, funciones y procedimientos para ser utilizada por otro software como una capa de acceso y transferencia de datos. El otro mecanismo es por la vía de ***servicios web*** (*webservices*) que utilizan un conjunto de protocolos y estándares que sirven para intercambiar datos entre aplicaciones.

Esta forma de acceder de manera masiva a los registros administrativos, está orientada a usuarios con conocimientos específicos y experimentados en tecnologías de información. Además, requieren de una fuerte inversión en infraestructura de ***datacenter***, como se denomina al Centro de Proceso de Datos, disponible en servidores ubicados físicamente en las instalaciones de organización o bien externalizados a terceros en otra ubicación física o en la "*nube*" (*Cloud computing*), donde se concentran los recursos necesarios para recibir, almacenar, procesar y transferir los datos a través de internet.

Con el fin de capturar el valor agregado de los registros administrativos, algunos proveedores del Estado invierten en esta infraestructura para procesar, integrar, sistematizar y consolidar la información de manera coherente en un solo repositorio de datos que la procese en línea y entregue acceso a esta información con el suficiente nivel de granularidad y periodicidad que permita explorar nuevas oportunidades de negocios, encontrar equilibrios eficientes y competitivos para evaluar y/o estimar los riesgos financieros, operacionales, segmentar empresas con respecto al historial de ventas, tipo de productos, tipo de mandante y otras variables disponibles en el repositorio consolidado.

En el caso específico de las licitaciones para contratación pública, el valor agregado que proporcionan los registros administrativos se convierte en una herramienta que puede aportar información de valores estimados de los "activos licitados" por los demás proveedores (Milgrom, 1980), logrando reducir la incertidumbre y aumentar el beneficio esperado en cada licitación (Wilson, 1960 y 1970).

1.1.4 Antecedentes en España y otros países

En este punto se aporta información obtenida de los sitios oficiales de entidades de diferentes países, los cuales describen iniciativas que rigen la contratación pública mediante la aplicación de normas alineadas a las recomendaciones internacionales OCDE.

1.1.4.1 España

El 9 de marzo de 2018 entra en vigor en España la nueva Ley 9/2017 de Contratos del Sector Público, la cual incorpora al ordenamiento jurídico español las últimas Directivas del Parlamento Europeo en relación con esta materia. Esta ley crea la Oficina Independiente de Regulación y Supervisión de la Contratación (OIReScon) que tiene por mandato, entre otros, elaborar el "Informe Anual de Supervisión de la Contratación Pública de España". Según este informe, en el año 2019 en España se efectuaron 129.594 licitaciones, por un importe total de 72.527,27 millones de euros (en PBL, sin incluir impuestos).

Según la LEY 9/2017[2] de la Agencia Estatal (BOE-A-2017-12902), *"La legislación de contratos públicos, de marcado carácter nacional, encuentra, no obstante, el fundamento de muchas de sus instituciones más allá de nuestras fronteras, en concreto, dentro de la actividad normativa de instituciones de carácter internacional, como es el caso de la OCDE, de UNCITRAL -en el ámbito de la ONU-, o, especialmente, de la Unión Europea. La exigencia de la adaptación de nuestro derecho nacional a esta normativa ha dado lugar, en los últimos treinta años, a la mayor parte de las reformas que se han ido haciendo en los textos legales españoles".*

El portal electrónico de licitaciones (*e-Procurement*) de España para la contratación pública está centralizado en www.contrataciondelestado.es que reúne a todas las autonomías, el sector público estatal, entidades locales e incluso algunas universidades. El sistema de codificación estándar de categorización de actividades económicas de la Unión Europea es usado para clasificación de productos, servicios y obras en contratos públicos vía licitaciones, este código es el CPV (*Common Procurement Vocabulary*) a 8 dígitos.

1.1.4.2 Otros países

En Colombia, el estamento encargado de la Contratación Pública es "La Agencia Nacional de Contratación Pública - Colombia Compra Eficiente" (ANCP - CCE). El portal de acceso a licitaciones es *ColombiaCompra*[3], creado por medio del Decreto Ley 4170 de 3 de noviembre de 2011. Utiliza el Código de Bienes y Servicios de Naciones Unidas UNSPSC v14_0801[4] para codificar sus licitaciones, al igual que lo hace Chile en varios de los procedimientos administrativos.

México es también un país miembro de la OCDE desde 1994. Su caso es particular, tanto por el tamaño de su economía en América Latina como por el hecho de que **no usa** el mismo mecanismo de licitaciones y que tampoco usa el código de productos estándar UNSPSC de la ONU o su equivalente CPV de la Unión Europea, la alternativa usada por este país es el Código Único de las Contrataciones Públicas (CUCOP). En México, la contratación pública electrónica se implementa en un portal del tipo mercado en línea (*marketplace*), es un tipo de sitio web de comercio electrónico multicanal en el que la información sobre productos o servicios es proporcionada por múltiples terceros. En este *marketplace*, las empresas y personas físicas se registran como proveedores del Estado y suben sus productos a un

[2] https://www.boe.es/buscar/act.php?id=BOE-A-2017-12902
[3] https://colombiacompra.gov.co/secop/secop
[4] https://colombiacompra.gov.co/sites/cce_public/files/cce_clasificador/unspsc_spanish_v14_0801.pdf

catálogo web donde las entidades públicas son compradores directos, sin licitación de por medio.

Este proceso está regulado por la Ley Modelo de Adquisiciones, Arrendamiento de Bienes Muebles y Prestación de Servicios de las Entidades Federativas, propuesta por el IMCO (Instituto Mexicano para la Competitividad). Según esta entidad IMCO, en esta iniciativa se enfrentan los principales problemas que afectan la eficiencia en el proceso de compras gubernamentales, aborda las mejores prácticas vistas en otros países y asume las recomendaciones de la Ley Modelo de la Organización de las Naciones Unidas. Este portal de compras oficial, que depende de Hacienda, se conoce con el nombre de CompraNet. Es una Tienda Digital del Gobierno Federal, administrado y operado por la Unidad de Política de Contrataciones Públicas (UPCP).

1.1.5 Proceso de contratación pública

La mejora continua en la contratación pública va siendo guiada por las recomendaciones y buenas prácticas OCDE a través de la creación de comisiones que mantienen versiones actualizadas de cada manual y recomendación. En 2019 se actualizó la revisión de contratación pública (OECD, 2007b) sobre el "Progreso en la implementación de la recomendación de la OCDE de 2015" (OECD, 2019) .

Este informe de 2019 presenta el progreso realizado por los países de la OCDE y otras economías en su adhesión a la Recomendación del Consejo sobre Contratación Pública de 2015. La recomendación proporciona orientación estratégica para abordar los desafíos que se encuentran en la contratación pública e identifica las buenas prácticas de contratación para garantizar un uso estratégico y holístico de la contratación pública. Este informe analiza la pertinencia continua de la recomendación, qué tan ampliamente se ha difundido y si requiere actualización o revisión.

El informe actualizado del "Comité de Gobernanza Pública" de la OCDE de 2015 sobre "*la aplicación de recomendaciones del Consejo y mejora de la integridad en la contratación pública [C(2008)105]*", identificó los principales problemas a los que se enfrentan los países para mejorar sus sistemas de contratación pública, así como las posibles áreas de mejora [C(2012)98 y C(2012)98/CORR1].

En resumen, esta Recomendación de 2015, considera lo siguiente:

- favorece la asignación adecuada de los recursos públicos, proponiendo recurrir a la contratación pública como herramienta estratégica;
- aporta rentabilidad, pues fomenta una mayor eficiencia en el gasto público: un ahorro de un 1% representa 43.000 millones de euros al año en los países de la OCDE;
- atenúa riesgos como los de la ineficiencia o la corrupción, a menudo muy presentes en proyectos de contratación de grandes infraestructuras y otros de gran complejidad.

Para la OCDE, el contexto general de la **Contratación pública** "se refiere a la compra por parte de los gobiernos y las empresas estatales de bienes, servicios y obras. Dado que la contratación pública representa una parte sustancial del dinero de los contribuyentes, se espera que los gobiernos la lleven a cabo de manera eficiente y con altos estándares de conducta a fin de garantizar la alta calidad de la prestación del servicio y salvaguardar el interés público".

Según el Comité de Gobernanza Pública y con el apoyo del Comité de Competencia de la OCDE, acuerdan que, a efectos de la presente Recomendación (2015), se utilicen las siguientes definiciones, que se mantienen vigentes hasta hoy:

> **Contratación pública (CP): "**se refiere al proceso de identificación de necesidades, la decisión acerca de la persona (física o jurídica) más adecuada para cubrir estas necesidades y, por último, la comprobación de que el bien o prestación se entregan en el lugar correcto, en el momento oportuno, al menor precio posible, y que todo ello se hace con ecuanimidad y transparencia".
>
> **Contratación electrónica (CE)**, en inglés *e-Procurement*, se refiere a la introducción de las tecnologías digitales para la sustitución o el rediseño de los procedimientos en soporte papel presentes a lo largo del proceso de contratación pública.
>
> **Ciclo de la contratación pública (CCP)**, se refiere a la cadena de actividades relacionadas entre sí que comienza por la evaluación de necesidades, pasa por la fase de concurso y adjudicación, y abarca finalmente la gestión contractual y de los pagos, junto con las oportunas tareas de seguimiento o auditoría.

Este documento de recomendaciones del Comité de Gobernanza Pública de la OCDE 2015 no define explícitamente el concepto de Licitación, pero sí la menciona como el elemento clave y central para poner en práctica las recomendaciones actualizadas sobre la mejora de la integridad en la contratación pública, formulada en 2008.

> **Licitación Pública**: "Las licitaciones mediante concurso deberán ser el método habitual en la contratación pública, como instrumento adecuado que son para lograr la eficiencia, combatir la corrupción, obtener unos precios justos y razonables y garantizar unos resultados competitivos".

1.2 Teoría de subastas (licitaciones, un caso particular)

Para contextualizar, se considera la definición de subasta propuesta por McAfee y McMillan (1987): *"la subasta es la institución de mercado que cuenta con un conjunto explícito de reglas que determinan la asignación de recursos y precios basándose en las pujas presentadas por los participantes"* (McAfee & McMillan, 1987).

El trabajo de Durá Juez (2003) sobre la *"teoría de subastas y la reputación del vendedor"*, precisa que las subastas pueden tomar la forma de una licitación o bien de un remate:

- Licitación: se da cuando el objetivo de la subasta es la "adquisición de activos", sean estos productos, servicios, ejecución de una obra o provisión de suministros, en cuyo caso el comprador será el subastador (mandante) mientras que los vendedores serán los postores (proveedores) que deben pujar por adjudicarse (ganar) la licitación.
- Remate: se da cuando el objetivo es la venta de activos, en cuyo caso los roles se invierten, pues el vendedor es subastador y el comprador es el postor. Este tema no se estudia en el presente TFM.

En el trabajo de Durá Juez (2003), se plantean los elementos teóricos de las licitaciones que permiten hacer seguimiento a la reputación del postor (vendedor), bajo el supuesto que el postor es un agente de "corta vida" que participa en pocas licitaciones (pero que en teoría conoce la historia de las licitaciones anteriores), mientras que los compradores son agentes que tienen "larga vida" porque licitan periódicamente la adquisición de activos.

Este planteamiento implica que existe una manera de hacer trazabilidad de licitaciones, vendedores, compradores y del precio de los activos. El inconveniente de este supuesto es que es difícil hacer trazabilidad y seguimiento de los postores de manera individual porque son de "corta vida", pero se podría caracterizar a un postor prototipo que herede todos los atributos del segmento del proveedor al cual se relaciona la segmentación de los activos.

La experiencia de Paul Milgrom planteada en su libro "*Poniendo en práctica la teoría de la subasta*" (Milgrom, 2004), pone de relieve en el capítulo 6 que en realidad es poco habitual que la clave del éxito de una subasta se deba a su diseño novedoso e inteligente. En cambio, se considera que la clave está en conseguir buenos incentivos y asignaciones justas, mantener bajos los costes de licitación, alentar a los postores adecuados a participar, garantizar la integridad del proceso y asegurarse de que el postor ganador sea alguien idóneo que entregue los servicios pactados según lo comprometido.

La revisión que este autor aborda en su libro de todos los tipos de licitaciones de contratación pública electrónica (*e-Procurement*) incluye las que se realizan en la plataforma ChileCompra[5] y en las equivalentes de otros países OECD (Milgrom, 2004). La mayoría de esos casos, en su versión inicial, se ajustan a los desarrollados en los capítulos 7 y 8 del mismo libro, donde Milgrom aporta contribuciones al diseño de subastas en el que plantea diferencias entre subastas de unidades múltiples (muchos productos en una sola transacción) y subastas de un solo producto (caso de mercancías, sector industrial específico, pesca, agricultura, infraestructura, espectro de radio, energía eléctrica, gas, agua, letras del tesoro y otras aplicaciones). Para mayor detalle del caso chileno se sugiere al lector revisar en la página 14 los antecedentes del tópico "Escenario actual en Chile", "Descripción del proceso licitatorio de contratación pública y las licitaciones" y el apartado "Canasta de productos y servicios de una licitación a 8 dígitos según el Código ONU".

Otro concepto importante es la denominada **maldición del ganador**, usado por primera vez por Wilson en 1969 y refiriéndose posteriormente a él en sus artículos de 1981,1982 y 2004, donde se concluye que los postores más optimistas suelen pujar muy bajo, y como el ganador es aquel postor con la estimación más optimista respecto del valor promedio del activo, al adjudicarse la licitación, el ganador se dará cuenta de que ha sido el más optimista en comparación con el resto que ha sido más pesimista, por ende pensará que ha ofertado muy por debajo del valor real y que dejará de ganar la diferencia que hay con respecto al segundo mejor postor. Para más información se sugiere revisar el Anexo 1: La maldición del ganador, el escenario sin asimetrías de información en la contratación pública y el valor explícito para los 4 actores del ecosistema de licitaciones. .

Por otro lado, el comité de competencia de la OCDE realizó en octubre de 2006 un debate (OECD, 2007a) sobre políticas en los mercados de licitaciones (*Competition in bidding markets*) donde se plantea el especial interés por el diseño de subastas en diferentes aspectos tales como favorecer la competencia, la cooperación entre empresas (consorcios), reducir la colusión y el abuso de posición dominante, evitar el *lobbie* (cabildeo) y mejorar la información pública disponible antes de ofertar (simetrías de la información pública). El documento final de esta reunión hace referencia al trabajo de Milgrom, Wilson, Klemperer y varios otros autores como elementos clave para el diseño de subastas.

[5] ChileCompra: Es la plataforma web chilena que gestiona todo el proceso licitatorio de contratación pública electrónica (*e-Procurement*), desde que se publica hasta que se adjudica,

No hay un total consenso en el uso de subastas para los contratos públicos, hay opiniones diversas. A modo de ejemplo, Chong et al. (2013) analizan empíricamente sobre 76 mil contratos públicos del Estado Francés, que en el sector construcción no está claro cuán eficiente resulta una subasta frente a una negociación; en cambio, para el sector privado, la negociación es notoriamente más eficiente y la más usada. En efecto, sostienen que las subastas en el segmento construcción e infraestructura presentan varias deficiencias de procedimientos e incluso algunas funcionan mal cuando los proyectos son complejos y es ahí donde las condiciones de negociación son más eficientes.

Lalive & Schmutzler (2011), por su parte, analizan la contratación pública de servicios ferroviarios en Alemania donde se da la oportunidad de evaluar los dos modos de contratación, por subastas y por negociación, bajo el mismo marco legal e institucional. Aquí los resultados indican algo distinto a lo señalado por Chong et al. (2013), ya que observan un aumento del 16% en la frecuencia del servicio que ha sido subastado respecto del contrato negociado, y en este mismo sentido el precio del contrato subastado es un 25% menor que en uno negociado.

1.3 Código estándar de productos UNSPSC y CPV

El código de productos y servicios estándar UNSPSC de la ONU, o su equivalente CPV de la Unión Europea, son recomendaciones OCDE a países miembros para codificar y estandarizar los productos, servicios y obras que son licitados en contratos públicos. Codificar facilita la identificación de los productos objeto de compraventa en la contratación pública para evitar errores de interpretación y confusión al definir el producto o servicio licitado. Ambos códigos UNSPSC y CPV son un estándar válido que además de codificar, facilita las estadísticas sectoriales por grupo de productos, la trazabilidad del gasto público y todo aquello que permita fortalecer la gobernanza de los países que hacen un esfuerzo para implementar estas recomendaciones.

1.3.1 UNSPSC: Código estándar de productos y servicios de Naciones Unidas

El UNSPSC es una metodología uniforme de codificación verificada y recomendada por la ONU. Su uso práctico y más frecuente es para clasificar productos y servicios que se licitan en un contrato público. Lo primero a saber es que es jerárquico, la manera más fácil de interpretar el UNSPSC es de lo más general a lo más particular.

Tabla 2: UNSPSC – Primer nivel por grupo de conceptos y letras
Fuente: Guía para codificación de bienes y servicio con el UNSPSC (Gobierno de Colombia, 2013)

Materia prima
Segmento 10 - 15
A - Materias primas, productos químicos, papel, combustible
Equipo industrial
Segmentos 20 - 27
B - Herramientas y equipos industriales
Componentes y Suministros
Segmentos 30 - 41
C - Suministros y componentes
D - Suministros y equipos de construcción, edificaciones y transportes
Productos de uso final
Segmentos 42 - 60
E - Productos farmacéuticos, suministros y equipos de ensayo, de laboratorio y médicos
F - Suministros y equipos de servicios, limpieza y comida
G - Suministros y equipos tecnológicos, de comunicaciones y de negocios
H - Suministros y equipos de defensa y seguridad
I - Suministros y equipos de consumo, domésticos y personales
Servicios
Segmentos 70 - 94
J – Servicios

En la Tabla 2, el primer nivel se divide en letras de la A hasta la J, que se agrupan en 6 conceptos: Materia Prima, Equipos Industriales, Componentes y Suministros, Productos de Uso Final y Servicios.

El UNSPSC codifica los productos a 8 dígitos porque cumple con el objetivo práctico de identificar con precisión el producto a licitar. Podemos verificar en la Tabla 3 que a mayor cantidad de dígitos se es más específico. El primer nivel usa 1 dígito y tiene 9 Secciones, el segundo nivel usa 2 dígitos y tiene 56 divisiones, y así sucesivamente, como se indica en la Tabla 3, hasta llegar al quinto nivel que usa 8 dígitos y tiene 53.317 productos y servicios.

Tabla 3: Número de dígitos por nivel del UNSPSC
Fuente: www.un.org[6], www.mercadopublico.cl. (*) corresponde a la cantidad real de grupos que tiene cada nivel, con referencia a las licitaciones realizadas en Chile durante los años 2018 a 2020

Nivel del código UNSPSC	Dígitos	Grupos		Especificidad	Sensibilidad
		Teórico UNSPSC	Reales (*) Chile		
Sección	1 díg. **1**xxxxxxx	9	9	Baja	Alta
División	2 díg. **12**xxxxxx	56	55	Baja	Alta
Clase	4 díg. **1234**xxxx	476	354	Media	Media
Subclase – Familia	6 díg. **123456**xx	4.294	1.930	Alta	Baja
Ítem (producto o servicio)	8 díg. **12345678**	53.317	13.596	Alta	Baja

En la Tabla 3 vemos que el grupo teórico a 8 dígitos tiene 53.317 productos y servicios; sin embargo, en la práctica las licitaciones en Chile sólo usan el 25% del UNSPSC; es decir, en los 36 meses analizados sólo se han usado 13.596 códigos a 8 dígitos. Esto no es un error, es sólo para dar cuenta de qué productos y servicios son requeridos por el Estado de Chile a través de licitaciones. Es más, dadas las características particulares de cada país y su normativa, estos pueden usar partes del UNSPSC que otros no, de ahí que la codificación jerárquica permite la comparación entre países en términos de características agregadas y no sólo por importes agregados de las transacciones.

Cuando el objetivo es realizar estudios o análisis de cualquier índole, es recomendable usar el UNSPSC a 1, 2 o 4 dígitos según sea la precisión que se desea lograr. En este sentido, se puede ver en la Tabla 3 en las columnas Especificidad / Sensibilidad, una mayor especificidad se corresponde con una menor sensibilidad de los modelos. No se recomienda realizar estudios sobre una codificación a 6 u 8 dígitos porque, según hemos comentado en el párrafo anterior, hay códigos que no se usan y estos pueden causar inconvenientes a la hora de implementar el modelo cuando se presentan casos nuevos que no han sido objeto de transacción anteriormente. En el anexo 3 se puede encontrar una versión reducida del UNSPSC a dos dígitos.

1.3.2 CPV: Vocabulario común de contratos públicos en la Unión Europea

El CPV (*Common Procurement Vocabulary*) que se resume en este apartado proviene del Manual del Vocabulario Común de Contratos Públicos (European Commission, 2008), que rige desde 2008 en su versión más actualizada y aprobada por el Consejo sobre los procedimientos de los contratos públicos del parlamento europeo[7].

> *El reglamento (CE) n° 213/2008 de la comisión de las comunidades europeas de 28 de noviembre de 2007, modifica el Reglamento (CE) n° 2195/2002 del Parlamento Europeo y del Consejo, por el que se aprueba el Vocabulario común*

[6] Clasificador ONU Productos y servicios https://www.un.org/es/procurement/info/unccs.shtml
[7] Diario Oficial de la Unión Europea, reglamento (CE) No 213/2008 de la comisión de 28 de noviembre de 2007

de contratos públicos (CPV), y las Directivas 2004/17/CE y 2004/18/CE del Parlamento Europeo y del Consejo sobre los procedimientos de los contratos públicos, en lo referente a la revisión del CPV.

La revisión final del CPV y sus recomendaciones actualizadas al 2017 fue encargado por la comisión europea en la dirección general de asuntos de mercado interno, industria, emprendimiento y pymes (European Commission, 2017 *Final report, Revision of CPV* "*Consultancy Services for Common Procurement Vocabulary expert group*").

El objetivo del CPV es normalizar los términos que utilizan los poderes y entidades adjudicadoras al describir el objeto de los contratos. Para ello se introduce en la contratación pública un solo sistema de clasificación y se ofrece así un instrumento útil a los usuarios potenciales que intervienen en los procedimientos de adjudicación de contratos. El uso de unos códigos normalizados facilita la aplicación de las normas de publicidad y hace más sencillo el acceso a la información.

En la actual versión del CPV, el término "producto" hace referencia tanto a mercancías como servicios, que se hereda de dos clasificadores internacionales: una es la *Clasificación Central de Productos* (CCP) elaborada por la ONU para supervisar el comercio mundial y favorecer un marco de comparación internacional de estadísticas referentes a mercancías, servicios y activos; la otra es la *Clasificación Industrial Internacional Uniforme* (CIIU Rev. 3), nomenclatura para clasificar actividad económica. Su equivalente en Europa es la *Clasificación Europea de Actividades Económicas* (NACE Rev. 1 de 1990) que ofrece mayor detalle en supervisión de las economías europeas; ésta finalmente ha dado origen a la versión de la *Clasificación de Productos por Actividades* (CPA).

En el siguiente diagrama se muestran las distintas versiones de los clasificadores de productos y actividad económica que dan origen a la última versión del CPV. Una versión reducida a dos dígitos del CPV está disponible en el Anexo 4: Divisiones del CPV de 2008.

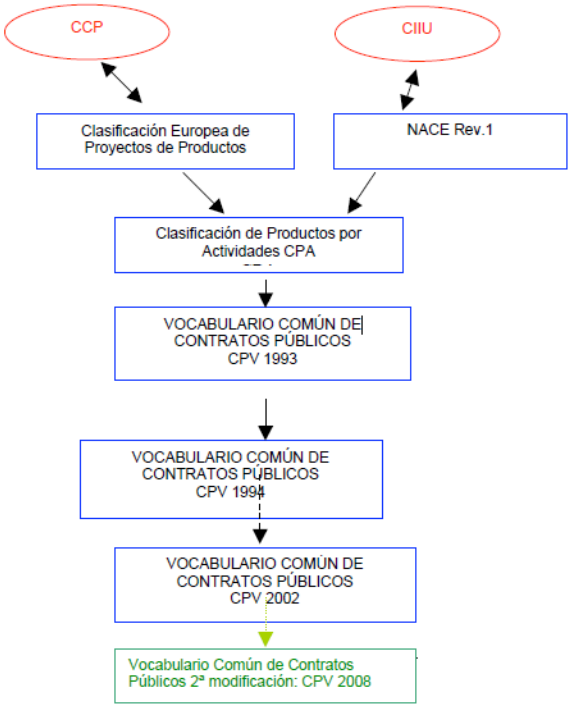


Figura 1: Revisión de versiones del vocabulario común de contratos públicos (CPV)

1.4 Canasta de productos y servicios de una licitación a 8 dígitos UNSPSC

Uno de los objetivos prácticos de la licitación es identificar con precisión cada elemento de la canasta para evitar errores de interpretación respecto a qué producto o servicio se está licitando realmente. Por lo tanto, el o los productos y servicios de la canasta deben estar codificados a 8 dígitos con el código jerárquico internacional UNSPSC. A continuación, se revisan y describen algunas características de las licitaciones y del código UNSPSC que dan cuenta de la importancia de cada uno y cómo se relacionan para proporcionar una herramienta útil de gestión y de análisis.

1.4.1 Características de una licitación

Para caracterizar una licitación utilizaremos los conceptos de Tipo de Adjudicación, código UNSPSC y Canasta:

• El Tipo de Adjudicación es un atributo que toma un valor en la etapa de adjudicación. A partir del resultado y evaluación de las propuestas y pujas de todos postores, cada licitación toma uno de los siguientes estados: {Desierta, Simple, Parcial o Múltiple}.

Desierta: en ocasiones, no se adjudica ninguno de los productos o servicios de la canasta, con lo cual la licitación es declarada desierta, debido a que ningún postor realiza una oferta, o las ofertas realizadas no cumplen las condiciones técnicas y administrativas requeridas por el mandante.

Adjudicación simple: corresponde a la asignación del 100% del producto o servicio licitado a un solo vendedor (postor o proveedor).

Adjudicación parcial: corresponde a la asignación de una porción (parte o fracción) del producto licitado a un solo proveedor. En este caso la parte o fracción faltante no es adjudicada.

Asignación múltiple: consiste en asignar el mismo producto a varios oferentes de modo que se reparten todos los recursos disponibles entre ellos.

• El Código UNSPSC:

Según se ha revisado en el apartado “Código estándar de productos UNSPSC y CPV” (pág. 20), es un código estándar de productos y servicios de las Naciones Unidas a 8 dígitos (*The United Nations Standard Products and Services Code*® - UNSPSC). Es una metodología uniforme de clasificación utilizada para clasificar productos y servicios fundamentada en una estructura lógica y jerárquica.

• La canasta de productos y servicios de una licitación:

En marketing y fidelización, Agrawal & Srikant (1994) definen la canasta de compras (*basket-market*) como el conjunto de productos reunidos en una sola compra (transacción en un instante de tiempo). En el ámbito de las subastas y contratos públicos, la canasta toma el nombre de *licitación* y conserva todas las propiedades que dan origen al análisis de canasta de mercado (*market basket analysis*, MBA).

1.4.2 Especificidad y sensibilidad del código UNSPSC

La siguiente tabla muestra la cantidad de grupos que hay por nivel del clasificador jerárquico UNSPSC a 1, 2, 4, 6 y 8 dígitos, seguido del nivel de especificidad, sensibilidad y objetivo. Los grupos reales corresponden al caso de Chile entre los años 2018 y 2020.

Tabla 4: Código estándar de productos y servicios UNSPSC.
(*) Los grupos reales corresponden al caso de Chile entre los años 2018 y 2020.
Fuente: www.un.org[8], www.mercadopublico.cl

Dígitos	Nivel del clasificador UNSPSC	Grupos Teóricos UNSPSC	Grupos Reales (*) Chile	Especificidad	Sensibilidad	Objetivo
1 díg. 1xxxxxxx	Sección	9	9	Baja	Alta	Análisis y estudios (Datos agregados)
2 díg. 12xxxxxx	División	56	55	Baja	Alta	
4 díg. 1234xxxx	Clase	476	354	Media	Media	
6 díg. 123456xx	Subclase - Familia	4.294	1.930	Alta	Baja	Precisión (evitar errores)
8 díg. 12345678	Ítem (producto o servicio)	53.317	13.596	Alta	Baja	

Según se indica en la tabla anterior, con el código UNSPSC a 8 dígitos el nivel de precisión es muy detallado, de modo que se logra una alta especificidad (mucho detalle) y baja sensibilidad (modelos poco sensibles a la tendencia o patrón).

Una de las condiciones básicas en una licitación es identificar con precisión los productos y servicios para evitar errores de interpretación. El código UNSPSC admite hasta 53.317 posibles productos, permite gran detalle en la clasificación. Sin embargo, se puede constatar que en la práctica, entre 2018 y 2020, en Chile sólo han usado 13.596 códigos, lo cual representa un 25% del total, para identificar sus productos[9].

Cuando el objetivo es analizar y estudiar un conjunto de licitaciones se recomienda agrupar los productos y servicios a 1, 2 o 4 dígitos del UNSPSC, de modo que se pueda trabajar con información agregada que facilite la investigación y evite los casos particulares.

Para el caso de estudio de este TFM, vamos a modelar y caracterizar una canasta a 2 dígitos del código UNSPSC, de modo que se identifican 55 grupos reales de productos y servicios con suficiente nivel de agregación para hacer modelos sensibles a los patrones y tendencias de los datos agregados. Con ello se pretende cumplir los objetivos planteados en el clasificador de licitaciones (ver fila “2 dígitos” en la Tabla 4).

1.4.3 Canasta mono producto (licitación de un producto o servicio)

En el caso de Chile (años 2018 a 2020), el 86% de las licitaciones consta de una canasta con sólo un producto o servicio. Éste se acompaña de especificaciones técnicas y administrativas que lo describen, y debe estar codificado a 8 dígitos. La Tabla 5 identifica, a modo de ejemplo real, tres licitaciones de un producto cada una: la primera columna muestra la referencia de la licitación o identificador de la canasta, la segunda columna indica el producto codificado a 8 dígitos, seguidos del nombre del producto y una descripción genérica.

La codificación a 8 dígitos aporta una identificación en detalle del producto o servicio. Esto nos permite verificar en Tabla 5 que la segunda licitación, ref. 5488-12-LP21, presenta un error en la codificación, ya que el código UNSPSC 43212110 ha sido mal asignado porque corresponde a la “Adquisición de impresoras” (es un producto de uso final). Se puede verificar

[8] Clasificador ONU Productos y servicios https://www.un.org/es/procurement/info/unccs.shtml
[9] Fuente Mercado Público, Chile, base de datos de licitaciones (2018 a 2020).

a partir de la descripción que en rigor se solicita un servicio de “arriendo de impresoras” y lo correcto hubiese sido considerar el Código UNSPSC 80161800 para el producto: “Servicios de alquiler o arrendamiento de equipo de oficina”.

Tabla 5: Ejemplo de canasta mono producto, tres licitaciones codificadas a 8 dígitos

Licitación ID (Canasta)	**Código UNSPSC 8 dígitos**	**Nombre Producto (UNSPSC)**	**Descripción**
3413-20-L121	52131702	Barras de cortinas	Kits soporte para cortinas de madera
5488-12-LP21	43212110	Impresoras multifunción o multifuncionales	Arriendo de impresoras; valor mensual; ofertar sumatoria total formulario n°4 en pesos.
2920-50-L121	80141607	Producción de eventos	Servicio de producción de evento

1.4.4 Canasta multi producto (licitación de dos o más productos y servicios)

En el caso de Chile (años 2018 a 2020), el 14% de las licitaciones son canastas que tienen más de un producto y/o servicio (multi producto). Es habitual encontrar dos casos de canastas multi producto como las que se comentan a continuación.

El primer caso es aquella canasta que identifica varios productos con el mismo código a 8 dígitos. La diferencia puede estar en las especificaciones técnicas y/o administrativas que el proveedor debe cumplir; por ejemplo, “*51102710 Antisépticos basados en alcohol o acetona*” puede variar en cantidad, volumen, envase, u otra especificación técnica. Respecto a las especificaciones administrativas, se puede requerir que el suministro sea mensual, por única vez, en plazo indefinido, sin quiebre de stock, stock de seguridad del 5%, etc.

La siguiente tabla muestra el ejemplo de una licitación con tres productos codificados con el mismo código UNSPSC, pero difieren en el formato o presentación (especificación técnica o administrativa).

Tabla 6: Ejemplo de canasta multi producto (1 código a 8 dígitos UNSPSC)

Licitación ID (Canasta)	**Código UNSPSC 8 dígitos**	**Nombre Producto**	**Descripción**
4940-2-LQ22	51102710	Antisépticos basados en alcohol o acetona	Alcohol Etílico 70% FC250 ML D
	51102710	Antisépticos basados en alcohol o acetona	Alcohol Etílico 96% FC X LT C
	51102710	Antisépticos basados en alcohol o acetona	Alcohol Gel 1000 ML

Otro caso muy frecuente en una licitación multi producto, es una canasta de varios productos que en su mayoría tienen distinto código UNSPSC a 8 dígitos. En la Tabla 7 se muestra una licitación de ejemplo real, que consta de una canasta de 6 productos distintos cuando se han codificado a 1, 2, 6 y 8 dígitos.

Tabla 7: Ejemplo de canasta multi producto (varios códigos a 8 dígitos UNSPSC)

Código UNSPSC				**Nombre Producto**	**Descripción**
1 dígito Sección	**2 dígitos** División	**6 dígitos** Familia	**8 dígitos** Ítem		
5	**51**	**512412**	**51241226**	Preparados tópicos de urea	Urea 30% en Crema Base
		511715	**51171504**	Antiácidos de bicarbonato de sodio	Sodio Bicarbonato 1 Gr Papelillos
		511716	**51171631**	Preparado laxante de polietilenglicol	Polietilenglicol 3350 SO 17 Gr
			51171630	Aceite mineral	Azufrada 6% CJ X 50 Gr
		511027	**51102702**	Agua destilada para irrigación	Agua Oxigenada 10 Vol Fco 110
			51102710	Antisépticos basados en alcohol o acetona	Alcohol Absoluto Farmacopea

El código UNSPSC permite relacionar qué concepto corresponde a un determinado nivel de agrupación. Por ejemplo, si la licitación anterior se codifica a un dígito, tendrá un grupo con el código “5” y su concepto será “Productos de uso final”. A continuación, se listan todos los casos de este ejemplo cuando se agrupan a 1, 2, 4, 6 y 8 dígitos:

- A nivel de sección (un dígito) - un grupo,
  - 5: Productos de uso final
- A nivel de división (dos dígitos) - un grupo,
  - 51: Medicamentos y productos farmacéuticos
- A nivel de clase (cuatro dígitos) - tres grupos,
  - 5124: Fármacos que afectan a los oídos, los ojos, la nariz y la piel
  - 5117: Medicamentos que afectan al sistema gastrointestinal
  - 5110: Medicamentos antiinfecciosos
- A nivel de familia de productos (seis dígitos) - cuatro grupos,
  - 512412: Agentes dermatológicos
  - 511715: Antiácidos y antiflatulentos
  - 511716: Laxantes
  - 511027: Antisépticos
- A nivel de ítem o producto (ocho dígitos) - 6 productos (no hay grupos),
  - 51241226: Preparados tópicos de urea
  - 51171504: Antiácidos de bicarbonato de sodio
  - 51171631: Preparado laxante de polietilenglicol
  - 51171630: Aceite mineral
  - 51102702: Agua destilada para irrigación
  - 51102710: Antisépticos basados en alcohol o acetona

Con este ejemplo comprendemos por qué el UNSPSC es jerárquico y tiene sentido plantear su utilidad. Si el objetivo es realizar un análisis o estudio, éste debe hacerse con datos agregados que agrupan productos y servicios a 1, 2, o 4 dígitos. En cambio, cuando los productos se codifican a 6 u 8 dígitos, el objetivo es identificar el producto en detalle para evitar errores operativos en la licitación.

1.4.5 Representación vectorial de una canasta

Al realizar estudios y análisis de las licitaciones, los productos y servicios de la canasta deben estar codificados y agrupados a 1, 2 o 4 dígitos del UNSPSC o CPV, con el fin de manipular los datos y aplicar modelos multivariantes. La representación vectorial de una canasta depende del nivel de agrupación que se haya escogido para el análisis. Por ejemplo, si realizamos el análisis a nivel de Sección (1 dígito) el vector tendrá 9 valores, dando lugar a una matriz con 9 variables binarias (ver Tabla 8). Si utilizamos un código a nivel de División (2 dígitos) el vector tendrá 55 valores (ver Tabla 9). Sí usamos un código a nivel de Clase (4 dígitos el vector tendrá 354 variable binarias.

En la Tabla 8 se muestra un ejemplo con 3 licitaciones codificadas a un dígito. De esta forma el vector se construye con la información de la licitación, asignando un 1 al descriptor que tenga asociado un producto y un 0 en caso contrario.

Tabla 8: Ejemplo de canasta mono producto como vector a un dígito UNSPSC

Licitación ID (Canasta)	**UNSPSC 8 dígitos**	**UNSPSC 1 dígito**	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
3413-20-L121	**5**2131702	5	0	0	0	0	1	0	0	0	0
5488-12-LP21	**4**3212110	4	0	0	0	1	0	0	0	0	0
2920-50-L121	**8**0141607	8	0	0	0	0	0	0	0	1	0

En este ejemplo, vemos que la primera licitación, con ref. 3413-20-L121, si se agrupa a un dígito asignaremos un 1 en el descriptor asociado a “X5” de un total de 9 posibles y 0 en las otras 8 posiciones. Pero en caso de codificar a dos dígitos, esta misma licitación asignará un 1 en el descriptor asociado al código “X52” y los 54 restantes quedarán en 0 (cero).

En la Tabla 9 se presenta el mismo ejemplo, codificado a dos dígitos, aunque dado el ancho del vector se omiten algunas variables (descriptores).

Tabla 9: Ejemplo de canasta mono producto como vector a dos dígitos UNSPSC

Licitación ID (Canasta)	**UNSPSC 8 dígitos**	**UNSPSC 2 dígito**	X_{10}	X_{11}	---	X_{42}	X_{43}	X_{44}	---	X_{51}	X_{52}	X_{53}	---	X_{78}	X_{80}	X_{81}	---	X_{92}	X_{93}	X_{94}
3413-20-L121	**52131702**	52	0	0	…	0	0	0	…	0	1	0	…	0	0	0	…	0	0	0
5488-12-LP21	**43212110**	43	0	0	…	0	1	0	…	0	0	0	…	0	0	0	…	0	0	0
2920-50-L121	**80141607**	80	0	0	…	0	0	0	…	0	0	0	…	0	1	0	…	0	0	0

En definitiva, en el caso de licitaciones y otros del tipo canasta, la representación vectorial asigna muchos “ceros” y pocos “unos”. Por lo tanto, la matriz resultante será de baja densidad o “rala”, término que se aplica cuando los componentes, partes o elementos están más separados de lo regular en su clase. Éste es uno de los inconvenientes que deberá resolver la técnica seleccionada para el análisis de canasta y diseño del clasificador de licitaciones.

2 OBJETIVOS

A partir de este contexto de mercado, el escenario actual de contratos públicos del Estado de Chile y la de otros países que se rigen por las recomendaciones OCDE, este TFM se propone cumplir con un objetivo general y varios objetivos específicos.

Objetivo General:

Diseñar, construir y validar un clasificador de licitaciones multivariante que sea eficiente y robusto para la contratación pública del Estado de Chile, capaz de aportar valor a todos los actores del ecosistema de licitaciones. Este Clasificador de Licitaciones debe ser una herramienta genérica, sencilla de poner en práctica y cuyo resultado sea fácil de interpretar para un analista de datos. Además, debe facilitar la solución de los inconvenientes de manipulación de datos de manera práctica y eficiente, y se requiere también que aporte información segmentada a todos los agentes sin pérdida de información.

Objetivos específicos:

1. Relacionar, medir y comparar la eficiencia de tres técnicas de análisis exploratorio multivariante en la etapa de análisis de canasta de licitaciones MBA (reglas de asociación AR, análisis de correspondencia múltiple MCA y análisis de componentes principales PCA).
2. Proponer la técnica más eficiente para la etapa de análisis de canasta de licitaciones MBA, cuando todas las variables son categóricas.
3. Con el fin de validar la incorporación de nuevas herramientas, se busca aportar elementos teóricos que permitan cuestionar o promover las características y potencialidades del MCA como técnica multivariante para el análisis de canasta de mercado MBA.

4. Describir el marco regulatorio y recomendaciones que rigen las compras públicas a través del mecanismo de licitaciones vigente, su contexto de mercado y sus equivalencias en los países miembros OCDE.
5. Describir el marco de recomendaciones internacionales de la EUROSTAT que aportan a la estandarización y validación de las estadísticas europeas basadas en Registros Administrativos de compras públicas.
6. Identificar los supuestos de la teoría de subastas más relevantes que se consideran en la contratación pública que sean capaces de aportar valor a los agentes o actores del ecosistema de licitaciones a través del clasificador de licitaciones.
7. Comprender e identificar el potencial de los códigos de productos del UNSPSC de la ONU o su equivalente CPV de la Unión Europea, para codificar y estandarizar los productos y servicios licitados en contratos públicos.
8. Identificar beneficios y oportunidades de negocio que aporta el clasificador a los cuatro actores del ecosistema de licitaciones.

3 MATERIALES Y MÉTODOS

El punto de partida de este TFM es el marco teórico visto en la introducción y antecedentes, que plantea un escenario basado en las recomendaciones internacionales de a) la OCDE en materia de "*contratación pública"*, b) EUROSTAT para *"estadísticas sobre registros administrativos"*; c) el "*código UNSPSC y CPV"* para la codificación de productos y servicios para facilitar la "*representación vectorial de la canasta"*; finalmente, d) la puesta en práctica de la "*teoría de subastas"* de Milgrom, Wilson, Klemperer y otros autores que plantean un escenario competitivo que puede conducir a estimar el valor de los activos segmentados.

La base de datos utilizada contiene 330.000 licitaciones del Estado de Chile registradas en 36 meses consecutivos correspondientes a los años 2018, 2019 y 2020. Esta información se ha obtenido de la plataforma de contratación electrónica (*e-Procurement*) que sigue las recomendaciones y estándares internacionales en materia de contratación pública usadas por todos los países OCDE. En el caso chileno, la canasta de productos y servicios licitados es codificada con según el código UNSPSC.

La manipulación y preprocesamiento de los datos se ha realizado siguiendo la metodología KDD (descrita en la pág. 29), se han utilizado herramientas SQLserver, Excel, TextPad8 y el software estadístico R-Studio para modelar el problema de las licitaciones a partir del análisis de canasta con las técnicas RA, MCA y PCA, segmentación con HCPC y clasificación con SVM, redes neuronales, árboles de decisión y *random forest*.

Considerando el marco teórico y materiales disponibles, lo que se busca es un conjunto de métodos multivariantes que permitan modelizar el problema de la diversidad de licitaciones a partir de un análisis de canasta de mercado MBA, que aporte una respuesta segmentada y genere valor añadido a todos los agentes y actores del ecosistema de licitaciones.

En este trabajo se revisan en un contexto práctico del análisis de canasta MBA tres técnicas para comparar sus ventajas e inconvenientes. De este modo se podrá determinar, en el caso de licitaciones, cuál de las tres es la más apropiada para construir un clasificador de licitaciones, cuando las variables de la canasta son categóricas y han sido agrupadas a dos dígitos del código UNSPSC:

- **RA:** Association Rules,

- **PCA:** Principal Component Analysis
- **MCA:** Multiple Correspondence Analysis

En este sentido, se busca reforzar y dar un contexto al hecho de que el análisis de correspondencias múltiple MCA es una técnica estadística multivariante muy útil para quienes trabajan con datos categóricos de varios niveles (Greenacre, 2008). Por su parte, Abdi & Valentin (2007) recomiendan el MCA para analizar un set de observaciones de variables de selección múltiple con dos niveles (“0”, “1”). Aunque en su origen el MCA fue desarrollado hace más de 50 años por Jean-Paul Benzécri para trabajar en el mundo de las ciencias sociales, ahora su ámbito de aplicación es más amplio y se ajusta muy bien a procesos industriales, agrícolas, logística, “marketing y fidelización”, encuestas, servicios y medicina, entre varias otras disciplinas. En nuestro caso aplicado a licitaciones resulta fundamental para el análisis de canasta de mercado MBA y para construir el clasificador de licitaciones.

Con el fin de validar la incorporación de nuevas herramientas, también buscamos dejar los elementos teóricos que permitan cuestionar o promover las características y potencialidades del MCA como técnica multivariante para el análisis de canasta. La herramienta tradicional e indiscutida por muchos años ha sido las reglas de asociación AR, que ha aportado con sencillez al conocimiento e información para el tomador de decisiones, principalmente en el ámbito del comercio minorista donde tuvo su origen.

3.1 Metodologías de preprocesamiento KDD vs CRISP-DM

Antes de aplicar cualquier modelo o técnica, conviene plantearse una de las dos metodologías para el preprocesamiento y manipulación de datos más usadas en minería de datos es KDD o CRISP-DM. Fayyad, Piatetsky-Shapiro y Smyth plantean la metodología KDD (*knowledge discovery in databases*) (Fayyad et al., 1996) como una herramienta que permite extraer conocimiento en las bases de datos para descubrir patrones en forma de reglas o funciones. El KDD sigue una secuencia de pasos para preprocesar los datos antes de desarrollar los modelos (ver Figura 2). Esta metodología tiene más sentido cuando se aplican técnicas de *machine learning* en un contexto de experimentación, laboratorio o de investigación que dan acceso a una muestra de datos para un estudio o para validar un modelo. Cuando el flujo de información es constante y se ha sistematizado el modelo como parte del proceso productivo de información, el KDD se vuelve similar al caso de CRISP-DM (ver Figura 3).

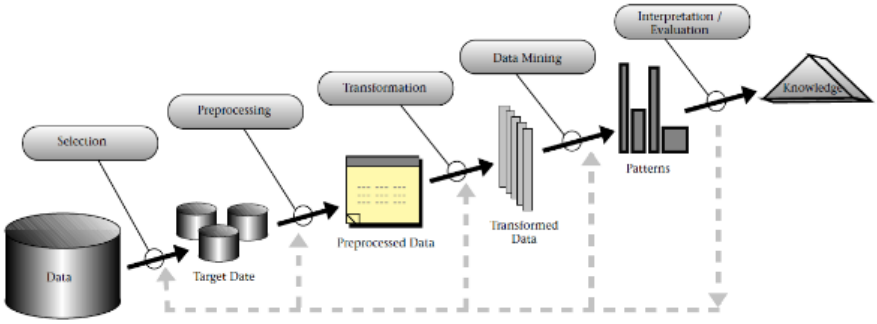


Figura 2: Diagrama KDD: descripción general de los pasos que componen el proceso KDD para extraer conocimiento de las bases de datos.

Otra metodología muy usada es la CRISP-DM (del inglés, *Cross Industry Standard Process for Data Mining*). Se trata de un modelo estándar creado en 1996 por un consorcio de

empresas y financiado por la Unión Europea, que describe el proceso que utilizan los expertos en minería de datos, dividido en 6 fases: 1) Comprensión del negocio, 2) Comprensión de los datos, 3) Preparación de los datos, 4) Fase de Modelado, 5) Evaluación y 6) Implantación del modelo, más una componente cíclica que permite aprender de la experiencia tras la aplicación del modelo (ver Figura 3)[10]. Esta metodología es adecuada cuando el contexto es para grandes proyectos que involucran muchos sub-proyectos que se relacionan y comparten infraestructura y recursos en común. En este contexto, sus usuarios se enfocan en modelar herramientas de minería de datos y que requieren acceder a la misma información sobre bases de datos transaccionales corporativas.

Figura 3: Diagrama de proceso que muestra la relación entre las diferentes fases de CRISP-DM - Fuente: Wikipedia.

3.2 Análisis de canasta de mercado MBA

El análisis de canasta de mercado (en inglés, *Market Basket Analysis*, MBA) es un concepto desarrollado en la década de los años 1990 para aportar información al tomador de decisiones en el ámbito del comercio minorista (*Retail*, supermercados, grandes tiendas) y se valida como una herramienta efectiva para una vasta lista de disciplinas y aplicaciones.

La técnica de minería de datos más utilizada en un MBA son las reglas de asociación AR (*mining association rules*) con su eficiente y eficaz algoritmo *apriori*, que se valida por su excelencia y por la gran cantidad de artículos científicos que respaldan su uso práctico y efectividad en análisis de canasta MBA. En este apartado se centra en la revisión del trabajo de Kaur & Kang (2016) que compara varios algoritmos y sus modificaciones para mejorar el rendimiento del AR analizando varios casos y aplicaciones MBA. A continuación, se presentan dos definiciones básicas:

Canasta de compras (*basket market*): varios autores coinciden en la siguiente definición: "es *el conjunto de productos reunidos en una sola compra (transacción en un instante*

[10] CRISP_DM: Wikipedia, https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

de tiempo)" (Agrawal, et al., 1993a; Agrawal & Srikant, 1994; Bayardo & Agrawal, 1999; Bradlow et al., 2017; Dorn et al., 2008; Kaur & Kang, 2016; Pinho L., 2010)

Análisis de canasta de mercado (MBA *basket market analysis*): *"investigación aplicada sobre las compras realizadas por un cliente o tipo de cliente (segmento) durante un periodo de tiempo"* (Agrawal & Srikant, 1994; Kaur & Kang, 2016).

Los autores Agrawal et al. (1993a) son algunos de los primeros en introducir el concepto de minería de reglas de asociación AR y su algoritmo *apriori*, que da origen a la técnica, tal como se conoce hoy. El origen y motivación principal surge de resolver los problemas que tienen los directores de supermercados, que a menudo deben lidiar con una gran cantidad de productos y tomar diferentes decisiones, por ejemplo: ¿qué productos colocar en venta?; ¿cómo diseñar los cupones de ventas?; ¿cómo colocar la mercadería en los estantes para maximizar las ventas?, etc. El AR se conoce y se referencia usualmente como el problema de la canasta de mercado (*basket-market*) (Agrawal, et al., 1993a) a partir del cual se publican diversos artículos relacionados con su teoría, aplicaciones y mejoras.

Otra aportación de Agrawal, et al. (1993b) es el concepto de *database mining* como plataforma de trabajo para aplicar técnicas de aprendizaje automático (*machine learning*), entre ellas el AR, y obtener patrones y/o reglas dentro de datos transaccionales masivos, que pueden abordarse en problemas de clasificación, asociación y secuenciación.

Por su parte, otro estudio (Agrawal & Srikant, 1994), a través de las reglas de asociación AR, desarrolla uno de los problemas aplicados más importantes de la minería de datos que se mantiene vigente en el tiempo y que se conoce con el nombre de "Análisis de Canasta de Mercado" MBA. Es usual encontrar artículos basados en el modelo MBA, porque es el más usado y tradicional.

La gran ventaja del análisis de canasta MBA con reglas de asociación AR, es que puede procesar muy bien varios miles de productos con algoritmos optimizados que obtienen la combinatoria (*NP-Hard*) de todos los productos que se venden juntos en una transacción (Bayardo & Agrawal, 1999) y que puede relacionarse con el perfil del cliente (Chang et al., 2007).

Determinar el perfil del cliente es otro propósito del MBA, ya que permite detectar patrones de compra y caracterizar el prototipo de cliente, que se infiere del comportamiento de compra anterior de los clientes por medio de análisis de agrupamiento (segmentación) y análisis de reglas de asociación (Chang et al., 2007).

Una consecuencia natural, al identificar el perfil del cliente, deriva en el concepto de marketing tradicional para desarrollar estrategias segmentadas y aplicadas a la promoción de productos que se venden juntos (Griva et al., 2018). En paralelo, se desarrolla el concepto de marketing y fidelización (*loyalty*), hoy conocido como marketing relacional (*relationship marketing*), que a partir de las estrategias segmentadas y el perfil del cliente reconoce otras 7 variables de percepción del cliente que deben medirse, por medio de formularios y entrevistas, para determinar un modelo conceptual de lealtad del cliente (*conceptual model of customer loyalty*): (a) imagen, (b) percepción de calidad, (c), valor (d) satisfacción, (e) emoción, (f) confianza y (g) compromiso (Gaur et al., 2013).

En el marketing aplicado (Agrawal, et al. 1993b) sobre base de datos (*database marketing*), es usual el análisis de la canasta MBA tradicional sobre un "*set de datos estáticos"* para detectar patrones de compra del cliente y ayudar al analista de marketing a comprender el comportamiento de los clientes. Hay varias técnicas y algoritmos disponibles para realizar la

minería de Reglas de Asociación, cuyo objetivo principal es establecer la relación entre los artículos que se venden juntos en la tienda minorista o *retailer* (Kaur & Kang, 2016). El *Retailer* es el comercio minorista que vende directamente a público "*business to customer"* (B2C) ya sea en tienda o por internet, a diferencia del "*busines to business"* (B2B) que vende de mayoristas a intermediarios o distribuidores esencialmente desde bodega.

El trabajo de Kaur & Kang (2016) revisa varios estudios previos sobre las reglas de asociación AR como técnica aplicada al MBA, que es también conocido como el aprendizaje de reglas de asociación o análisis de afinidad (AR *Association Rules*). Es una técnica de minería de datos que se puede utilizar en varios campos, como el marketing, bioinformática, educación, ciencias nucleares, etc. Esta investigación de Kaur & Kang (2016) relaciona las versiones de AR comparando las propuestas de varios autores que mejoran la capacidad de respuesta del algoritmo y su implementación. Además, se plantean desarrollos y adaptaciones con nuevos enfoques de la Serie Apriori, el algoritmo AIS, el algoritmo Apriori, el algoritmo FP-Tree (*Frequent Pattern-Tree Algorithm*), y el algoritmo RARM (*Rapid Association Rule Mining*). Como conclusión, los autores de este artículo señalan que el algoritmo Apriori original sigue siendo el de mejor rendimiento comparativo entre todos los algoritmos revisados y que además es el más sencillo de implementar y de interpretar.

Un ejemplo típico de minería de reglas de asociación aplicado al análisis de canasta MBA, puede ser algo parecido a lo siguiente: "El 30% de las compras que contienen cerveza y patatas fritas, también contienen cacahuete salado" y "el 2% de todas las compras del supermercado contienen los tres productos"; esta descripción corresponde a una regla, entre muchas otras, que cumplen la condición de soporte y confianza .

Esta afirmación se puede expresar como una regla:
Si A entonces B (*c, s*)
Donde:
- A es el conjunto de productos (atributos) de la condición de la regla, denominado Antecedente.
- B es el conjunto de productos (atributos) de la conclusión de la regla, denominado Consecuente.
- c (30%) se denomina confianza de la regla.
- s (2%) se denomina soporte de la regla.

Este ejemplo de *Mining Association Rule* (AR) aplicado a MBA, está extraído de la guía docente del curso Minería de Datos impartido en la Universidad de Santiago de Chile, por el profesor Max Chacón Pacheco (2004)

La validez de los modelos AR se basan en una fuerte componente estadística para justificar la calidad de sus resultados. Dorn et al. (2008) y Pinho (2010) revisan el trabajo de varios autores que promueven indicadores y conceptos habitualmente usados en AR:

- Hay dos parámetros de entrada del modelo, "Soporte" y "Confianza", que ayudan a seleccionar reglas válidas.
- Es necesario evaluar algunos indicadores para asegurar la buena calidad del conjunto de reglas seleccionadas.
- De la bibliografía recopilada por los autores, se citan los siguientes indicadores de calidad: *Lift*, *Gini*, Convicción, *Laplace*, Ganancia, Entropía, Chi-Cuadrado y la Métrica de Piatetsky-Shapiro. Algunos de estos indicadores son redundantes dependiendo del ámbito de aplicación, de modo que deben elegirse los más pertinentes. De esta forma, los indicadores logran acotar y podar una selección de patrones que tienen mayor probabilidad de ser reglas (*minimal support or minimal confidence*).

La tesis doctoral de Pinho (2010) explica y define dos parámetros iniciales (soporte y confianza) para reducir el número de reglas de alta combinatoria y analiza estadísticamente cuatro medidas de interés para obtener reglas de asociación válidas y de calidad.

- *Support* (soporte): es un parámetro que contabiliza el número de transacciones en las cuales los ítems presentes en una regla ocurren simultáneamente, en relación con el número total de transacciones.
- *Confidence* (confianza): es un parámetro que indica el porcentaje de transacciones que contienen conjuntamente el antecedente y el consecuente en relación al número de transacciones que contienen la parte antecedente.
- *Lift* (sustentación): en una regla de asociación A → B, este parámetro representa en qué grado "B" tiende a ser frecuente cuando "A" ocurre, o viceversa (bidireccional).
- *Conviction* (convicción): es una medida que evalúa en qué grado el antecedente **influye** en la ocurrencia del consecuente. A diferencia del *lift*, la convicción es una medida unidireccional, o sea, el resultado de *conviction* (A → B) será diferente del de *conviction* (B → A).
- Test χ^2: es un test estadístico que compara las frecuencias obtenidas con las esperadas. Dicho test suele ser utilizado para evaluar el nivel de correlación o asociación del término antecedente con el término consecuente de una regla.
- *Coverage* (cobertura): es el soporte de la parte izquierda de la regla (antecedente). Se interpreta como la frecuencia con la que el antecedente aparece en el conjunto de transacciones.

Para dar un contexto práctico e intuitivo, en el Anexo 2: "Análisis de Canasta MBA aplicado a licitaciones", se calculan las reglas de asociación AR de un set de licitaciones a partir del procedimiento usado por Gil Martínez (2020), con el cual se contrastan empíricamente las ventajas y desventajas de la técnica AR y su algoritmo Apriori. El procedimiento de Cristina Gil usa la librería "*arules*" de RStudio, que despliega una selección de reglas filtradas considerando dos parámetros (soporte y confianza) y tres indicadores de calidad (*coverage*, *lift* y el test de Fischer).

3.2.1 Reglas de asociación AR vs. PCA y MCA

En este TFM se emplea MCA y PCA como alternativa al AR en el análisis de canasta MBA. AR fue desarrollado con el objetivo de evaluar todas las combinaciones de productos que se venden juntos en una transacción y seleccionar un conjunto reducido de reglas de buena calidad. Sin embargo, AR se aleja del propósito de un análisis exploratorio amplio que considere toda la información y características de los datos, lo cual se puede lograrse con PCA o MCA.

En general, el MBA con AR puede presentar dificultades para hacer frente a la alta correlación de variables, valores faltantes y la falta de capacidad para detectar cambios en la tendencia de patrones. Estas condiciones se pueden mitigar o resolver con las propiedades multivariantes del MCA o PCA.

Por otro lado, independientemente de la técnica de minería de datos a utilizar, Yin & Jaynak (2015) identifican cinco prioridades, denominadas las cinco V's, (Figura 4) que deben trabajarse en un contexto *big data* para mantener una solución equilibrada que facilite el análisis y extracción de información, reglas y conocimiento (Yin & Kaynak, 2015):

Volumen: a mayor cantidad de datos se requiere una mayor inversión en instalaciones, seguridad, almacenamiento, procesamiento y automatización.

Veracidad: en caso de datos obtenidos a partir de sensores, estos son necesarios para validar la calidad de los datos, es decir, su precisión e incertidumbre.

Velocidad: necesidad de algoritmos ágiles para gestionar altas tasas de datos nuevos, dinámicos y patrones cambiantes.

Variedad: necesidad de modelos flexibles capaces de trabajar con datos de distinta naturaleza: imágenes, sensores, video, audio, mensajes estructurados y no estructurados, etc.

Valor: hay que plantearse para quién toma valor estos datos; ¿es oportuno?

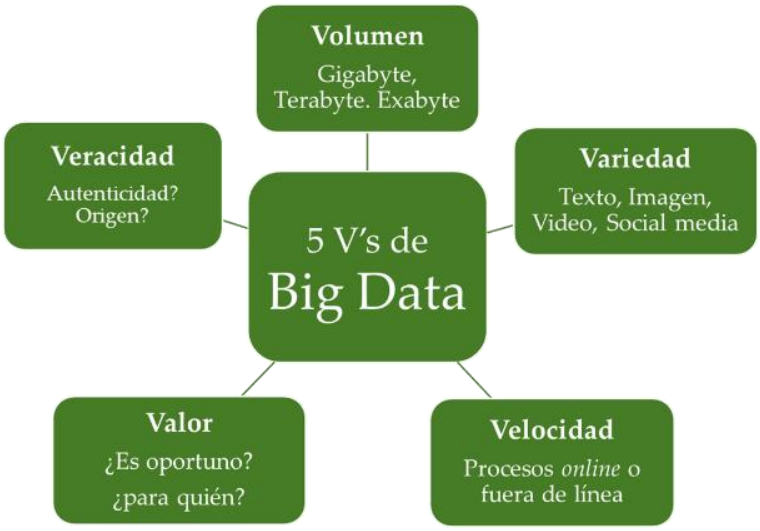


Figura 4: Cinco V's de Big data.

Yin & Kaynak (2015) comentan cómo trabajar estas cinco cuestiones, no sólo para el caso del análisis de canasta en el comercio minorista sino también para otras disciplinas, que de manera sistemática puedan colaborar en mejorar las técnicas existentes y fomentar nuevas ideas entre el mundo académico, industrial y empresarial.

En este aspecto, resulta fundamental el aporte del *database mining* discutido por varios autores (Agrawal et al., 1993a; Agrawal et al., 1993b; Fayyad et al., 1996), que proponen algunos de los elementos necesarios que debieran ser los más adecuados para el procesamiento de datos y el quehacer metodológico cuando estamos en un contexto *big data*.

A partir de estos antecedentes las técnicas basadas en variables latentes surgen como una alternativa viable al análisis de canasta tradicional MBA con AR, como son el análisis de componentes principales PCA de variable continua (Abdi & Williams, 2010; Wold et al., 1987) y el análisis de correspondencia múltiple MCA de variable categórica (Abdi & Valentin, 2007; Greenacre, 2008).

Tanto PCA como MCA, al igual que AR, permiten relacionar los productos de la canasta que se venden juntos y extraer información característica. El rendimiento de los tres modelos mejora si las variables están codificadas y agrupadas en familias de productos (sin tanto detalle y datos agregados, por ejemplo, a 1, 2 o 4 dígitos del UNSPSC). En este aspecto, llevado al extremo, la gran ventaja del análisis de canasta MBA con AR es perfectamente adecuado para casos de varios miles de productos con una mínima agrupación, logrando mucha eficiencia y alto nivel de detalle. En cambio, PCA y MCA requieren que las variables

estén agrupadas homogéneamente; por tanto, un PCA podría procesar bien hasta un par de miles de variables, y varios cientos en el caso de MCA (ver Tabla 10).

El inconveniente principal a considerar en un PCA y MCA para el análisis de canasta MBA, es que son modelos sensibles a la inercia (varianza) de las variables, de modo que ambos métodos deben trabajar sobre variables y categorías agrupadas con "inercias" homogéneas (ver propiedad "Categorías de Alta Inercia", pág. 38).

Algunas consideraciones:

Se puede distinguir del análisis de canasta de licitaciones MBA, revisado en la bibliografía, la dualidad existente entre un ***análisis exploratorio amplio*** del PCA o MCA frente a la ***especificidad*** y gran detalle del AR. Tal vez no se trata de elegir cuál de los tres es el mejor modelo, sino en saber que, para obtener un modelo global más eficiente, estas técnicas pueden usarse juntas (en paralelo, combinadas o secuenciadas). De este modo, el asunto estará en diseñar y decidir en qué etapa del proceso del análisis de canasta de licitaciones conviene usar un modelo MBA tradicional con AR y cuándo conviene usar PCA o MCA. La siguiente tabla muestra un resumen comparativo de las características de cada técnica.

Tabla 10: Comparación de tres técnicas estadísticas: AR, PCA y MCA, para el análisis de canasta de licitaciones MBA.
(*) La columna Nivel de agrupación UNSPSC muestra que PCA y MCA podrían trabajar bien hasta una agrupación a 4 dígitos, y AR hasta 6 dígitos.

Técnica	Variable	Representación Vectorial	Análisis Estático (no supervisado)	(*) Nivel de Agrupación UNSPSC
AR	Cualitativa: Binaria (Ordinal)	Lista de reglas (Selección de alta calidad)	Combinatoria de dos o más variables de mínima agrupación (alto detalle)	1,2,4,6
PCA	Cuantitativa: Continua – Mixta	*Scores y loading*	Análisis Exploratorio - Variables agrupadas (poco detalle, máxima generalidad)	1,2,4
MCA	Cualitativa: Categórica (Ordinal)	*Scores y loading*	Análisis Exploratorio - Variables agrupadas (poco detalle máxima generalidad)	1,2,4

3.3 Análisis de Correspondencia Múltiple (MCA) aplicado al análisis de canasta

En este apartado se revisan los aportes de varios autores al concepto MCA como herramienta multivariante. Además, se valida su técnica y metodología, que en nuestro caso es útil para representar en una canasta los productos y servicios que están presentes en una licitación.

Los orígenes teóricos de esta técnica fueron desarrollados a principios de los años 1970 para dar respuestas en las ciencias sociales. Fue el matemático y lingüista francés Jean-Paul Benzécri quien dio un impulso real a las aplicaciones modernas del MCA. En el libro de Peña (2002) se presenta el análisis de correspondencias múltiple MCA como una técnica descriptiva para representar tablas de contingencia, donde recogemos las frecuencias de aparición de dos o más variables categóricas o cualitativas (discreta, binaria, dicotómica, tricotómica, etc.).

Un objetivo para el análisis de correspondencias múltiple MCA es la representación en un espacio multidimensional reducido, de la relación que existe entre las categorías de dos o más variables no métricas. Greenacre (2008) afirma que MCA es una técnica estadística útil para quienes trabajan con datos categóricos. El método es especialmente eficaz para analizar las tablas de contingencia con datos de frecuencias numéricas.

En la práctica, el MCA puede utilizarse en cualquier contexto donde la representación vectorial y matricial sea un set de observaciones de n variables categóricas con múltiples categorías ordinales. Por ejemplo: set de licitaciones, cuestionario de selección múltiple, respuestas de encuestados, etc. En el caso de variables continuas como la edad, estas se pueden discretizar en tramos homogéneos para convertir todas las variables a categóricas.

3.3.1 Propiedades del MCA: inercia de las categorías y variables

En un modelo MCA, la inercia es un elemento clave que puede afectar de manera relevante a la calidad e interpretación de los resultados. Las categorías que tienen baja frecuencia serán representadas con una alta inercia, causando una gran distorsión al modelo. Es decir, tienen una elevada influencia en los resultados. En nuestro caso de licitaciones, para mitigar la alta inercia de las categorías, las variables de la canasta son agrupadas codificando a 2 dígitos del código UNSPSC (ver línea 2, nivel de "División", Tabla 11). En este nivel se consigue modelar un vector de 55 variables binarias o dicotómicas, con una frecuencia promedio de 2.824 licitaciones por grupo o variable real (mínimo de 233 y máximo de 37.184), lo que representa una alta frecuencia y una inercia baja o muy baja entre grupos. En cambio, si usáramos el código UNSPSC a 8 dígitos, el vector se formaría con 13.596 variables binarias (Ver Tabla 11, nivel de "ítem"), con una frecuencia promedio de 56 licitaciones por grupo o variable real (mínimo de 1 y máximo de 12.617), lo cual representa una frecuencia muy baja en promedio por variable, y alta inercia entre los grupos.

Tabla 11: Frecuencia promedio e inercias por grupo según el Nivel del Código UNSPSC. (*) La frecuencia promedio por grupo real (en cada nivel) se refiere al promedio de veces que ha aparecido cada grupo en las licitaciones realizadas en Chile durante los años 2018 a 2020. (**) corresponden a códigos de productos o categorías que tienen una licitación.

Dígitos código UNSPSC		Nivel	Grupo Teórico UNSPSC	Grupo Real (*) Chile	Promedio de licitaciones por grupo real (*) Chile	Min.	Máx.	Inercia MCA
1 díg.	**1**xxxxxxx	Sección	9	9	15.260	7.781	80.715	Muy Baja
2 díg.	**12**xxxxxx	División	56	55	2.824	233	37.184	Baja
4 díg.	**1234**xxxx	Clase	476	354	581	(**) 1	16.859	Media
6 díg.	**123456**xx	Subclase	4.294	1.930	135	(**) 1	13.730	Alta
8 díg.	**12345678**	Ítem	53.317	13.596	26	(**) 1	12.617	Muy Alta

Representación matricial del set de licitaciones

Al codificar una canasta a dos dígitos del UNSPSC, la representación matricial de un set de licitaciones, caso concreto (Chile, años 2018-2020) tiene las siguientes características:

- Cada observación (licitación) es un vector o canasta definido por 55 variables binarias.
- Cada variable tendrá dos categorías, "0: no está presente en la licitación" y "1: sí está presente en la licitación".
- El 86% de las observaciones son licitaciones mono producto, es decir, sólo una variable toma el valor "1" y las 54 restantes el valor "0".
- El otro 14% corresponde a licitaciones multi producto donde dos o más variables toman el valor "1" y el resto es "0" (en este caso, las observaciones multi producto tienen en promedio 2.6 variables en "1").

En consecuencia, la matriz que modela el set de licitaciones se caracteriza por ser una matriz de baja densidad o rala (muchos ceros y pocos unos) donde las variables son categóricas (binarias o dicotómicas) con valores “0: no está presente” y “1: sí está presente”.

Matriz de licitaciones de baja densidad (rala)

Es importante considerar que la canasta representada como vector, asigna muchos "ceros" y pocos "unos", de modo que la matriz resultante será de baja densidad o rala. Éste será uno de los inconvenientes que debe resolver la técnica seleccionada en el análisis de canasta de licitaciones MBA.

Una matriz de baja densidad, afecta de distintas maneras según la técnica seleccionada en un análisis de canasta MBA. Para este propósito, por ejemplo, cuando trabajamos con un modelo MBA con AR, una matriz rala es una ventaja porque es más fácil encontrar reglas válidas debido a que hay menos combinaciones de productos que se venden juntos. En cambio, al usar un modelo PCA de variable binaria (matriz rala de ceros y unos), afectará negativamente la sensibilidad del modelo por la alta frecuencia de los “ceros” y habrá menor sensibilidad para detectar la información de las variables que toman el valor “uno”. Sin embargo, si consideramos una matriz rala de variable categórica (en este caso de licitaciones es discreta, dicotómica o binaria), entonces el análisis exploratorio debe usar un análisis de correspondencia múltiple MCA que aporta una riqueza y una herramienta muy robusta en el análisis de canasta.

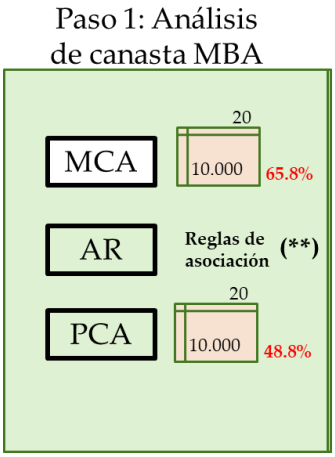


Figura 5: Análisis Exploratorio y reducción de la dimensionalidad, eficiencia del análisis de canasta MBA con AR, MCA y PCA.
(**) El análisis de canasta MBA con reglas de asociación AR, considera sólo canastas de dos o más productos y selecciona aquellas que tienen más probabilidad de ser reglas.

De esta forma vemos en la figura anterior que el modelo MCA es más apropiado y se ajusta mejor al problema de licitaciones porque captura más variabilidad de las variables categóricas y considera todas las observaciones (todos los casos).

3.3.2 Reducción de la Dimensionalidad para Categorías de Alta Inercia en MCA

La representación de cualquier expresión vectorial y matricial de observaciones de n variables con múltiples categorías ordinales (set de licitaciones, formulario de selección múltiple o encuestas), nos permite identificar aquellas categorías de menor frecuencia. Luego, ¿qué hacer cuando una categoría ***j*** de una variable ***i*** es de baja frecuencia (alta inercia) o muy poco representativa del total? Greenacre (2008), describe en la siguiente propiedad un criterio a seguir:

Propiedad de categorías de alta inercia

El modelo MCA se aplica a una matriz con *n* variables cualitativas, cada una de ellas con *m* categorías. Greenacre (2008), sugiere:

> "Si una categoría *j* de la variable *i* es de baja frecuencia debe agruparse con otra categoría próxima (contigua), para evitar que la inercia de la categoría *j* se incremente".

El modelo MCA, como cualquier modelo multivariante, es muy sensible a las observaciones extremas (outliers severos o moderados). Esta **propiedad de categorías de alta inercia** se aplica a las categorías de baja frecuencia (alta inercia) porque quedarán muy por fuera de rango tal como lo hace una observación extrema. Por ello, el modelo MCA mejora cuando se agrupan variables y categorías de baja frecuencia con otra contigua o próxima a ella.

En el supuesto que haya una categoría que tenga una baja frecuencia respecto de las otras, ésta y su alta inercia afectará la capacidad del modelo para capturar toda la variabilidad de los datos. ¿Cuál es el umbral de corte para agrupar o no agrupar dos variables o categorías contiguas?, no hay un criterio o recomendación explícita que fije arbitrariamente un valor umbral de corte cuando se evalúa una variable a la vez, pero existe el criterio de agrupación de Ward que considera de golpe todas las variables y plantea un criterio de agrupación.

3.3.2.1 Agrupación de Ward

A partir de una representación matricial de observaciones de n variables con múltiples categorías ordinales, podemos aplicar una agrupación de Ward que se define a continuación para evitar variables y categorías de alta inercia.

La **agrupación de variables de Ward** es un método jerárquico donde los grupos se unen según el criterio de inercia mínima, que tiene en cuenta los pesos de todos los nodos de un árbol jerárquico. La agrupación de Ward es un algoritmo que evalúa estadísticamente todas las categorías y variables considerando sus pesos para agrupar las variables menos significativas.

Según Greenacre (2008, capítulo 15) la agrupación de variables de Ward es un caso aplicado de la propiedad de categorías de alta inercia, pues es equivalente a hacer una descomposición de la inercia con relación a cada nodo del árbol. Este método de Ward se usa como clúster jerárquico en la función HCPC de la librería FactoMineR (RStudio) que usaremos en este TFM para el análisis de canasta de licitaciones y sirve tanto para agrupar variables categóricas de un modelo MCA como para variables continuas de un modelo PCA.

Peña (2002, p. 252) explica esta propiedad de alta inercia, considerando lo siguiente: "*El análisis de conglomerados de variables es un procedimiento exploratorio que puede sugerir procedimientos de reducción de la dimensión, que podríamos construir a partir de una medida de distancia entre dos variables x_j y x_h representando cada variable como un punto en R^n y calculando la distancia euclídea entre los dos puntos, para construir una matriz de distancias o similitudes entre variables y aplicar a esta matriz un algoritmo jerárquico de clasificación*".

Según Peña (2002, p. 245) el método de Ward, *"para un agrupamiento jerárquico, parte de los elementos directamente, en lugar de utilizar la matriz de distancias, y se define **W** como una **medida global de la heterogeneidad** de una agrupación de observaciones en grupos.*

Esta medida, W, se obtiene a partir de la Ecuación 1, como la suma de las distancias euclídeas al cuadrado entre cada elemento y la media de su grupo, donde $\bar{x}_g$ es la media del grupo g".

$$W = min \sum_g \sum_{i \in g} (x_{ig} - \bar{x}_g)'(x_{ig} - \bar{x}_g) \qquad \text{(Ec. 1)}$$

Tras aplicar el método de Ward, se logra agrupar categorías y reducir la dimensionalidad, esto implica obtener un buen índice ***W*** de la medida global de heterogeneidad conservando sólo las variables y categorías estadísticamente significativas.

3.3.2.2 Matriz de Burt

La **matriz de Burt** se define como una matriz simétrica, sobre la cual se obtienen las componentes principales y las inercias de un modelo MCA (Greenacre, 2008, capítulo 18). Está formada por una tabla de contingencia de doble entrada que resulta del cruce de todos los pares de variables. En la diagonal se hallan los cruces de las variables con ellas mismas.

Tabla 12: Matriz de Burt. Ejemplo Greenacre (2008, capítulo 18)

1T	1t	1C	1?	2T	2t	2C	2?	3T	3t	3C	3?	4T	4t	4C	4?
2501	0	0	0	172	1107	1131	91	355	1710	345	91	1766	538	40	157
0	476	0	0	7	129	335	5	16	261	181	18	128	293	17	38
0	0	79	0	1	6	72	0	1	17	61	0	14	21	38	6
0	0	0	362	1	57	108	196	7	96	55	204	51	45	2	264
172	7	1	1	181	0	0	0	127	48	4	2	165	15	0	1
1107	129	6	57	0	1299	0	0	219	997	61	22	972	239	13	75
1131	335	72	108	0	0	1646	0	24	989	573	60	760	616	84	186
91	5	0	196	0	0	0	292	9	50	4	229	62	27	0	203
355	16	1	7	127	219	24	9	379	0	0	0	360	14	1	4
1710	261	17	96	48	997	989	50	0	2084	0	0	1348	567	23	146
345	181	61	55	4	61	573	4	0	0	642	0	202	286	73	81
91	18	0	204	2	22	60	229	0	0	0	313	49	30	0	234
1766	128	14	51	165	972	760	62	360	1348	202	49	1959	0	0	0
538	293	21	45	15	239	616	27	14	567	286	30	0	897	0	0
40	17	38	2	0	13	84	0	1	23	73	0	0	0	97	0
157	38	6	264	1	75	186	203	4	146	81	234	0	0	0	465

La matriz de Burt (Tabla 12 y Tabla 13) es el resultado de la multiplicación de una matriz binaria ***F*** consigo misma (B = F'F). La matriz binaria ***F*** se forma a partir de una representación vectorial de un set de datos de un formulario de selección múltiple, set de licitaciones o el set de respuestas de una encuesta según corresponda, donde todas las variables son categóricas. En este caso es recomendable haber aplicado previamente una agrupación de Ward que asegure que la matriz de Burt considere sólo las variables y categorías estadísticamente significativas.

Tabla 13: Matriz binaria F. Ejemplo Greenacre (2008, capítulo 18).

Preguntas				Pregunta 1				Pregunta 2				Pregunta 3				Pregunta 4			
1	2	3	4	T	t	C	?	T	t	C	?	T	t	C	?	T	t	C	?
1	3	2	2	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
2	3	3	2	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0
4	3	3	2	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0
4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
1	3	2	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0
⋮	⋮	⋮	⋮		⋮				⋮				⋮				⋮		

En la primera columna "*Preguntas*" (Tabla 13) del ejemplo de Greenacre (2008, capítulo 18), se registran las primeras 6 respuestas de N = 3.418 encuestados de un cuestionario de selección múltiple que consta de 4 preguntas (variables/columnas: "*Pregunta 1*", "*Pregunta 2*", "*Pregunta 3*", "*Pregunta 4*") y cada una con cuatro opciones de selección múltiple (categorías: T, t, C, ?). Las respuestas forman una matriz binaria F_{Nx16} de N encuestados y 16 categorías.

Otro ejemplo Nishisato (2022, capítulo 5), tiene una matriz binaria ***F*** que se forma a partir de un set de datos con Q = 3 variables y C = 3 categorías cada una, en total T = Q·C = 9 categorías, por lo que la matriz de Burt es de 9 filas x 9 columnas (ver Tabla 14). En esta matriz se pueden identificar en la diagonal las frecuencias de las 9 categorías.

La forma de leer la matriz de Burt de este ejemplo, es la siguiente: la frecuencia de la categoría Q1C1 con Q3C1 es 2 (fila 1 y columna 7, en rojo) y corresponde al número de veces que ambas categorías aparecen juntas en un set de datos de selección múltiple.

Tabla 14: Matriz de Burt. Ejemplo Nishisato (2022, capítulo 5).

		Q1			Q2			**Q3**		
		C1	C2	C3	C1	C2	C3	**C1**	C2	C3
Q1	**C1**	3	0	0	2	1	0	**2**	1	0
	C2	0	3	0	0	2	1	1	1	1
	C3	0	0	3	0	1	2	1	1	1
Q2	C1	2	0	0	2	0	0	1	1	0
	C2	1	2	1	0	4	0	1	2	1
	C3	0	1	1	0	0	3	2	0	1
Q3	C1	2	1	1	1	1	2	4	0	0
	C2	1	1	1	1	2	0	0	3	0
	C3	0	1	1	0	1	1	0	0	2

Debemos indicar que por notación la matriz de Burt está compuesta por QxQ sub matrices (en el ejemplo es 3x3, ver Ecuación 2), donde Q es el número de variables y $\boldsymbol{C_{ij}}$ es una tabla de contingencia de doble entrada que relaciona las categorías de la variable ***i*** con las categorías de ***j***, en el caso de ***i=j*** es una matriz diagonal $\boldsymbol{D_{ii}}$, que indica la frecuencia de cada categoría en la variable ***i***.

$$\begin{bmatrix} D_{11} & C_{12} & C_{13} \\ C_{21} & D_{22} & C_{23} \\ C_{31} & C_{32} & D_{33} \end{bmatrix} \qquad \text{(Ec. 2)}$$

3.3.3 Reducción de la Dimensionalidad sobre Variables Latentes en MCA

A partir de la **matriz de Burt** (simétrica), se obtienen las componentes principales y las inercias de un modelo MCA, al igual que otros modelos basados en variables latentes, el modelo MCA puede reducir la dimensionalidad al considerar sólo las J variables latentes relevantes aplicando el criterio de Nishisato y/o el coeficiente Alpha de Cronbach, además, veremos que hay una relación entre ambos índices.

3.3.3.1 Criterio $\lambda > 1/Q^2$ (Nishisato): Índice de correlación promedio al cuadrado

Dada una matriz de Burt de un modelo MCA (ver Tabla 14), Nishisato (1980) sugiere que solo J dimensiones con valor propio que excedan $\lambda > 1/Q^2$ son interesantes de conservar en el modelo, donde $\sqrt{\lambda} > 1/Q$ es el índice de correlación promedio (Nishisato, 1994, 2014, 2022)

Tabla 15: Varianza explicada por el modelo MCA.
Ejemplo: Nishisato (2022, Capitulo 5)

Dimensión	Valor propio λ	% de varianza	Sum % de varianza
Dim. 1	0.4545	48.7	48.7
Dim. 2	0.2620	28.1	76.7
Dim. J=3	0.1138	12.2	88.9
Dim. 4	0.0680	7.3	96.2
Dim. 5	0.0343	3.7	99.9
Dim. T-Q=6	0.0009	0.1	100.0

En este sentido, en la tabla de varianza explicada (Tabla 15), los valores propios sobre una matriz de Burt del ejemplo Nishisato (2022, capítulo 5), se muestran los siguientes resultados sobre las componentes principales. Cuando Q = 3 variables, el criterio $\lambda_{MCA} > 1/Q^2 = 0.1111$ se cumple en la dimensión J=3, esta dimensión explica un 12.2% de variabilidad del modelo y acumula el 88.9% de información.

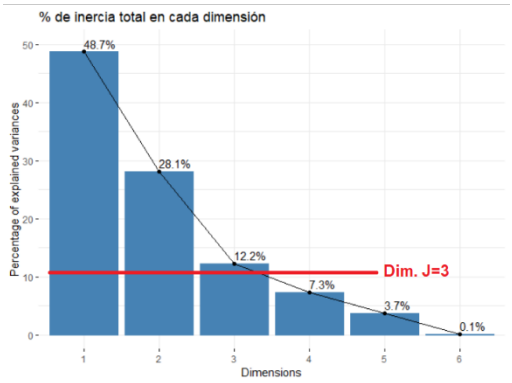


Figura 6: Gráfico de Sedimentación Modelo MCA.
Ejemplo: Nishisato (2022, Capitulo 5)

El Gráfico de sedimentación (Figura 6) muestra una línea roja horizontal que representa el criterio de Nishisato, donde se cumple la condición $\lambda_{J=3} > 1/Q^2 = 1/3^2 = 0.0111$, esto determina que J = 3 son las dimensiones que interesa conservar, las cuales explican el 88.9% de la variabilidad.

Nota 1: El criterio $\lambda_{MCA} > 1/Q^2$ de Nishisato para un MCA de variable categórica, es equivalente al criterio Kaiser en un modelo PCA de variable continua, que considera relevantes las dimensiones que tengan una variabilidad mayor a la unidad $\lambda_{PCA} > 1$.

Nota 2: de la Tabla 15, se puede deducir que sólo 6 dimensiones (variables latentes) explican el 100% de información del modelo. Según el criterio Nishisato ($\lambda > 1/Q^2$), las dimensiones 4 a 6 no son interesantes porque no cumplen el criterio, aun cuando su valor propio es $\lambda > 0$.

Nota 3: ¿Qué pasa con las otras K=3 dimensiones o variables latentes de la matriz de Burt? Estas dimensiones (dim7 a dim9) no aportan información al modelo MCA porque su valor propio λ es 0 (nulo). Se puede verificar que estas dimensiones corresponden a las K=3 categorías de "*menor inercia*".

3.3.3.2 Alpha de Cronbach en MCA: Índice de consistencia interna

En un modelo MCA, el Alpha de Cronbach es un índice de consistencia interna, en inglés *Internal consistency reliability*, que mide la confiabilidad del constructo o variable latente

(dimensión) basándose en las correlaciones de las variables originales. Se trata de evaluar cuánto mejora el ajuste del modelo si se excluye alguna de las dimensiones no relevantes (reducción de la dimensionalidad). Cuanto mayor sea el resultado de α más confiable será la prueba, por lo que se pretende encontrar un Alpha de Cronbach lo más cercano a 1 y mayor que cero (Nishisato, 1980).

Nota metodológica: A menudo las citas traducidas al castellano de varios de los autores que miden la consistencia interna con el Alpha de Cronbach, usan correctamente una traducción literal de *reliability* como fiabilidad, pero en el ámbito de la estadística nos encontramos con otra disciplina que también traduce *reliability* como fiabilidad y se define como "la probabilidad de que un sistema y sus componentes realicen adecuadamente su función prevista a lo largo del tiempo". En este TFM usaremos la traducción de *reliability* como confiabilidad para distinguirla de la fiabilidad que investiga distribuciones de probabilidad y tasa de fallo de las componentes en sistemas complejos.

Se puede obtener la confiabilidad α (índice de consistencia interna) de cada una de las variables latentes del modelo MCA, con el Alpha de Cronbach, a partir de la *Ecuación 3* en función de la suma de las varianzas del ítem sobre la varianza total (Cronbach 1951), de la *Ecuación* 4 con el índice de correlación promedio y su valor singular (Nishisato 1980) o de la *Ecuación 5* con la suma de las correlaciones cuadráticas del ítem.

$$\alpha = \frac{Q}{Q-1}\left(1 - \frac{\sum_q S_q^2}{S^2}\right) \qquad \text{(Ec. 3)}$$

$$\alpha_k = \frac{Q}{Q-1}\left(1 - \frac{1}{Q\sqrt{\lambda}_k}\right) \qquad \text{(Ec. 4)}$$

$$\alpha_k = \frac{\mathrm{Q}}{\mathrm{Q}-1}\left(1 - \frac{1}{\sum_{q=1}^{Q} r_{qk}^2}\right) \qquad \text{(Ec. 5)}$$

En la Tabla 16 se muestra el valor máximo de confiabilidad acumulada Σα = 1.30 calculado hasta la dimensión J = 3. Sin embargo, podríamos optar por no incluir esta dimensión dado que el índice de consistencia interna (Alpha de Cronbach) α_3=0.02 es casi nulo, pero vemos que el aporte de variabilidad de la dimensión 3 es del 12.2%, por lo cual deberíamos conservar esta dimensión en el modelo MCA.

Tabla 16: índice de consistencia interna: Alpha de Cronbach para un Modelo MCA

Dimensión k	Valor propio λ	% de varianza	Sum % de varianza	Valor Singular $\sqrt{\lambda}$	Alpha de Cronbach α	Σα
dim 1	0.4545	48.7	48.7	0.6742	0.76	0.76
dim 2	0.2620	28.1	76.7	0.5119	0.52	1.28
J = dim 3	**0.1138**	**12.2**	**88.9**	**0.3373**	**0.02**	**1.30**
dim 4	0.0680	7.3	96.2	0.2608	-0.42	0.88
dim 5	0.0343	3.7	99.9	0.1853	-1.20	-0.32
dim 6	0.0009	0.1	100.0	0.0305	-14.91	-15.22

En rigor este coeficiente α no puede volverse negativo ya que se define como una relación entre dos números que teóricamente son positivos. Sin embargo, cuando descomponemos los datos en componentes, no es raro que obtengamos valores negativos para el alfa de Cronbach (Nishisato, 2022).

En teoría, α_k alcanza su máximo valor de 1 cuando todos los ítems están perfectamente correlacionados, y se vuelve negativo cuando la suma de las correlaciones al cuadrado del

ítem *k* $\left(Q\sqrt{\lambda_k} = \sum_{q=1}^{Q} r_{qk}^2 < 1\right)$ es menor que 1, ver Ecuación 5 (Greenacre & Blasius 2006, capitulo 7, página 173 - 174). Esta expresión es equivalente a la afirmación de Nishisato (2014) que sostiene que α se vuelve negativo cuando el valor singular $\sqrt{\lambda}$ es menor que 1/Q ($\sqrt{\lambda} < 1/Q = 1/3$), ver *Ecuación 4*.

Del párrafo anterior, para el caso de este ejemplo, en la práctica podemos confirmar que el valor más alto de α es 0.76, que se alcanza en la primera dimensión, obviamente porque las Q variables no están perfectamente correlacionadas. Pero si se cumple el criterio de Nishisato $\sqrt{\lambda} > 1/Q$, será el punto de corte para seleccionar las J=3 variables latentes relevantes del modelo MCA.

Reemplazando la suma de las correlaciones cuadráticas $Q\sqrt{\lambda_k} = \sum_{q=1}^{Q} r_{qk}^2$ en la Ecuación 5 es equivalente a la Ecuación 4, además la doble sustitución en la Ecuación 3 de la varianza $Q^2\sqrt{\lambda} = S^2$ y la suma de varianzas de los k ítems $Q = \sum_{k=1}^{Q} S_k^2$ es equivalente a la Ecuación 4**.**

En la Ecuación 6 (Greenacre & Blasius, 2006, p. 173; Greenacre, 2006, capítulo 7) se establece la relación entre el Alpha de Cronbach y el criterio de Nishisato, que corresponde al índice de correlación promedio (lado izquierdo de la Ecuación 6) y es igual a 1/Q (Nishisato, 1994, 2014, 2022). Por tanto, de esta ecuación se puede inferir el razonamiento de Nishisato donde sugiere que solo las J dimensiones con valor propio λ que excedan $\sqrt{\lambda} > 1/Q$ son interesantes de conservar en el modelo MCA. Es equivalente a sugerir que el modelo debe conservar sólo las dimensiones cuya confiabilidad sea positiva ($\alpha_k > 0$, Alpha de Cronbach mayor que cero).

$$\frac{\sum_{i=1}^{Q}\sqrt{\lambda_i}}{(T-Q)*\sqrt{\lambda_i}} = \frac{1}{Q*\sqrt{\lambda_i}} < 1 \;\forall i = 1 \dots Q$$

$$\frac{\sum_{i=1}^{Q}\sqrt{\lambda_i}}{(T-Q)} = \frac{1}{Q} < \sqrt{\lambda_i} \qquad (6)$$

Q: variables cualitativas del modelo MCA
K: valores propios no nulos $\lambda_k > 0$ (K = T - Q = 9 - 3 = 6)
T: total categorías (Q·C = 3*3 = 9)
J: variables latentes que cumplen el criterio de Nishisato y confiabilidad α>0

Nota metodológica: Algunos autores como Nishisato (1980), Greenacre (2008) y Lê et al. (2008) usan la matriz de Burt o la matriz binaria para obtener las componentes principales y las inercias del modelo MCA, el resultado no es el mismo para uno u otro caso, luego λ_k será la k-ésima inercia principal (valor propio) de la matriz de Burt y $\sqrt{\lambda_k}$ la k-ésima inercia principal de la matriz binaria. Esta distinción es importante porque algunos softwares o librerías entregan resultados según usan una u otra matriz. Por otro lado, los autores usan una notación coherente según el desarrollo de sus definiciones, por lo cual suele confundir el hecho de que algunos usen para calcular la confiabilidad α_k el valor propio λ_k o bien con el valor singular $\sqrt{\lambda_k}$. En este TFM, consideramos la notación de Nishisato (1980), donde la confiabilidad α_k deberá calcularse según la Ecuación 4 con $\sqrt{\lambda_k}$ según hemos presentado en el desarrollo metodológico de este apartado. En ambos casos, siendo consistente con la notación que cada autor da a su desarrollo metodológico, la confiabilidad α_k calculada obtendrá el mismo valor para cada dimensión *k;* esta afirmación se puede comprobar en la Tabla 28 (pág 61) donde se analiza esta condición particular.

4 RESULTADOS Y DISCUSIÓN

4.1 Análisis de canasta de licitaciones MBA: Comparación entre AR, MCA y PCA

Al revisar el aspecto teórico y metodológico aplicado del análisis de canasta MBA con reglas de asociación AR, se confirma que el resultado esperado de esta técnica es la capacidad de seleccionar un conjunto de patrones que tengan una alta probabilidad de ser reglas. En particular, el algoritmo *Apriori* de un AR evalúa todas las combinaciones para luego podar hasta obtener reglas de buena calidad y con la mejor combinación de productos que se venden juntos en una canasta.

La consecuencia de la poda es que reduce el dominio de solución a unos pocos casos (las mejores reglas) dejando fuera de la selección observaciones o segmentos de licitaciones que eventualmente podrían ser de interés. Una condición obvia del AR es que sólo trabaja con canastas de dos o más productos, lo cual descarta las licitaciones mono producto que corresponden al 86% de las licitaciones en Chile durante los años 2018 a 2020.

Respecto a los modelos basados en variables latentes tenemos el PCA, técnica desarrollada para variables continuas, y el MCA para variables categóricas. Para el caso de licitaciones, las variables de la canasta se interpretan mejor como variables categóricas binarias {“0”, “1”} por lo cual el modelo MCA sería el más aconsejable para realizar un análisis de canasta MBA.

En el punto 5.4.2 se compara el dominio de solución PCA y MCA con los datos de licitaciones registradas entre los años 2018 a 2020. Los resultados confirman que el análisis exploratorio y la reducción de la dimensionalidad de 55 variables categóricas a 20 variables latentes del modelo PCA, es un 17% peor que el obtenido con MCA; es decir, un modelo PCA captura menos variabilidad de los datos que un modelo MCA cuando todas las variables son categóricas. En consecuencia, en una primera instancia descartamos el modelo PCA de variable continua binaria, ya que técnicamente no es el más adecuado cuando “todas” las variables son categóricas.

4.2 Validación Metodológica: Clasificador de Licitaciones a un dígito UNSPSC

En este punto se aplica la teoría de subastas de Milgrom y Wilson en el contexto de las licitaciones en Chile y la de otros países miembros OCDE que se rigen por las mismas recomendaciones y estándares internaciones de contratación pública. También se revisa en un contexto práctico la teoría del análisis de canasta MBA en relación con los modelos AR, PCA y MCA para comparar ventajas e inconvenientes y justificar que, en el caso de licitaciones, MCA es más apropiado para construir un clasificador de licitaciones múltiple ya que logra una precisión mayor al 95% de aciertos en 15 segmentos, cuando las variables de la canasta son categóricas y han sido agrupadas a un dígito del código UNSPSC.

Los resultados se fundamentan en el marco teórico que busca reforzar y dar un contexto al MCA como técnica estadística multivariante muy útil para quienes trabajan con datos categóricos, que en nuestro caso resulta ser relevante en el análisis de canasta de licitaciones y en la construcción del clasificador de licitaciones. Aunque MCA fue desarrollado hace más de 50 años por Jean-Paul Benzécri para trabajar en aplicaciones del mundo de las ciencias sociales, ahora su ámbito de aplicación es más amplio y se extiende a otros sectores como por ejemplo en procesos industriales, agrícolas, logística, “marketing y fidelización”, servicios y medicina entre varios otros.

Con el fin de validar la incorporación de nuevas herramientas, buscamos plantear un debate técnico que cuestione o promueva las características y potencialidades de un MCA como técnica multivariante para el análisis de canasta MBA, donde la herramienta tradicional e indiscutida ha sido el AR que ha aportado con sencillez al conocimiento e información para el tomador de decisiones en el ámbito del comercio minorista.

A continuación, se describe la secuencia metodológica para construir y validar un clasificador de licitaciones codificado a un dígito del UNSPSC, que parte con la selección de una muestra, seguido de un análisis de canasta, segmentación y clasificación.

4.2.1 Selección de la muestra de tamaño 6.000

Se selecciona una muestra aleatoria de 6000 observaciones (licitaciones) representativas de la base de datos de 330.000 licitaciones del Estado de Chile entre los años 2018 a 2020. Según revisamos en el código UNSPSC Productos y Servicios a un dígito, en este nivel corresponde a licitaciones que se agrupan en 9 variables (productos) y dos categorías {0, 1} cada una.

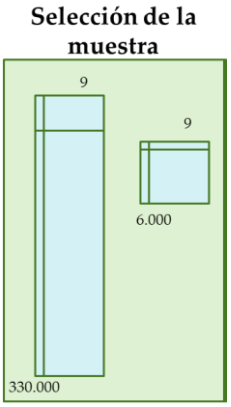


Figura 7: Selección de la Muestra 6000 (Nivel de Sección: a un dígito)

La muestra debe estar necesariamente equilibrada, ya que el 86,4% de las licitaciones tienen un solo producto y el 13,6% dos o más. Esto permite mejorar la calidad y precisión del resultado al construir el clasificador de licitaciones.

Para equilibrar la muestra se ha empleado la función set.seed(123) y set.seed(1234) de R-Studio para obtener dos submuestras de 3.000 observaciones cada una. La primera submuestra selecciona aquellas licitaciones que tienen un solo producto en la canasta (Tabla 17(a) en verde) y la segunda las que tienen dos o más (Tabla 17(b) en azul). En la Tabla 17, podemos ver 6 ejemplos de cada submuestra con su respectivo identificador (Id) y un total de productos licitados, en rojo vemos qué productos están presentes en la licitación.

Tabla 17: Muestra equilibrada de 6.000 licitaciones (Nivel de Sección: a un dígito UNSPSC).
(a) Submuestra de 3000 licitaciones mono producto (b) Submuestra de 3000 licitaciones Multi Producto.
Se muestran seis ejemplos de cada submuestra.

(a)

Id	X1	X2	X3	X4	X5	X6	X7	X8	X9	Total
147530	0	0	0	0	0	0	1	0	0	1
322843	0	0	0	0	0	0	1	0	0	1
191170	0	0	0	0	0	0	0	0	1	1
356797	0	0	0	0	0	0	1	0	0	1
377054	0	0	0	0	0	0	0	1	0	1
64985	0	0	0	0	1	0	0	0	0	1

(b)

Id	X1	X2	X3	X4	X5	X6	X7	X8	X9	Total
85989	0	1	1	0	0	0	0	0	0	2
258137	1	1	1	1	1	1	0	0	0	6
254307	1	0	0	1	0	0	0	0	0	2
259129	0	0	0	1	1	0	0	0	0	2
339126	1	0	0	1	0	1	0	1	0	4
265448	1	0	0	1	0	0	0	0	0	2

La siguiente tabla muestra la frecuencia de cada variable y la inercia de sus dos categorías cuando se han considerado las dos submuestras de 3.000 licitaciones cada una.

Tabla 18: Frecuencia e Inercia de 9 variables y 2 categorías - para una muestra de 6000 licitaciones (Nivel de Sección: a un dígito UNSPSC)

Variables {X_1 a X_9}	Productos X_1 a X_6						Servicios X_7 a X_9				
Categoría {0 o 1}	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Frecuencia	Inercia
0: no tiene	4692	4728	4612	3316	4191	5649	5005	5202	5811	Alta	Baja
1: sí tiene	1308	1272	1388	2684	1809	315	995	798	189	Baja	Alta
Total	6000	6000	6000	6000	6000	6000	6000	6000	6000		

Tabla 19: Matriz de Burt (18x18) - para una muestra de 6000 licitaciones (Nivel de Sección: a un dígito UNSPSC)

	X_1_0	X_1_1	X_2_0	X_2_1	X_3_0	X_3_1	X_4_0	X_4_1	X_5_0	X_5_1	X_6_0	X_6_1	X_7_0	X_7_1	X_8_0	X_8_1	X_9_0	X_9_1
X_1_0	4692	0	3871	821	3813	879	2756	1936	3473	1219	4519	173	3806	886	3982	710	4514	178
X_1_1	0	1308	857	451	799	509	560	748	718	590	1130	178	1199	109	1220	88	1297	11
X_2_0	3871	857	4728	0	3938	790	2798	1930	3277	1451	4523	205	3845	883	4006	722	4549	179
X_2_1	821	451	0	1272	674	598	518	754	914	358	1126	146	1160	112	1196	76	1262	10
X_3_0	3813	799	3938	674	4612	0	2710	1902	3188	1424	4425	187	3707	905	3896	716	4435	177
X_3_1	879	509	790	598	0	1388	606	782	1003	385	1224	164	1298	90	1306	82	1376	12
X_4_0	2756	560	2798	518	2710	606	3316	0	2495	821	3191	125	2468	848	2667	649	3151	165
X_4_1	1936	748	1930	754	1902	782	0	2684	1696	988	2458	226	2537	147	2535	149	2660	24
X_5_0	3473	718	3277	914	3188	1003	2495	1696	4191	0	3964	227	3355	836	3538	653	4031	160
X_5_1	1219	590	1451	358	1424	385	821	988	0	1809	1685	124	1650	159	1664	145	1780	29
X_6_0	4519	1130	4523	1126	4425	1224	3191	2458	3964	1685	5649	0	4686	963	4883	766	5467	182
X_6_1	173	178	205	146	187	164	125	226	227	124	0	351	319	32	319	32	344	7
X_7_0	3806	1199	3845	1160	3707	1298	2468	2537	3355	1650	4686	319	5005	0	4253	752	4823	182
X_7_1	886	109	883	112	905	90	848	147	836	159	963	32	0	995	949	46	988	7
X_8_0	3982	1220	4006	1196	3896	1306	2667	2535	3538	1664	4883	319	4253	949	5202	0	5026	176
X_8_1	710	88	722	76	716	82	649	149	653	145	766	32	752	46	0	798	785	13
X_9_0	4514	1297	4549	1262	4435	1376	3151	2660	4031	1780	5467	344	4823	988	5026	785	5811	0
X_9_1	178	11	179	10	177	12	165	24	160	29	182	7	182	7	176	13	0	189

La matriz de Burt de 18x18 es una representación matricial, de una tabla de doble entrada, de las frecuencias entre las 18 categorías que se han modelado. La diagonal de la matriz coincide con los valores de la Tabla 18 de frecuencias. En verde están representadas las 9 sub matrices diagonales de 2x2 que corresponden a la tabla de contingencia de cada variable *Xi* con ella misma y sus dos categorías.

4.2.2 Paso 1: MCA - Análisis de Canasta a un dígito UNSPSC

Dada la muestra aleatoria de 6000 observaciones (licitaciones) y una canasta de 9 variables, el modelo MCA forma una matriz de Burt de 18x18 (con Q = 9 variables), el criterio de Nishisato sugiere considerar sólo las J = 5 dimensiones con valor propio mayor a $\lambda > 1/Q^2 = 0.012$.

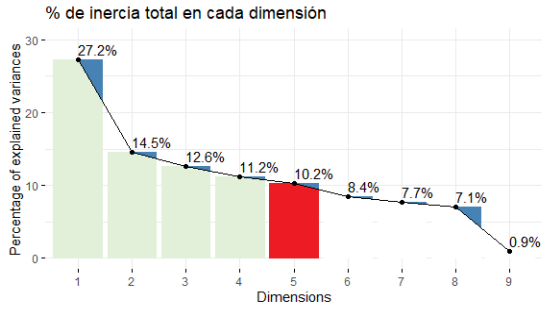


Figura 8: Gráfico de Sedimentación Modelo MCA (Nivel de Sección: a un dígito UNSPSC)

Tabla 20: Varianza explicada por el modelo MCA (Nivel de Sección: a un dígito UNSPSC)

Q	Var λ	% var.	Σ % var.	$1/Q^2$
Dim.1	0.034	27.2	27.2	
Dim.2	0.018	14.5	41.8	
Dim.3	0.015	12.6	54.3	
Dim.4	0.014	11.2	65.4	
J=Dim.5	**0.013**	**10.2**	**75.9**	**0.012**
Dim.6	0.011	8.6	84.5	
Dim.7	0.009	7.4	91.9	
Dim.8	0.009	7.1	98.9	
Dim.9	0.001	1.1	100.0	

El Gráfico de sedimentación muestra 9 dimensiones de valores propios no nulos y porcentaje de inercia explicada. En la Tabla 20 de varianza acumulada total, se indica que J = 5 dimensiones explican el 75.9% de la variabilidad capturada por el modelo.

En la siguiente figura se representa el proceso de selección de la muestra de 6000 observaciones equilibradas y, en el paso 1, se muestra la matriz de *scores* de 6000x5 que es el resultado del análisis exploratorio de un modelo MCA.

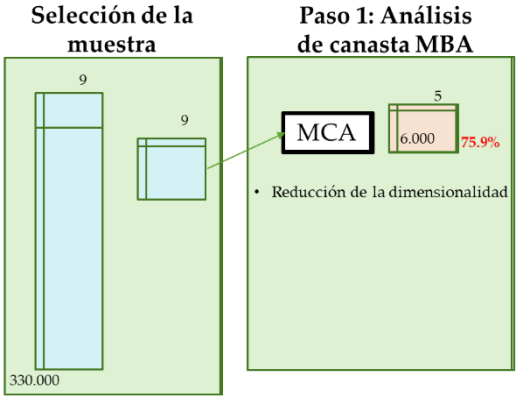


Figura 9: Paso 1: Análisis de canasta MBA - Modelo MCA (Nivel de Sección: a un dígito UNSPSC)

4.2.2.1 Análisis exploratorio MCA

La Figura 10 y 11 muestran los gráficos de *loadings* (variables) y *scores* (observaciones), a partir de los cuales se pueden interpretar las variables latentes del modelo MCA.

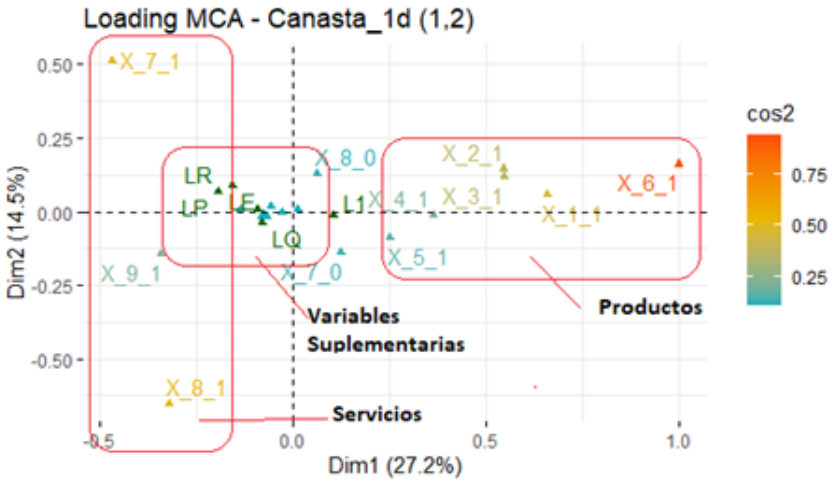


Figura 10: Gráfico de loadings (variables) del modelo MCA en la dimensión (1,2) - Canasta a un dígito UNSPSC.
cos2: es un indicador que corresponde al gradiente de contribución de calidad que la dimensión k aporta a las Q variables.

En este contexto, se entiende por calidad cos2 a la contribución de la dimensión J a las Q variables y viene dada por la proporción de inercia de una variable explicada por una dimensión (contribución relativa por dimensión). Las menos representadas por estas dos dimensiones son las variables X8 y X9 (cerca del centro (0,0)) y de color azulado, por lo que habrá otras dimensiones que aporten mayor calidad a estas variables.

La 1ª Variable latente representa una variable subyacente que podemos interpretar en el gráfico de *loadings* (Figura 10) como Productos vs. Servicios: representa el 27.2% de la información del modelo, distingue entre productos y servicios. Muestra, de mayor a menor inercia, los 6 Productos a la derecha (X_{1_1}, X_{2_1}, X_{3_1}, X_{4_1}, X_{5_1}, X_{6_1}), los 3 servicios a la izquierda (X_{7_1}, X_{8_1}, X_{9_1}) y al centro, con baja inercia, las variables suplementarias de tipo de licitación (L1, LE, LP, LR, LQ). En adelante estas variables suplementarias no serán parte del análisis de canasta debido a que no aportan inercia y no son relevantes para la discusión.

La 2ª Variable latente interpreta un constructo sobre los Servicios: representa el 14.6% de la información, se separan claramente ciertos servicios en forma vertical, de arriba hacia abajo (X_{7_1}, X_{8_1}, X_{9_1}).

El gráfico de *Scores* (Figura 11) para las 6.000 observaciones proyectadas en la primera y segunda dimensión se ven claramente tres nubes de puntos (muestra sólo aquellas con gradiente de calidad cos2>0.1), que en apariencia corresponden a tres segmentos. Sin embargo, en el apartado HCPC para la segmentación de canasta a 1 dígito (ver p. 50) se comprobará que, al considerar 5 dimensiones, la cantidad óptima de segmentos es cinco.

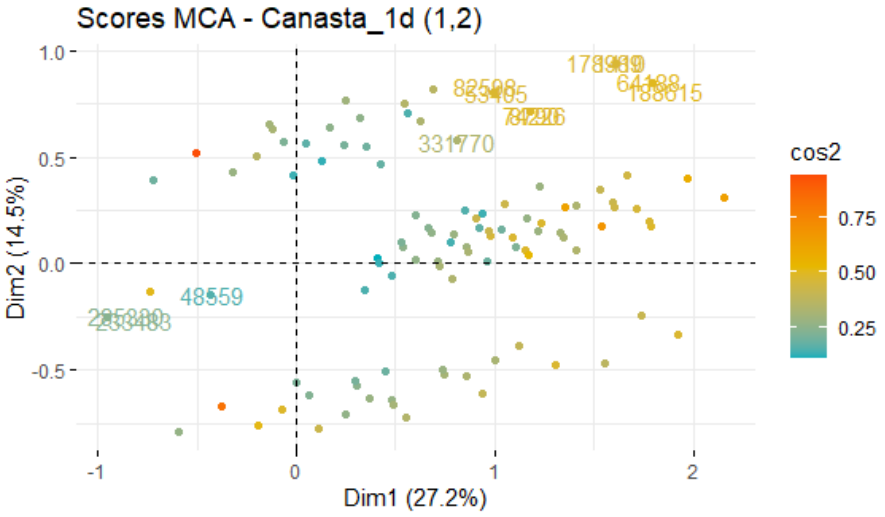


Figura 11: Gráfico de Scores (observaciones) del modelo MCA en la dimensión (1,2) - Canasta a un dígito UNSPSC.

El indicador cos2 corresponde al gradiente de contribución de calidad que la dimensión k aporta a las N observaciones. En este gráfico de scores sólo se proyectan las observaciones de calidad mayor a 0.1 (cos2>0.1) para evitar la superposición de puntos de baja calidad de estas dimensiones.

La 3ª Variable latente (Figura 12) no representa un constructo bien definido por lo que se considera Mixto: representa el 12.5% de la información. Esté gráfico muestra con mayor inercia tres elementos, X_{5_1} y X_{9_1} a la izquierda frente a X_{6_1} a la derecha.

La 4ª Variable latente (constructo) Servicio está determinada por $\mathbf{X_{9\_1}}$: representa el 11.1% de la información, muestra con mayor inercia el servicio X_{9_1}.

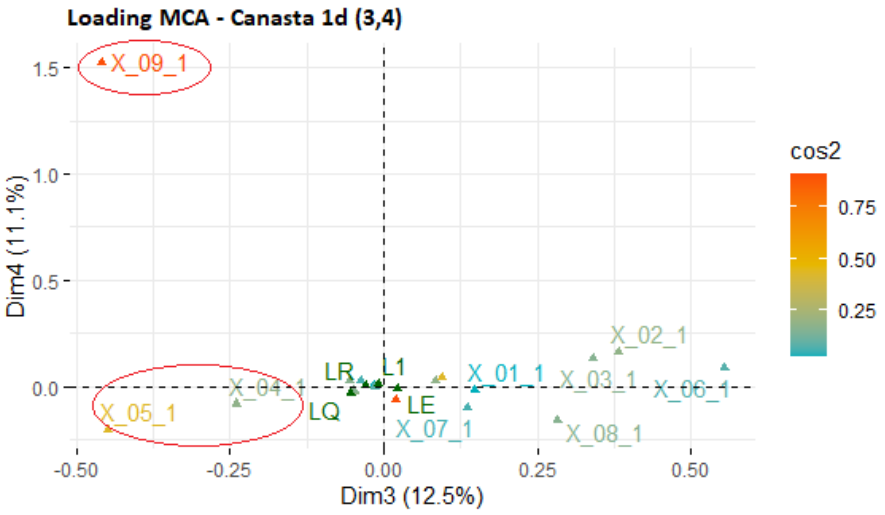


Figura 12: Gráfico de Loading (variables) del modelo MCA en la dimensión (3,4) - Canasta un dígito UNSPSC.

cos2: Es un indicador que corresponde al gradiente de contribución de calidad que la dimensión k aporta a las Q variables

En los gráficos de *loading* y *scores* de la dimensión 3 y 4 (Figura 12 y 13), se encuentran agrupadas y alejadas del resto de puntos, un conjunto de licitaciones que en ambos gráficos están acotadas por la elipse en rojo y se destacan por tener una contribución a la calidad cos2 muy alta en estas dimensiones cercana al 0.75, corresponden a las variables con más peso X_{9_1}, X_{5_1}, X_{4_1} y a las licitaciones {1348, 93173, 95279, 142796, 234908} (ver Tabla 21)

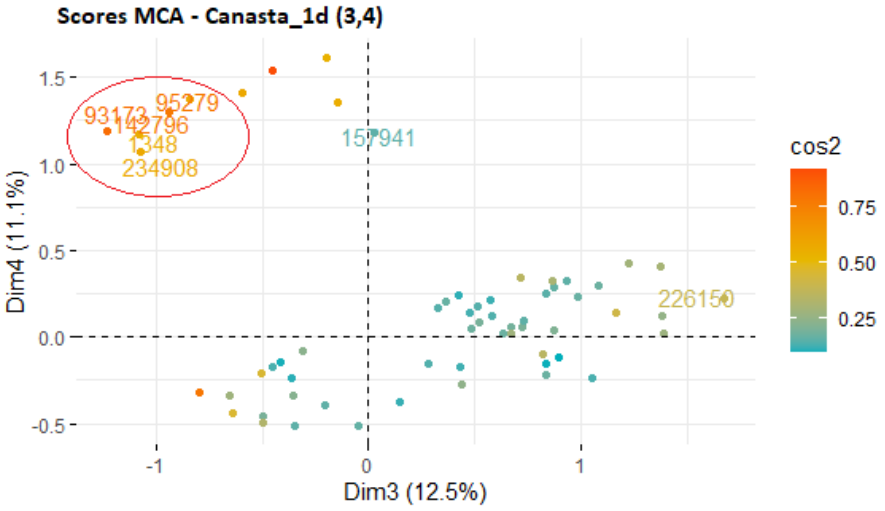


Figura 13: Gráfico de Scores (observaciones) del modelo MCA en la dimensión (3,4) – Canasta un dígito UNSPSC.

El indicador cos2 corresponde al gradiente de contribución de calidad que la dimensión k aporta a las N observaciones. En este gráfico de scores sólo se proyectan las observaciones de calidad mayor a 0.1 (cos2>0.1) para evitar la superposición de puntos de baja calidad de estas dimensiones.

En la siguiente tabla se muestra un ejemplo de 5 licitaciones, en términos generales, en este segmento predomina con mayor presencia un servicio X_{9_1} que se licita junto con otros dos productos X_{5_1}, X_{4_1}. En efecto, a simple vista en la Figura 13, estas 5 canastas podrían pertenecer a un segmento en particular, dado que son similares y determinan un patrón predominante que se licita junto.

Tabla 21: Selección de 5 observaciones, con una canasta predominante de los productos {X9, X5, X4}

Id	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Tipo Licitación Variable suplementaria
1348	1			1	1				1	LE
93173				1	1				1	L1
95279					1				1	L1
142796					1				1	L1
234908				1	1		1		1	LE

4.2.3 Paso 2: HCPC - Segmentación de Canasta a un dígito UNSPSC

Este apartado está enfocado en obtener un modelo de segmentación jerárquica de licitaciones (a un dígito del código UNSPSC), que permita encontrar segmentos que caractericen 15 prototipos de canastas o clases.

Las 5 variables latentes relevantes representadas por la matriz de *scores* (resultado del paso 1) aportan el 75.9% de la información capturada por el modelo MCA, la matriz de *scores* es el elemento de entrada en la función HCPC para realizar una segmentación de 15 clases. La función HCPC de la librería FactoMineR de R-Studio tiene la particularidad que puede trabajar en una modalidad MCA si todas las variables son categóricas, o bien PCA si las variables de la matriz original son consideradas variables continuas (en este último caso, puede incluir algunas variables categóricas binarias).

En la siguiente figura, a partir de la matriz de *scores*, se representa la secuencia del paso 1 y 2 para conseguir una segmentación de la muestra en 'c' de 15 clases.

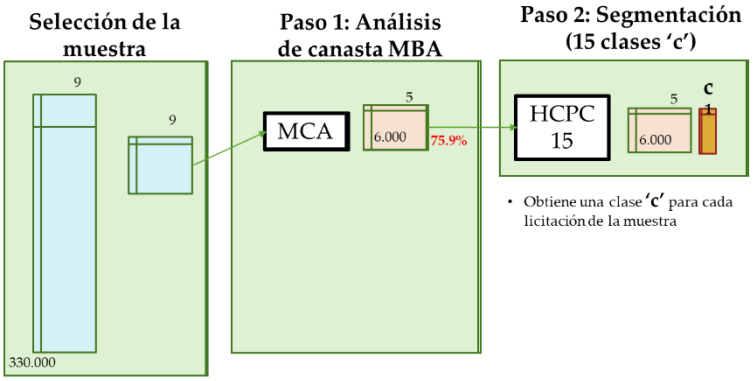
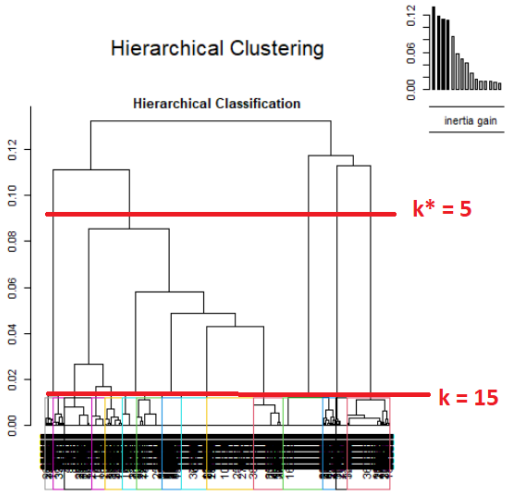
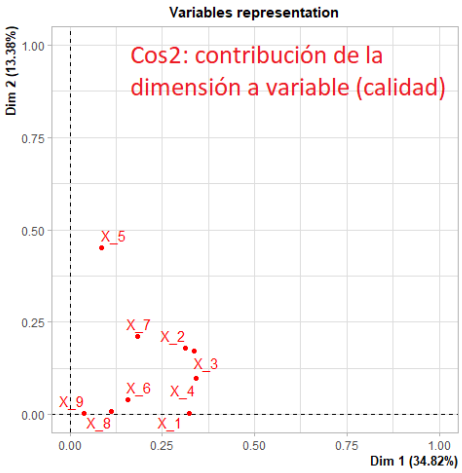


Figura 14: Paso 2: Segmentación de la muestra en 15 clases - Modelo HCPC (Nivel de Sección: a un dígito UNSPSC)

El Clúster jerárquico HCPC agrupa las observaciones de la muestra en 'c' clases. El óptimo del modelo, tal como se ve en el dendograma (Figura 15a), es k* = 5 clases, pero se necesita al menos k = 15 para forzar al modelo a representar más casos y más combinaciones de productos que se licitan juntos (sin perder sensibilidad del modelo).



(a)



(b)

Figura 15: Modelo Clúster Jerárquico HCPC - (Nivel de Sección: a un dígito UNSPSC). (a) dendograma de agrupamiento jerárquico con k=5 y k=15. (b) Gráfico de Contribuciones Dimensión (1, 2)*

En el gráfico de contribuciones (b), el indicador Cos2 corresponde a la contribución de calidad que la dimensión 1 y 2 aporta a cada variable. El concepto calidad cos2 o contribución de la dimensión se define en página 47.

En la siguiente tabla, están representadas las k=15 clases 'c' con la frecuencia de cada Variable. Podemos ver que la segmentación HCPC de las primeras 6 clases agrupa licitaciones en la que predomina claramente un producto (en naranja), en el caso de las otras 9 clases se produce una combinación de dos o más productos predominantes que se licitan juntos en una canasta.

Tabla 22: Frecuencia del segmento 'c' - Clúster Jerárquico HCPC (Nivel de Sección: a un dígito UNSPSC)

Clase 'c'	Frecuencia Clase	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
c1	758	14	24	10	23	0	0	755	0	0
c2	697	21	16	22	76	86	0	25	687	0
c3	189	10	10	11	24	28	7	7	12	185
c4	119	35	9	12	41	108	0	114	10	0
c5	436	0	0	0	0	434	0	0	0	0
c6	814	0	0	0	802	0	0	0	0	0
c7	344	0	0	0	325	325	0	0	0	0
c8	523	0	0	514	192	87	0	7	0	0
c9	467	0	458	0	179	69	0	4	0	0
c10	394	389	0	101	140	0	0	1	1	0
c11	313	304	0	50	176	304	0	0	10	0
c12	197	64	4	50	89	47	194	8	10	0
c13	260	0	256	256	166	49	0	10	10	0
c14	342	330	330	222	251	147	0	31	25	0
c15	147	104	134	111	126	70	142	21	21	0
	6.000	1.271	1.241	1.359	2.610	1.754	343	983	786	185

Vemos también que cada producto dominante en una clase 'c' tiene una frecuencia alta (mayor a 100) y se acompaña de otros con menor frecuencia. Esto implica que el producto dominante (en naranja) y su combinación con otros productos "es la característica o patrón de la clase".

4.2.4 Paso 3: SVM – Clasificador de Licitaciones a un dígito UNSPSC

La efectividad del paso 1 y 2 permite realizar en el paso 3 una buena estimación 'ĉ'. Para la clasificación de 'c', el modelo seleccionado es un clasificador SVM que obtiene un alto nivel de aciertos y es consistente con todo el proceso de trabajo realizado con los datos desde la selección de la muestra.

Para el paso 3 se ha utilizado la función ***ksvm*** de la librería "kernlab" de R-Studio, correspondiente a las máquinas de soporte vectorial de un método supervisado (SVM *Support Vector Machine*). En modo clasificador requiere dos entradas: una es la clase 'c' que viene de la segmentación provista por la función HCPC del paso 2 y la otra entrada es la matriz de *scores* del modelo MCA obtenida en el paso 1.

Para ajustar el modelo de clasificación SVM en el paso 3, la matriz de *scores* con 6.000 observaciones ha sido separada en dos submuestras aleatorias para usar una de 2.000 observaciones en el entrenamiento y otra de 4.000 para test.

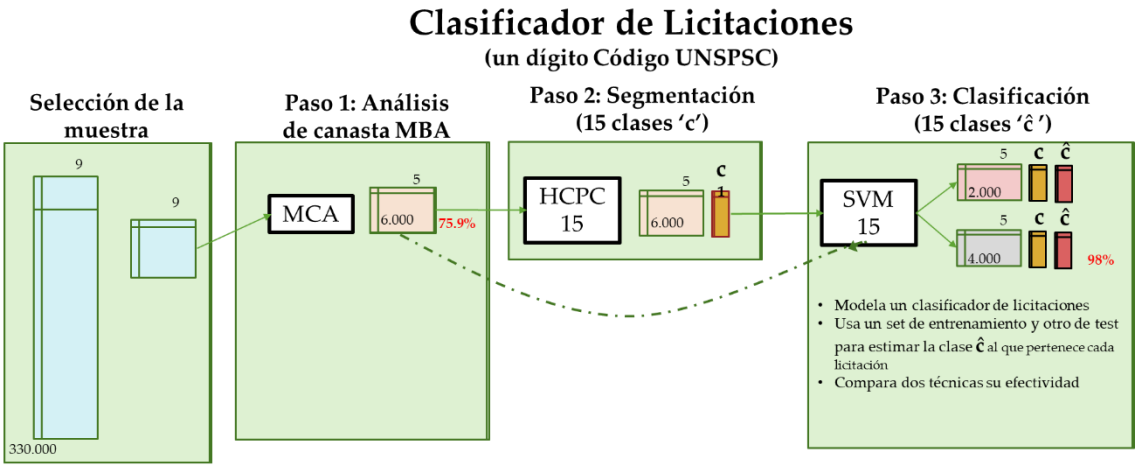


Figura 16: Paso 3: Clasificación de licitaciones - Modelo SVM (Nivel de Sección: a un dígito UNSPSC)

En la Figura 16 vemos que el análisis exploratorio (paso 1) aporta la matriz de *scores* con sus 5 variables latentes relevantes y la segmentación (paso 2) aporta la clase 'c'. Una vez ajustado el modelo (paso 3), el clasificador SVM logra una clasificación que estima la clase 'ĉ' ~ 'c', con un porcentaje de aciertos sobre el 95% de las 4.000 observaciones de la submuestra test.

Al aumentar el tamaño de la submuestra para entrenamiento sólo mejora marginalmente la tasa de aciertos. Por ejemplo, si la submuestra para el entrenamiento aumentara a 3.000 o a 4.000 observaciones la tasa de aciertos subiría al 98.6% y al 99.2% respectivamente.

4.3 Clasificador de Licitaciones a dos dígitos UNSPSC

En este apartado se pone a prueba la metodología para diseñar un clasificador de licitaciones basado en un análisis de canasta de licitaciones MBA sobre variables cualitativas, cuando los productos de la canasta han sido codificados a dos dígitos del código UNSPSC. Se busca obtener resultados para su interpretación con un clasificador eficiente y robusto, y verificar la confiabilidad del modelo MCA.

4.3.1 Selección de la muestra de tamaño 10.000

Se selecciona una muestra aleatoria de 10.000 observaciones (licitaciones) representativas de la base de datos de 330.000 licitaciones del Estado de Chile entre los años 2018 a 2020,

codificadas a dos dígitos del UNSPSC, que al nivel de "división" agrupa productos y servicios en 55 divisiones con dos categorías cada una {0: No está presente, 1: Sí está presente}.

La muestra debe estar equilibrada entre canastas mono productos y multi productos, esto es un requisito que permite mejorar la calidad y precisión del resultado al construir el clasificador, dado que el 86,4% de las licitaciones tienen un solo producto y el 13,6% dos o más.

Para equilibrar la muestra usamos la función set.seed(123) y set.seed(1234) de R-Studio para obtener dos submuestras de 5.000 observaciones cada una. La primera submuestra selecciona aquellas licitaciones que tienen un solo producto en la canasta y la segunda las que tienen dos o más.

4.3.2 Paso 1: MCA - Análisis de Canasta a dos dígitos UNSPSC

Dada la muestra aleatoria de 10.000 observaciones y al modelar una canasta de licitaciones a dos dígitos del código UNSPSC, se puede construir la matriz de Burt de 110x110 (con Q = 55 variables y dos categorías cada una, tal que T = Q·2 = 110). A partir de ésta se obtienen las componentes principales y las inercias de un modelo MCA.

4.3.2.1 Reducción de la dimensionalidad - Criterio $\lambda > 1/Q^2$ (Nishisato 1980)

Según el criterio $1/Q^2$, Nishisato (1980) sugiere que sólo son relevantes las J = 20 dimensiones con valor propio mayor a $\lambda > 1/Q^2 = 1/55^2 = 0.00033 = 3.310^{-4}$. Este criterio coincide con el Alpha de Cronbach que, para las mismas 20 dimensiones, tienen un índice de confiabilidad mayor que cero $\alpha_k > 0$ (positivo) y menor igual a cero $\alpha_k \leq 0$ (negativo) para todos los otros casos, ver Tabla 24.

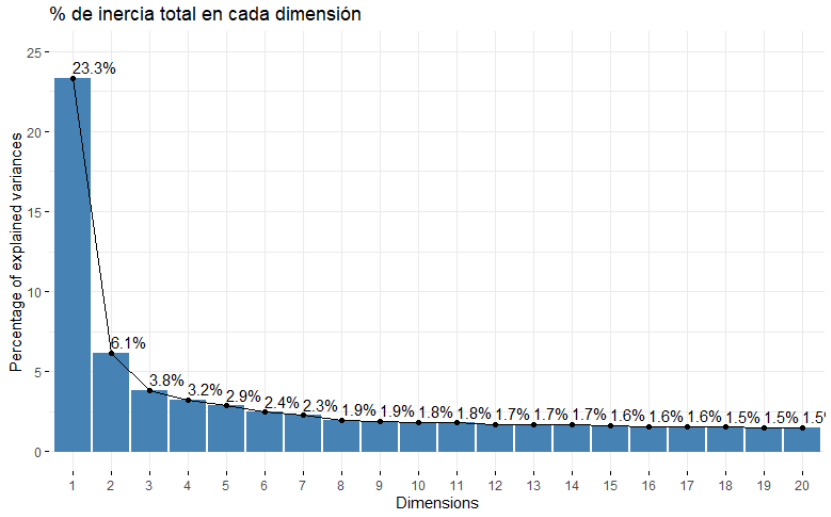


Figura 17: Gráfico de Sedimentación Modelo MCA (Nivel de División: a dos dígitos UNSPSC)

El Gráfico de sedimentación muestra las J = 20 dimensiones con su respectivo porcentaje de varianza explicada (inercia). En la Tabla 23 y Tabla 24 se confirma que la varianza acumulada total con J = 20 dimensiones, explican el 65.8% de la información capturada por el modelo MCA.

Tabla 23: Varianza explicada por el modelo MCA (Nivel de División: a dos dígitos UNSPSC)

dim	1	2	3	4	5	6	7	8	…	16	17	18	19	20
eigenvalue	5.2E-03	1.4E-03	8.5E-04	7.2E-04	6.4E-04	5.5E-04	5.1E-04	4.4E-04	…	3.5E-04	3.5E-04	3.4E-04	3.4E-04	3.3E-04
% of variance	23.3	6.1	3.8	3.2	2.9	2.4	2.3	1.9	…	1.6	1.6	1.5	1.5	1.5
S% of variance	23.3	29.5	33.3	36.5	39.3	41.8	44.0	46.0	…	59.7	61.3	62.8	64.3	65.8

En la *Tabla 24* se obtiene en la dimensión 20 el valor máximo de confiabilidad del coeficiente Alpha de Cronbach acumulado Σα = 3.489, que coincide con el criterio $\lambda > 1/Q^2$.

Tabla 24: Alpha de Cronbach para un Modelo MCA (nivel de división a dos dígitos UNSPSC)

Dimensión	Valor propio λ	% de varianza	Sum % de varianza	Valor Singular $\sqrt{\lambda}$	Alpha de Cronbach α	Σα
dim 1	5.24E-03	23.3	23.3	0.0724	0.763	0.763
dim 2	1.38E-03	6.1	29.5	0.0371	0.520	1.282
…						
dim 18	3.42E-04	1.5	62.8	0.0185	0.018	3.472
dim 19	3.38E-04	1.5	64.3	0.0184	0.012	3.484
J=dim 20	**3.34E-04**	**1.5**	**65.8**	**0.0183**	**0.005**	**3.489**
dim 21	3.27E-04	1.5	67.2	0.0181	-0.005	3.484
dim 22	3.22E-04	1.4	68.7	0.0179	-0.014	3.470
…						
dim 53	1.11E-04	0.5	99.1	0.0105	-0.740	-4.910
dim 54	1.04E-04	0.5	99.6	0.0102	-0.799	-5.710
Q=dim 55	9.12E-05	0.4	100.0	0.0095	-0.921	-6.631

En la Figura 18 podemos ver los pasos que van desde la selección de la muestra de 10.000 observaciones hasta el análisis exploratorio con MCA que reduce la dimensionalidad de 55 variables cualitativas a una matriz de *scores* 20 dimensiones, que capturan el 65,8% de la variabilidad del modelo.

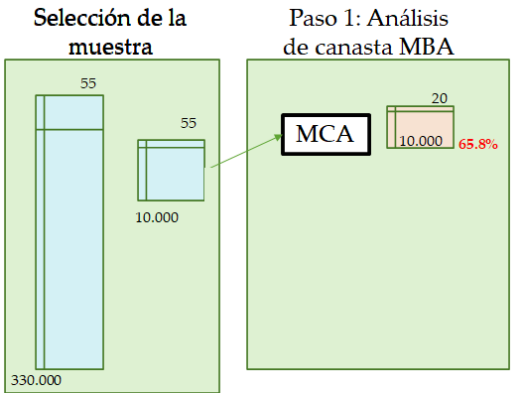


Figura 18: Análisis exploratorio MCA y reducción de la dimensionalidad

4.3.2.2 Gráficos de *loadings* (variables) y *scores* (observaciones) MCA

A partir de los gráficos de *loadings* y *scores* podemos interpretar las variables latentes del modelo MCA. En el gráfico de *loadings* (Figura 19a), las dos primeras dimensiones capturan el 29,5% de la información y están representadas las 55 variables categóricas, de las que se puede interpretar lo siguiente:

- Desde el centro (0,0) hacia la derecha en la dimensión 1, y proyectados sobre toda la dimensión 2, vemos todos **"los productos"** (dentro del rectángulo rojo grande), algunos de ellos con mucha inercia.
- Desde el centro (0,0) hacia la izquierda en la dimensión 1 y proyectados sobre la dimensión 2, vemos **"los servicios"** (dentro del rectángulo rojo pequeño). Los servicios están bien representados por ambas dimensiones, pero con menos inercia que los productos.
- Al centro del gráfico en torno al (0,0), en el círculo en verde, se encuentran representadas las variables de baja inercia que corresponden a la categoría "0: no está presente".

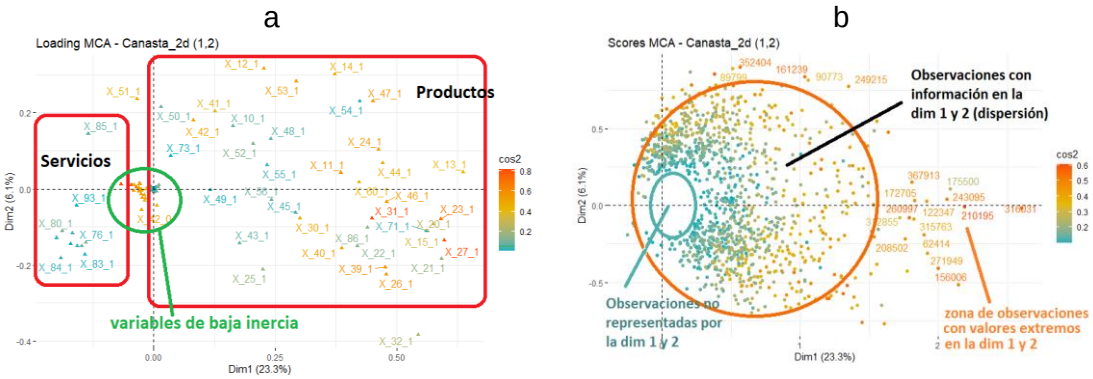


Figura 19: Gráfico de loadings (variables) y scores (observaciones) del modelo MCA en la dimensión (1, 2) - Canasta a dos dígitos UNSPSC. (a) Loading MCA (b) Scores MCA.

El indicador cos2 corresponde a la contribución de calidad que la dimensión aporta a las variables en (a) y a las observaciones en (b). La figura 19b de scores sólo grafica en ambas dimensiones los puntos de calidad mayor a 0.1 (cos2>0.1) para evitar la superposición de puntos de baja calidad de estas dimensiones.

Nota: Se debe interpretar correctamente el circulo verde, ya que el modelo MCA no captura información de la variable categórica “0: no está presente” porque tiene baja inercia respecto de la variable categórica “1: sí está presente” que tiene inercia media o alta y se representa por ambos rectángulos rojos de productos y servicios del gráfico de *loadings*.

En el gráfico de *scores* (Figura 19b) podemos ver que en el centro de ambas dimensiones (0,0) está la zona de observaciones que no están representadas (círculo pequeño en verde); dentro del círculo grande (en rojo) están representadas las observaciones que muestran una buena dispersión en la dimensión 1 y 2; fuera del círculo rojo están las observaciones extremas que están compuestas en su mayoría por productos con alta inercia en una o en ambas dimensiones.

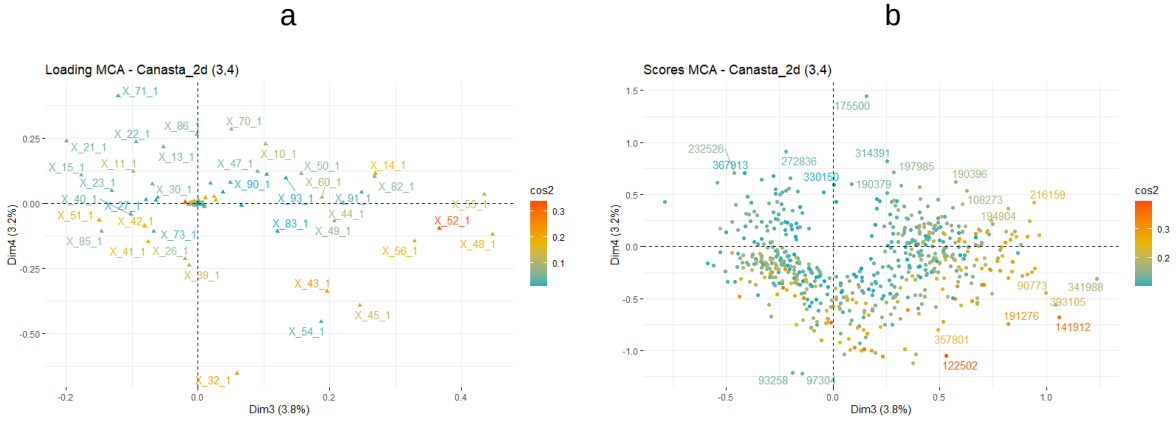


Figura 20: Gráfico de loadings (variables) y scores (observaciones) del modelo MCA en la dimensión (3,4) - Canasta a dos dígitos UNSPSC. (a) Loading MCA (b) Scores MCA.

El indicador cos2 corresponde a la contribución de calidad que la dimensión aporta a las variables en (a) y a las observaciones en (b). La figura 20b de scores sólo grafica en ambas dimensiones los puntos de calidad mayor a 0.1 (cos2>0.1) para evitar la superposición de puntos de baja calidad de estas dimensiones.

4.3.3 Paso 2: HCPC - Segmentación de Canasta a 2 dígitos UNSPSC

Este apartado está enfocado en obtener y validar un modelo de segmentación jerárquica de licitaciones a dos dígitos del código UNSPSC. La función HCPC de R-Studio permite

segmentar la muestra de 10.000 observaciones en 50 tipos de licitaciones (canastas) a partir de la matriz de *scores* reducida a J=20 dimensiones que aportan el 65.8% de la información capturada por el modelo MCA. Esta matriz será el elemento de entrada en la función HCPC para realizar una segmentación de 50 clases.

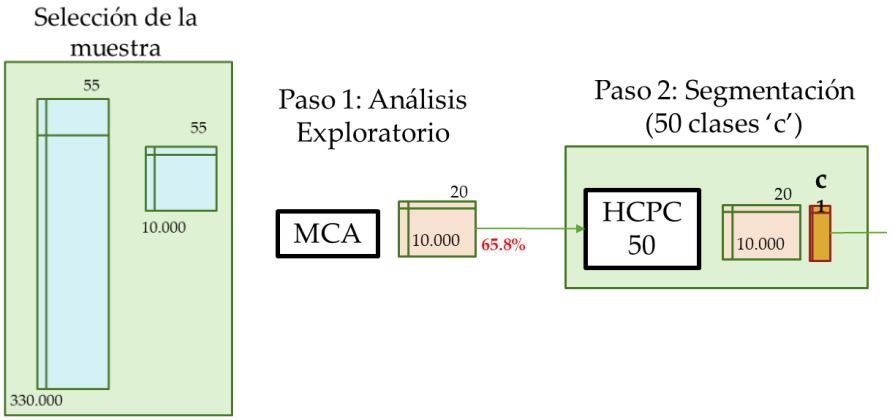


Figura 21: Paso 2 Segmentación de la muestra en 50 clases - Modelo HCPC (Nivel de División: a dos dígitos UNSPSC)

En la Figura 21 vemos que la función HCPC segmenta las 10.000 observaciones de la matriz de *scores* en 50 segmentos o clases; en otras palabras, cada observación de la muestra es asignada a uno de los 50 segmentos y su resultado es un vector columna denominado "c".

(a) (b)

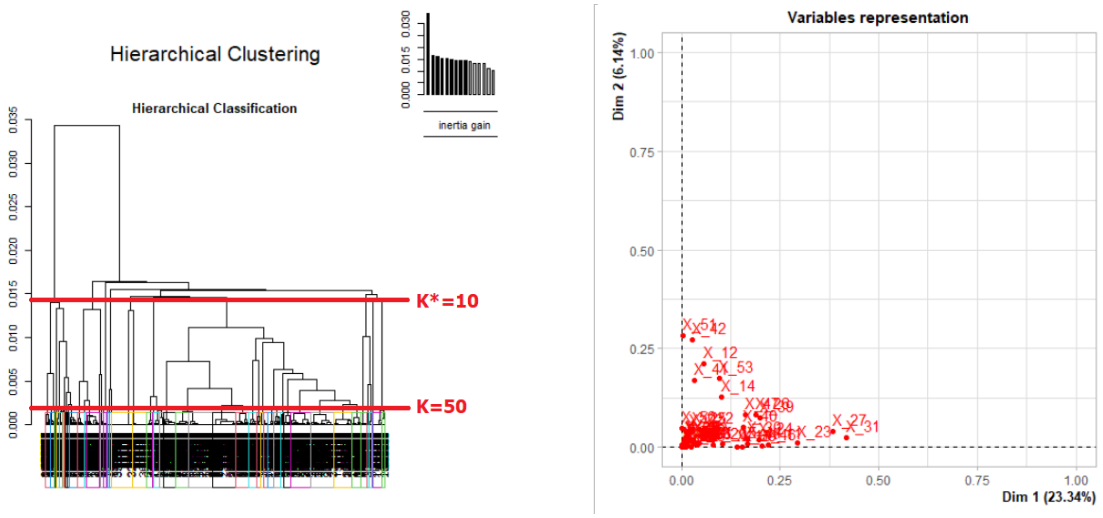


Figura 22: Modelo Clúster Jerárquico HCPC - (Nivel de División: a dos dígitos UNSPSC). (a) dendograma de agrupamiento jerárquico con k=10 y k=50. (b) Gráfico de Contribuciones Dimensión (1,2)*

El Clúster jerárquico HCPC (Figura 22a) tiene la opción de agrupar las observaciones de la matriz de *scores* en K tipos de canastas o prototipos (segmentos). El óptimo del modelo, tal como se ve en el gráfico, es k* = 10 segmentos, pero se necesita al menos K = 50 para forzar al modelo a representar más casos y más combinaciones de productos que se licitan juntos.

En el gráfico de contribuciones (Cos2), Figura 22b, el indicador Cos2 corresponde a cuanto contribuye la dimensión 1 y 2 a la calidad de cada variable. Las variables que están menos representadas por estas dos dimensiones son las más cercanas al centro (0,0); por su parte,

la dimensión 1 aporta una contribución de calidad de 0.45 a las variables X23, X27 y X31; y la dimensión 2 aporta una contribución de calidad de 0.25 a X51, X42 y X12.

La Tabla 25 indica la cantidad de licitaciones de la muestra asignadas a cada segmento; por ejemplo, el segmento c_1 y c_{13} tienen 8 y 735 elementos, respectivamente. Podemos verificar que hay algunos segmentos con pocas observaciones y otros con muchas, el ideal es que esté medianamente equilibrado entorno al promedio de 200 observaciones por segmento, pero eso no depende de la calidad de la segmentación ni de la muestra, sino que de efectivamente hay proporcionalmente pocas licitaciones en ese segmento.

Tabla 25: Segmentación de la muestra en 50 prototipos o canastas (cantidad de licitaciones por segmento)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
8	397	90	617	99	158	266	116	48	351	82	74	735	76	109	544	788	240	219	139	434	310	169	186	157
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
13	194	200	151	274	186	179	218	156	150	191	124	216	92	116	123	207	147	109	167	22	135	105	86	27

Se puede verificar la efectividad de la segmentación del paso 2, comparando tres de los cincuenta segmentos que se han modelado. Por ejemplo, c_7, c_{10} y c_{13} tienen asignadas 266, 351 y 735 licitaciones, respectivamente. Vemos en la Tabla 26 que:

- El segmento c_7 agrupa una canasta prototipo en la que predomina el producto X78 que está presente en 263 licitaciones, de las cuales 14 se licita junto con X25.
- Los segmentos c_{10} y c_{13} tienen dos productos en común que se licitan juntos, X42 y X51 (en color azul), pero ambos segmentos se diferencian en que este par de productos se licitan en combinación con otros productos.
- En general todos los segmentos se caracterizan en que predomina uno o dos productos, y se licitan acompañados con otro de menor frecuencia. Por ejemplo, en C_7 predomina X78 y en ocasiones se acompaña con X25, en C_{10} predomina X85 y en ocasiones se acompaña con X41, X42 y X51. Finalmente, en C_{13} predomina X51 que se acompaña con X11 y X42.

Tabla 26: Canasta de productos del segmento c_7, c_{10} y c_{13}.

C_7 #266		C_{10} #351				C_{13} #735		
X_25	X_78	X_41	X_42	X_51	X_85	X_11	X_42	X_51
14	263	42	77	121	338	166	196	597

Una vez caracterizados los 50 segmentos es interesante saber qué dimensiones aportan más información a cada segmento. Sabemos que la información del segmento está dispersa en varias dimensiones, pero no todas son significativas y no todas aportan información en la misma magnitud. A continuación, se muestran las primeras 5 dimensiones más significativas (con p-value ≈ 0) que más información aportan a estos tres segmentos c_7, c_{10} y c_{13}:

Tabla 27: dimensiones significativas para los segmentos c_7, c_{10} y c_{13}

C7 #266	v.test	C_{10} #351	v.test	C13 #735	v.test
Dim.7	40.6	Dim.18	48.4	Dim.49	17.6
Dim.31	40.1	Dim.24	24.6	Dim.2	15.2
Dim.34	28.4	Dim.27	24.3	Dim.54	14.7
Dim.32	28.1	Dim.30	23.9	Dim.53	12.7
Dim.10	12.3	Dim.22	16.9	Dim.11	8.4

El estadístico “v.test” de cada dimensión, es una prueba estadística que sigue una distribución normal, que puede considerarse significativa cuando toma un valor en módulo mayor a 2 ($|v.test| > 2$). Esto significa e interpreta que el segmento en cuestión tiene una coordenada significativamente diferente a cero en esa dimensión, es decir, la coordenada en esa dimensión contiene información de ese segmento.

Esto es útil para saber qué segmentos tienen grandes valores positivos o negativos en cada dimensión. Por ejemplo, estadísticamente la dimensión 7, 31, 34, 32 y 10 tienen mucha información asociada al segmento c_7. Pero debemos tener en cuenta que Nishisato considera relevantes sólo las primeras J=20 dimensiones porque la confiabilidad en ellas es positiva, en consecuencia, para el segmento c_7 debiéramos usar al menos las dimensiones 7 y 10.

4.3.4 Paso 3: SVM - Clasificador de Licitaciones SVM a dos dígitos UNSPSC

Este apartado tiene como objetivo modelar un clasificador que permita estimar una clase ‘$\hat{c}_k$’ para cada observación de la matriz de *scores* que previamente ha sido asignada a un segmento ‘c_k’ en el paso 2. Usamos la función ***ksvm*** de la librería "*kernlab*" de R-Studio. Es una máquina de soporte vectorial (SVM *Support Vector Machine*), que en modo clasificador requiere dos entradas para ajustar un modelo de clasificación; una es el vector columna ‘c’ que viene de la segmentación provista por la función HCPC del paso 2, y la otra entrada es la matriz de *scores* del modelo MCA del paso 1; el resultado del clasificador SVM es una estimación de la clase $\hat{c}$ (vector columna). La eficiencia del clasificador estará dada por el porcentaje de aciertos en la clasificación. El siguiente diagrama describe el proceso que se sigue desde la selección de la muestra hasta la clasificación.

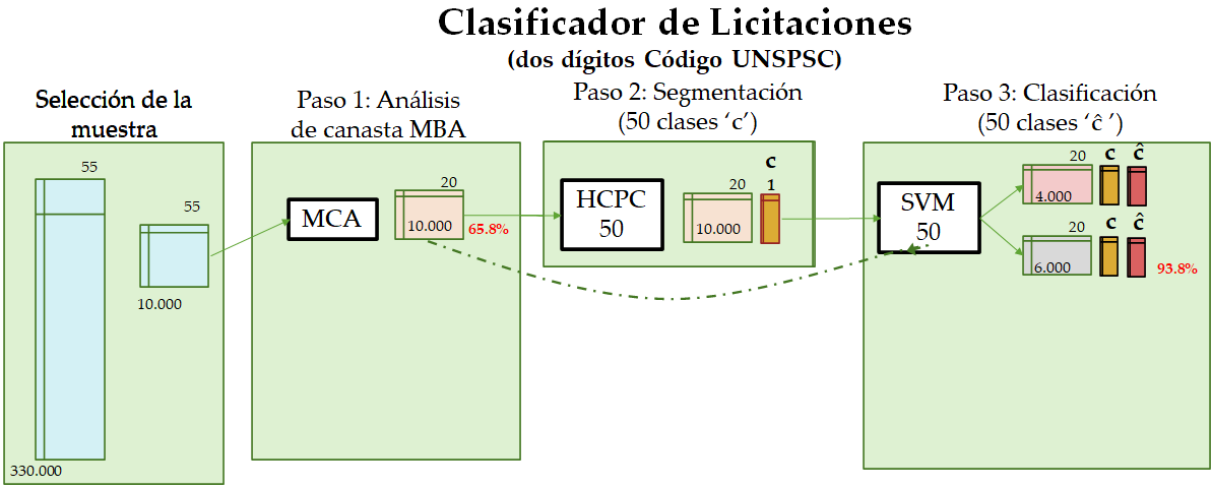


Figura 23: Paso 3: Clasificación de licitaciones - Modelo SVM (Nivel de División: a dos dígitos UNSPSC)

En la *Figura 23* vemos que el análisis exploratorio MCA aporta la matriz de *scores* con sus 20 variables latentes (en el paso 1), y la segmentación HCPC aporta el vector columna de segmentación ‘c’ (en el paso 2). Una vez ajustado el modelo SVM en el paso 3, se obtiene una estimación de la clase ‘ĉ’ ~ ‘c’ con un porcentaje de aciertos del 93,8%. Vemos también en el paso 3 que para ajustar el modelo de clasificación SVM, la matriz de *scores* ha sido separada en dos submuestras aleatorias para usar una de 4.000 observaciones en el entrenamiento (calibración) y otro de 6.000 para la validación (test).

El modelo de clasificación SVM se puede considerar eficiente y robusto; se comprueba empíricamente en el entrenamiento, cuando se remuestrea desde 1.000 observaciones, se

logra una efectividad sobre el 90,5%. Si el remuestreo es de 5.000 observaciones para el entrenamiento, la efectividad media será 94,3%. Ver Tabla 29 (página 63) para la tasa de aciertos según tamaño de muestra.

Por lo tanto, la sensibilidad del modelo SVM sigue siendo muy alta aun cuando la muestra sea más pequeña (por ejemplo 1.000 observaciones). Esta característica permite suponer que el clasificador con la técnica SVM es eficiente; por ende, sería factible aumentar el número de clases en el paso 2 sin afectar con ello la precisión del clasificador en el paso 3.

4.3.4.1 Clasificador de Licitaciones: Otras técnicas

Otro punto importante, para este paso 3 de clasificación, consiste en descartar el árbol de decisiones *TREE* y *random forest* RF como clasificador alternativo al SVM, dado que la efectividad de ambas técnicas resulta inapropiada para el caso de licitaciones, principalmente porque se presentan dificultades para interpretar los más de 180 nodos que se forman cuando el tamaño de muestra es 4.000 observaciones para la fase de entrenamiento y logra una efectividad máxima del 73% en la validación (ver Figura 24).

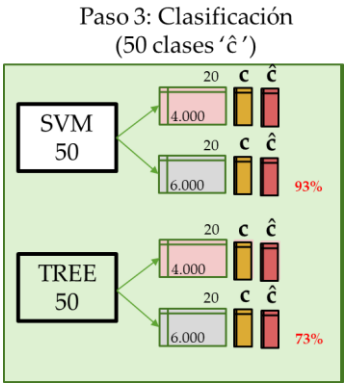


Figura 24: Efectividad del Árbol de Decisión TREE y Random Forest

La sensibilidad y efectividad del árbol de decisión *TREE* y *random forest* RF se ponen en duda cuando, bajo las mismas condiciones iniciales, la efectividad del modelo SVM es superior al *TREE* y al RF en más de un 20%. En efecto, se podrán descartar del todo como modelos alternativos porque, aun cuando se aumente el tamaño de la muestra, la efectividad del modelo *TREE* y RF sigue sin ser bueno. Por ejemplo, cuando la submuestra para el entrenamiento es de 8.000, la efectividad no supera el 80%, y si usamos las 10.000 observaciones de la matriz de *scores* en el entrenamiento, la efectividad no supera el 92% de aciertos (lo que es inapropiado porque no tenemos datos para validar el modelo).

En este sentido, el método de bifurcación y ramificación del árbol de clasificación *TREE* y *random forest* RF, es poco parsimonioso ya que aumenta exponencialmente el número de nodos en la medida que se aumenta el tamaño de la muestra y pierde aún más precisión cuando es necesario realizar una poda para evitar el sobreajuste.

4.3.5 Análisis "No Aciertos" en la clasificación: residuos del modelo SVM

Para analizar los "no aciertos" del modelo de clasificación con SVM, nos centramos en revisar ese 6.4% de observaciones mal clasificadas que corresponden a 384 licitaciones del total de la submuestra de validación. Nos percatamos que coincide con que 108 licitaciones (1.8%) están mal asignadas a la clase "C_{50}" y otras 126 (2.1%) están mal asignadas a las clases "C_{48},

C_{43}, C_{32}, C_{23}, C_{15}". La característica en común de estos residuos, es que son licitaciones de más de 10 productos, siendo que el promedio de las canastas multi productos es de 2,3.

La razón para revisar la tasa de "no aciertos" es para mejorar la efectividad del modelo. Sin embargo, tiene un propósito más práctico; es también para hacer frente al asunto que usualmente en las licitaciones hay una alta probabilidad de encontrarse con licitaciones que forman una combinación de productos que no se ha dado anteriormente.

Sabemos que el inconveniente de un modelo de clasificación, independiente de cuál es la técnica utilizada, es que una observación nueva será siempre asignada a una clase, aun cuando el dato tenga características que no han sido identificadas por el modelo o bien tiene patrones característicos de dos o más clases.

Otras 234 observaciones de la submuestra de validación (3.9%), son casos mal clasificados y corresponden a licitaciones cuya canasta está formada por muchos productos que presentan características similares de dos o más clases; es decir, el dato nuevo tiene características que pertenecen a más de una clase. Por lo cual, el modelo tiene dificultades para asignar a la clase correcta, esto se interpreta como una observación que está próxima a dos o más clases.

4.4 Interpretación y comparación de los resultados

La discusión técnica que se propone en este apartado es la mencionada en el objetivo específico *"3. Plantear y describir los elementos teóricos que permitan cuestionar, promover y validar las características y potencialidades del MCA como herramienta para el análisis de canasta MBA".* La idea va más allá del MCA aplicado a licitaciones visto en este TFM, sino que apunta a identificar los elementos teóricos del MCA que postulan varios investigadores relacionados, y que puede servir para validar su aporte al análisis de canasta de mercado MBA cuando la variable es categórica.

El paradigma inicial del análisis multivariante basado en variables latentes es obtener un *dominio de solución* de *scores* y *loadings*, que capture la máxima variabilidad posible de las Q variables categóricas originales. Para ello tenemos que considerar dos situaciones importantes que hemos trabajado en este documento, no como un objetivo en sí mismo, sino como un propósito que destaque las propiedades el modelo MCA, tal como se describe a continuación:

4.4.1 Relación entre reducción de dimensionalidad y confiabilidad del modelo MCA

El criterio $1/Q^2$ de Nishisato (1980) es ingenioso y facilita el cálculo de la confiabilidad en la ecuación Alpha de Cronbach (1951) con la que se mide el coeficiente de cada variable latente de un modelo MCA. Nishisato usa dos reemplazos en el Alpha de Cronbach, el de varianza $Q^2\sqrt{\lambda_k} = S^2$ y la suma de varianzas de los k ítems $Q = \sum_{k=1}^{Q} S_k^2$, para transformar la Ecuación 3 en la Ecuación 4 (consultar pág. 42).

Luego, al seguir el razonamiento de Cronbach, se seleccionan sólo aquellas dimensiones con $\alpha_k > 0$, condición que Nishisato interpreta cuando se cumple $\sqrt{\lambda_k} > 1/Q$, que toma el nombre de índice de correlación promedio (*average correlation ratio*). Para más detalles consultar la sección 3.3.3: Reducción de la Dimensionalidad sobre Variables Latentes en MCA (pág. 40).

4.4.2 Confiabilidad en el dominio de solución en MCA y PCA

¿Cuál es la máxima confiabilidad Σα del modelo? es 3.489 tanto para PCA como MCA, que se obtiene al seleccionar sólo J variables latentes con $\alpha > 0$ (ver Tabla 28, fila 20 en rojo). El

resultado es un *dominio de solución* representado por la matriz de *scores* y *loadings*, son una consecuencia de la reducción de la dimensionalidad $J = 20 < (T - Q) = (110 - 55)$.

Tabla 28: varianza explicada y Alpha de Cronbach del dominio de solución (a) PCA y (b) MCA

(a)

Dimensión PCA	Valor propio	% de varianza	Sum % de varianza	Valor Singular	Alpha de Cronbach α	Σα
1	3.98	7.2	7.2	1.9955	0.763	0.763
2	2.04	3.7	11.0	1.4289	0.520	1.282
....						
18	1.02	1.9	45.2	1.0087	0.018	3.472
19	1.01	1.8	47.0	1.0058	0.012	3.484
20	**1.01**	**1.8**	**48.8**	**1.0027**	**0.005**	**3.489**
21	0.99	1.8	50.7	0.9975	-0.005	3.484
22	0.99	1.8	52.5	0.9933	-0.014	3.47
53	0.58	1.1	98.0	0.7609	-0.740	-4.910
54	0.56	1.0	99.0	0.7485	-0.799	-5.71
55	0.53	1.0	100.0	0.7246	-0.921	-6.631

(b)

Dimensión MCA	Valor propio	% de varianza	Sum % de varianza	Valor Singular	Alpha de Cronbach α	Σα
dim 1	5.24E-03	23.3	23.3	0.0724	0.763	0.763
dim 2	1.38E-03	6.1	29.5	0.0371	0.52	1.282
...						
dim 18	3.42E-04	1.5	62.8	0.0185	0.018	3.472
dim 19	3.38E-04	1.5	64.3	0.0184	0.012	3.484
J=dim 20	**3.34E-04**	**1.5**	**65.8**	**0.0183**	**0.005**	**3.489**
dim 21	3.27E-04	1.5	67.2	0.0181	-0.005	3.484
dim 22	3.22E-04	1.4	68.7	0.0179	-0.014	3.47
...						
dim 53	1.11E-04	0.5	99.1	0.0105	-0.74	-4.91
dim 54	1.04E-04	0.5	99.6	0.0102	-0.799	-5.71
Q=dim 55	9.12E-05	0.4	100	0.0095	-0.921	-6.631

El *dominio de solución* obtenido a partir de un modelo MCA es equivalente (pero no igual) al obtenido por un modelo PCA, cuando se basan en la misma selección aleatoria de la muestra ¿en qué difieren?:

- El dominio de solución de un MCA captura estructuras de correlación lineales y no lineales, mientras que un PCA captura sólo correlaciones lineales (Greenacre & Blasius, 2006 p. 163). Se puede interpretar de la Tabla 28 que la varianza explicada en la k-ésima dimensión de un PCA es menor la obtenida por un MCA.

- Nishisato encuentra una relación entre los modelos MCA y PCA expresada en la suma de correlaciones al cuadrado de la k-ésima variable latente, que toma el valor según la siguiente ecuación:

$$\sum_{q=1}^{Q} r_{qk}^2 = \begin{cases} \lambda_k, PCA \\ Q\sqrt{\lambda_k}, MCA \end{cases} \qquad \text{(Ec. 7)}$$

- A partir de la Ecuación 7, la confiabilidad del modelo con $\alpha > 0$ da lugar al criterio de Nishisato (1980) para MCA (Ec. 8) y al de Kaiser (1960) para PCA (Ec. 9) cuando las variables han sido centradas y escaladas. A partir de estas ecuaciones podemos interpretar los criterios para seleccionar las variables relevantes en cada modelo.

$$\alpha_k = \frac{Q}{Q-1}\left(1 - \frac{1}{Q\sqrt{\lambda_k}}\right) > 0 \qquad \text{(Ec. 8)}$$

$$\sqrt{\lambda_k} > \frac{1}{Q}\ (Nishisato)$$

$$\alpha_k = \frac{Q}{Q-1}\left(1 - \frac{1}{\lambda_k}\right) > 0 \qquad \text{(Ec. 9)}$$

$$\lambda_k > 1\ (Kaiser)$$

4.4.3 Confiabilidad de Cronbach del modelo MCA y PCA son idénticos

Nishisato demuestra que la suma de correlaciones al cuadrado, Ecuación 10, toma un valor en función del k-ésimo valor propio λ_k^{MCA} del modelo MCA o su equivalente valor propio λ_k^{PCA} de un modelo PCA. Se confirma además que la igualdad en ambos valores se consigue cuando se modeliza sobre la misma muestra y las variables en PCA han sido centradas y escaladas (ver Tabla 28).

$$\sum_{q=1}^{Q} r_{qk}^2 = Q\sqrt{\lambda_k^{MCA}} = \lambda_k^{PCA} \qquad \text{(Ec. 10)}$$

Se puede confirmar que el valor calculado para la suma de correlaciones al cuadrado del modelo MCA y PCA es idéntico en ambos casos. Por ejemplo, para el caso de licitaciones, el primer valor singular para MCA es $\sqrt{\lambda_1^{MCA}} = 0.0724$, que multiplicado por Q=55 toma el mismo valor que el primer valor propio del PCA $\lambda_1^{PCA} = 3.98$. Por tanto, reemplazando en la Ecuación 8 y 9 respectivamente, el coeficiente de confiabilidad para la primera dimensión será idéntica para ambos modelos y se repetirá para todas las dimensiones.

4.4.4 Tamaño del dominio de solución MCA y PCA

La relación que se puede inferir de la Ecuación 11, está determinada en función del valor propio de la k-ésima variable latente con una confiabilidad $\alpha_k > 0$, donde el tamaño del dominio de solución del modelo MCA es $Q/\sqrt{\lambda_k^{MCA}}$ veces más pequeño que el de un PCA. Comprobar con los datos proporcionados en la Tabla 28.

$$\frac{Q}{\sqrt{\lambda_k^{MCA}}} * \lambda_k^{MCA} = \lambda_k^{PCA} \qquad \text{(Ec. 11)}$$

4.4.5 Efectividad del clasificador de licitaciones según MCA y PCA (resultado práctico)

Para la modelización se ha considerado la misma muestra de 10.000 licitaciones y 55 productos (codificados a 2 dígitos del código UNSPSC), un análisis de canasta con MCA o PCA (paso 1), segmentación HCPC (paso 2) y clasificación SVM (paso 3)

A continuación, se muestran los gráficos en la Figura 25 con la tasa de aciertos de cada modelización obtenida a partir de 50 muestras aleatorias con reposición de tamaño 1000, 2000, 3000, 4000 y 5000 para la etapa de entrenamiento y de test.

(a) (b)

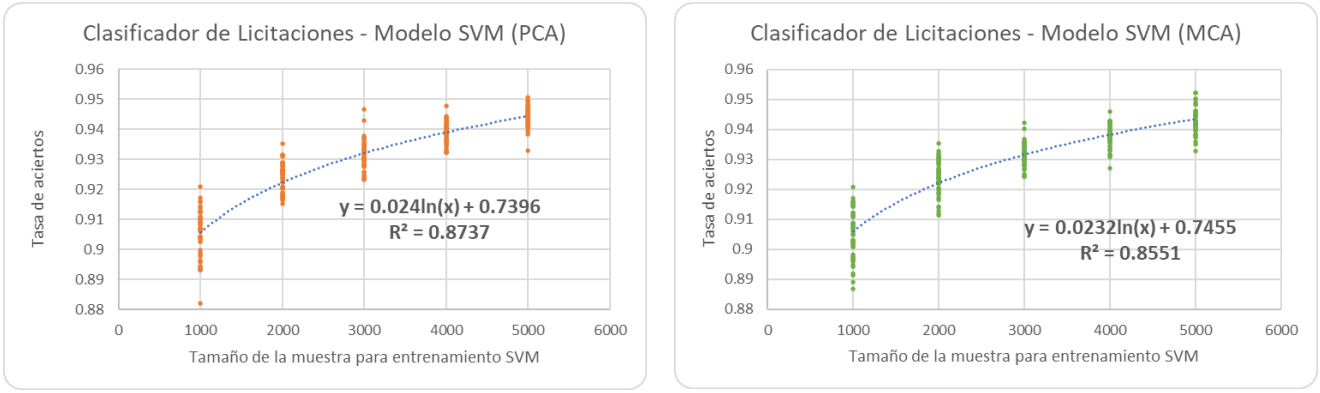


Figura 25: Tasa de aciertos clasificador de licitaciones - (a) Modelo PCA y (b) modelo MCA

Se puede interpretar de los resultados que no hay diferencia entre modelar un clasificador de licitaciones con PCA o MCA; a mayor tamaño de la muestra, ambos modelos mejoran su efectividad de aciertos; en promedio la tendencia es la misma en ambos casos, dado que la curva de regresión se ajusta con R^2 0.87 en PCA y 0.85 en MCA (siendo sus coeficientes de regresión similares). Se puede verificar también en la siguiente tabla que para el mismo tamaño de muestra el porcentaje promedio de aciertos es prácticamente igual (difieren en la tercera cifra significativa). El lector debe tener presente que se han realizado en total 500

muestras aleatorias con reposición, 250 para PCA y otras 250 para MCA y de estas corresponden 50 para cada tamaño de muestra.

Tabla 29: Tasa promedio de aciertos clasificador de licitaciones - (a) Modelo PCA y (b) modelo MCA

(a)

Tamaño Muestra (x)	Iteración	Tasa Promedio (y) PCA
1000	50	0.9048
2000	50	0.9243
3000	50	0.9314
4000	50	0.9387
5000	50	0.9443
	250	

(b)

Tamaño Muestra (x)	Iteración	Tasa Promedio (y) MCA
1000	50	0.9049
2000	50	0.9242
3000	50	0.9316
4000	50	0.9378
5000	50	0.9430
	250	

5 CONCLUSIONES

El clasificador de licitaciones multivariante, construido y modelado a partir de 3 años de licitaciones, logra procesar el 100% de casos provistos por una base de datos de 330.000 licitaciones, que abarca todo el ámbito de la contratación pública del Estado de Chile publicado en la plataforma electrónica (*mercadopublico.cl)*.

Este clasificador, a partir de una muestra de 10.000 licitaciones, segmenta y clasifica 50 tipos de canastas logrando una eficiencia promedio sobre un 93,8% cuando la canasta se define con 55 variables codificadas a 2 dígitos del UNSPSC. En la Figura 26 vemos que esta condición se cumple cuando hemos usado MCA para el análisis de canasta MBA, una segmentación jerárquica HCPC y una clasificación con SVM. En el paso 3, podríamos usar redes neuronales que son igualmente eficientes que un SVM y descartamos otras técnicas que son 20% menos eficientes como son los árboles de decisión (*TREE*) y *random forrest* que son basadas en ramificación y bifurcación, por lo generan demasiados nodos y producen baja tasa de aciertos (ineficiencias en la clasificación).

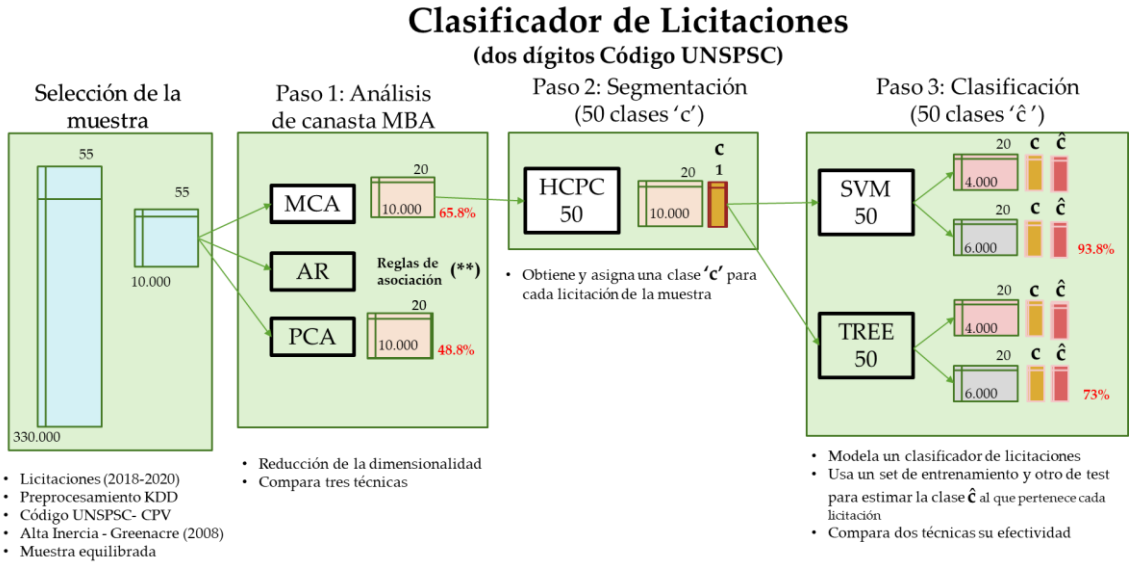


Figura 26: Diagrama de proceso del Clasificador de licitaciones en tres pasos (Nivel de División: a dos dígitos UNSPSC)

Para el caso del análisis de canasta del paso 1, se comparó la eficiencia y los aspectos teóricos de tres técnicas (AR, PCA y MCA), donde el MCA de variable categórica captura un 17% más de variabilidad de los datos que un PCA de variable binaria. Con respecto a AR, está técnica tiene la propiedad de seleccionar un conjunto muy específico de reglas válidas y de buena calidad, pero su resultado no proporciona una visión amplia y general tal como lo hace el MCA o PCA. Sin embargo, la ventaja del AR es que algunas de estas reglas seleccionadas pueden llegar a ser un patrón y un segmento interesante que un MCA o PCA no logra detectar, por lo que la técnica AR no debiera descartarse del todo.

Un elemento importante que se verifica en el punto 5.4.2 de este TFM, es que Nishisato (1980, 1994) determina la relación directa que existe entre los modelos PCA y MCA a través del índice de confiabilidad de consistencia interna (*internal consistency reliability*) o Alpha de Cronbach (1951) con el que se comprueba la confiabilidad del modelo. No obstante, esto aporta más interpretaciones dado que para ambos modelos el Alpha de Cronbach es idéntico en cada una de las k dimensiones, donde para cada k la suma de las Q correlaciones al cuadrado se cumple $\sum_{q=1}^{Q} r_{qk}^2 = Q\sqrt{\lambda_k^{MCA}} = \lambda_k^{PCA}$ (Ec.10). Este notable razonamiento es el que nos

permite validar que la eficiencia del clasificador licitaciones es estadísticamente igual si usamos un modelo MCA o PCA (en el paso 1). Vemos en la figura 27 que las tasas de aciertos promedio son 93.8% y 93.9% respectivamente, por lo que podemos usar indistintamente MCA o PCA ya que ambos conservan la misma información y confiabilidad para construir un modelo de clasificación de licitaciones.

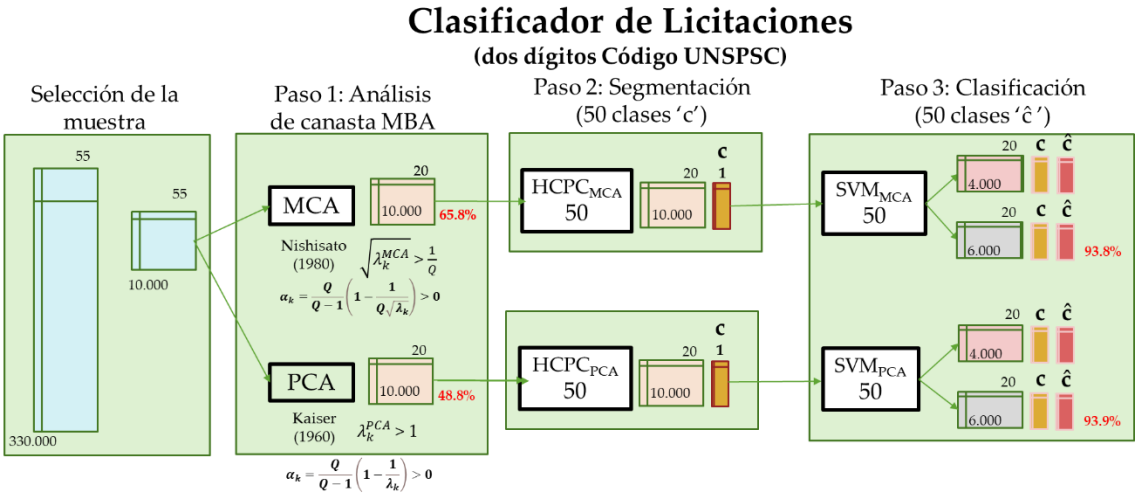


Figura 27: Eficiencia del clasificador de licitaciones SVM_{MCA} vs SVM_{PCA}

Los códigos estándar UNSPSC de la ONU y CPV para la Unión Europea son fundamentales para realizar el análisis multivariante, dado que permiten modelar una canasta de productos y servicios como un vector, donde cada grupo de producto o servicio es un descriptor del vector. De esta manera se pueden aplicar varias técnicas multivariantes al set de datos de licitaciones. Los códigos UNSPSC y CPV son jerárquicos y pueden trabajarse a 1, 2 y 4 dígitos cuando su uso es para realizar análisis y estudios con datos agregados o agrupados. En cambio, trabajar con datos desagregados a 6 y 8 dígitos es lo recomendable cuando el objetivo es más operativo, de gestión o transaccional para identificar con mucho detalle y precisión los productos y servicios de una licitación o canasta.

La estabilidad a largo plazo del proceso de contratación pública por la vía de licitaciones, se basa en tres pilares fundamentales: (1) la **"*teoría de subasta*"** estudiada por Milgron, Wilson, Klemperer y otros autores; (2) las recomendaciones internacionales de **"*gobernanza pública*"** de la OCDE; y (3) las buenas prácticas de la Eurostat para realizar estadísticas a partir de **"*registros administrativos*"**. Estos tres pilares se han trabajado juntos desde hace más de 20 años y ha sido el mecanismo más usado por los países que poco a poco van incorporando en sus procedimientos y procesos las recomendaciones vigentes que validan y trazan la hoja de ruta de su quehacer y que mejora sus estándares de calidad, transparencia y probidad en materia de contratación pública, que visibiliza el gasto público, la confianza de los contribuyentes y la comparación entre países.

El proceso de contratación pública y el clasificador de licitaciones de este TFM, tiene un valor explícito para los 4 actores del ecosistema de licitaciones, es una instancia para acceder a información oportuna, segmentada y ordenada, que de otro modo sería una barrera de entrada difícil de atravesar aun cuando la información siga siendo pública.

6 ANEXOS

6.1 Anexo 1: La maldición del ganador, el escenario sin asimetrías de información en la contratación pública y el valor explícito para los 4 actores del ecosistema de licitaciones.

En el contexto de mercado para la contratación pública, el proceso licitatorio es un mecanismo que promueve un escenario sin asimetrías de información, ya que las bases administrativas y técnicas de una licitación son requerimientos específicos que el postor debe evaluar al momento de ofertar. Esto implica que todos los postores tienen la misma información pública al hacer su oferta, aunque con información incompleta, ya que desconocen la información privada del precio de los demás postores. Es decir, el postor sólo conoce su precio y se expone a ofertar muy por debajo del valor de mercado, reduciendo su margen de utilidad medio, concepto que se conoce con el nombre de "maldición del ganador".

En un mercado competitivo y sin distorsiones, toma sentido el supuesto de que el proveedor se ve obligado a pujar con un precio que evite la "maldición del ganador". El segundo supuesto es que el clasificador podría aportar al postor información oportuna y segmentada de valores estimados de precios de los activos de otros jugadores (Milgrom 1980), logrando reducir la incertidumbre y aumentar su beneficio esperado en cada licitación (Wilson 1960 y 1970) [11].

La clave de ambos supuestos está en la Tabla 30 que muestra la segmentación en 50 grupos de licitaciones y en la Tabla 31 se pueden ver cuáles son los productos y servicios predominantes en tres de los cincuenta segmentos. Con ello, tanto el proveedor como el mandante, según afirma Wilson, podrán estimar con mayor facilidad los valores de precios de los activos del segmento y de los demás postores con la información histórica.

Tabla 30: Segmentación en 50 prototipos o canastas (cantidad de licitaciones por segmento)

Segmentación en 50 tipos de licitaciones (elementos en cada clase)																								
1	2	3	4	5	6	**7**	8	9	**10**	11	12	**13**	14	15	16	17	18	19	20	21	22	23	24	25
8	397	90	617	99	158	**266**	116	48	**351**	82	74	**735**	76	109	544	788	240	219	139	434	310	169	186	157
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
13	194	200	151	274	186	179	218	156	150	191	124	216	92	116	123	207	147	109	167	22	135	105	86	27

Tabla 31: Canasta de productos del segmento c7, c10 y c13

C_7 #266		C_{10} #351				C_{13} #735		
X_25	X_78	X_41	X_42	X_51	X_85	X_11	X_42	X_51
14	263	42	77	121	338	166	196	597

Las externalidades del clasificador se extienden a los demás actores que participan del ecosistema de licitaciones, que mantienen intereses y motivaciones diferentes.

En un contexto de mercado de licitaciones, todos los aspectos que se sugiere en la teoría de subastas, independiente de si el contrato público es subastado o negociado, el clasificador de licitaciones propuesto en este TFM facilita los elementos para construir una herramienta que

[11] El trabajo de Milgrom y Wilson logra el Nobel de economía en 2020 por su contribución a las mejoras en la teoría de subastas (licitaciones) y la invención de nuevos diseños de licitaciones

pueda **estimar** en cada segmento el valor real del precio promedio del activo que se ha adjudicado en el pasado.

Para el proveedor (o postor) es importante estar atento a su segmento de mercado, a su competencia, así como acceder oportunamente a las licitaciones de su sector. Saber cuál es la disponibilidad presupuestaria para el período en curso y conocer el gasto público anual histórico, lo que le permite estimar la demanda del segmento en el que participa.

El clasificador resuelve el gran dilema que enfrenta el postor (proveedor) en cada licitación:

a) dada una licitación que pertenece a un segmento, el postor debe evitar la "maldición del ganador", que consiste en estimar el **precio del activo** al que compra el Estado para poder hacer una buena oferta (puja) que lo sitúe en una posición donde tenga buenas posibilidades de adjudicación, justo por debajo del segundo mejor postor. El valor del activo por lo general es de un "Proyecto Llave en Mano" (Turn-Key Proyect); por ejemplo, debe incluir junto al equipamiento o servicio licitado, una capacitación, instalado en las oficinas o bodegas del comprador, ofrecer pólizas de garantías, seguros, etc. (requisitos que pueden ser técnicos y/o administrativos).

b) Por lo general, el postor tampoco conoce su segmento ni el de su competencia, dado que el precio, la competencia y el mandante de un producto o servicio cambia en función de la ubicación geográfica[12], los requisitos técnico-administrativos, el costo financiero (que está estrechamente ligado a los requisitos técnicos y administrativos, pues de ello deriva el tiempo que demora la entrega del producto, servicio o proyecto) y el diseño de la licitación. El diseño de subastas planteado por Milgrom y Wilson permite bajar las barreras de entrada a nuevos postores idóneos y a poner de relieve los incentivos correctos.

Por su parte, los mandantes (instituciones públicas que licitan en la plataforma de mercado público) tienen especial interés en saber cuántos proveedores potenciales habrá en una licitación que se clasifica en un segmento determinado. Esto es de interés para poner y definir bien los incentivos para atraer a proveedores idóneos, reducir los costes de entrada a la licitación (Milgrom, 2004), verificar cuán solventes son los postores, cuál es su experiencia y si el proveedor está habilitado para ofertar en mercado público.

El clasificador y la teoría de subastas ayudan a los mandantes a realizar un buen diseño de subastas, segmentando y clasificando todos los casos que se dan en la contratación pública por la vía de licitaciones. Permite también explorar un segmento y determinar el potencial de una licitación antes de diseñarla.

Por el lado de las compañías aseguradoras y entidades financieras, éstas tienen especial interés en saber **"en tiempo real"** (*on-line*): ¿qué proveedores se adjudican diariamente?, ¿qué licitaciones?, ¿en qué segmento?, ¿en qué rubros específicos?, etc. Todo ello con el propósito de ofrecer seguros, certificados de fianza, capital de trabajo y créditos de consumo entre otros servicios financieros.

El Estado, en su rol de gestor presupuestario, fiscalizador y controlador de contratos públicos, debe ajustarse a la normativa internacional de lavado de activos y seguir las recomendaciones de la OCDE en términos de transparencia en el gasto público. Este clasificador permite contar

[12] La ubicación geográfica determina un buen contexto para postores dominantes que son más fuertes en una zona determinada con un escenario donde competencia de mercado posee una mayor influencia y que eventualmente podrían acceder a un menor coste medio del activo.

con una herramienta rápida y de amplia cobertura que ayuda al seguimiento y monitoreo de transacciones fraudulentas que salen del patrón normal de compras del Estado.

6.2 Anexo 2: Análisis de Canasta MBA aplicado a licitaciones con reglas de asociación AR

En este anexo se presenta un análisis de canasta de mercado MBA aplicado a licitaciones, sobre la misma base de datos de licitaciones del Estado de Chile entre los años 2018 y 2020, se siguen las referencias bibliográficas del marco teórico y el procedimiento de Cristina Gil Martínez (2020) publicado en https://rpubs.com/Cristina_Gil/Reglas_Asociacion, que se aplica con la librería para reglas de asociación de R-Studio.

El análisis de canasta de licitaciones MBA con AR, se aplicado sobre 46.300 licitaciones que tienen 2 o más productos y servicios (en MBA no tiene sentido analizar canastas de 1 producto). Veremos a continuación las características y sensibilidad del modelo obtenido.

Al definir los parámetros soporte y una confianza, el modelo MBA obtiene varias reglas que cumplen la condición, por ejemplo:

- Si soporte s=0.02 (2%) y confianza c=0.7 (70%) resultan 55 reglas.
- Si soporte s=0.0065 (0.65%) y confianza c=0.7 (70%) resultan 1.721 reglas.

El siguiente ejemplo, que se corresponde con el gráfico de dispersión (Figura 28), muestra un conjunto de solución “válido” que filtra 1721 reglas según los parámetros soporte y confianza ya definidos y fijados, serán los indicadores de calidad propuestos los que dirán cuál de estas reglas serán de buena o mala calidad.

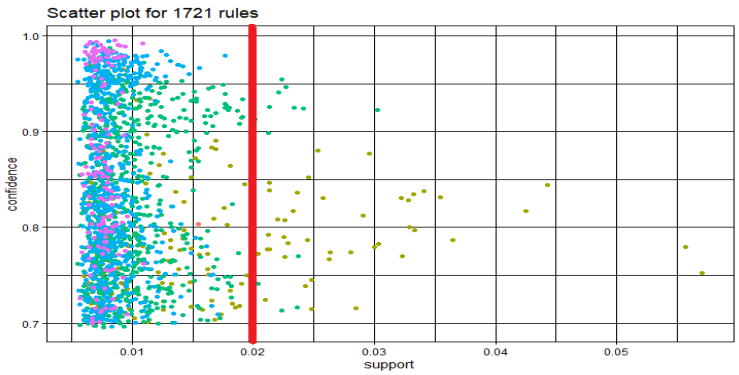


Figura 28: Gráfico de Reglas de Asociación (Soporte y Confianza) - aplicadas a licitaciones - MBA

En el gráfico de dispersión se observa cómo distribuye cada regla según el soporte y confianza calculado. Vemos que la densidad de puntos es mayor cuando el soporte es < 0.01 (1%); es menos denso entre 0.01< soporte < 0.02; por lo tanto, es evidente que a mayor soporte encontraremos menos reglas, pero siempre con una confianza mayor a 0.7 (70%).

El algoritmo *Apriori* de la librería *arules* de RStudio, después de evaluar los parámetros de soporte y confianza más cuatro indicadores calidad, obtiene los siguientes resultados:

- Obtiene una selección de 1721 reglas, de las cuales sólo 52 son de buena calidad.
- Combina 18 productos sobre canastas de 3 a 7 productos, lo cual implica que en teoría hay un total de 62.832 combinaciones posibles,

- Constatamos que 37 productos no generan reglas dado el soporte y la confianza definido. Por tanto, hay una gran cantidad de combinaciones que según esta condición no aportan valor.

En la práctica la conclusión es obvia, la selección de reglas NO permite una visión amplia de lo que ocurre con todas las canastas y la combinación de todos sus productos, dado que hay 37 = 55 -18 productos que no conducen a generar reglas. Sin embargo, se aprecia la efectividad y gran nivel de detalle de las AR para capturar 52 reglas de buena calidad sobre las más de 118 millones de combinaciones posibles.

En la Tabla 32 de frecuencias vemos que las reglas tienen de 3 a 7 productos en la canasta, incluidos el antecedente y el consecuente, cuando el soporte es del 0.65%, y entre 4 a 5 productos cuando el soporte es del 2%.

Tabla 32: Antecedente y consecuente de la regla en MBA con AR.

Regla {antecedente} →{consecuente}	Total Productos	Nº reglas Soporte > 0.65%	Nº reglas Soporte > 2.0%
2: $\{x_a, x_b\} \rightarrow \{x_g\}$	3	1	0
3: $\{x_a, x_b, x_c\} \rightarrow \{x_g\}$	4	154	41
4: $\{x_a, x_b, x_c, x_d\} \rightarrow \{x_g\}$	5	765	11
5: $\{x_a, x_b, x_c, x_d, x_e\} \rightarrow \{x_g\}$	6	645	0
6: $\{x_a, x_b, x_c, x_d, x_e, x_e\} \rightarrow \{x_g\}$	7	156	0
		1.721	52

La calidad de las reglas se verifica con el índice *lift* que debe ser muy superior a 1 (*lift*>>1). Por otro lado, para medir que la regla tenga alta probabilidad de ser un patrón se aplica un test de Fischer, donde un p-valor < 0.05 indica que se acepta la hipótesis nula H_0: "la regla es un patrón". En nuestro caso podemos ver en la siguiente tabla que las reglas son sólo de calidad "regular" dado que el *lift* promedio es de 3.4 pero estadísticamente se comprueba que las reglas obtenidas son un patrón porque su p-valor promedio es casi nulo.

Tabla 33: Soporte y confianza de la Regla, estadísticos de calidad

	Support (2%) $s > 0.02 = \frac{926}{46.300}$	Confidence (70%) $c > 0.7 = \frac{926}{1.323}$	coverage	lift	count	test Fisher
Min:	0.020 (2.0%)	0.71 (71%)	0.023	1.9	935	0E+00
1st Qu.:	0.022 (2.2%)	0.77 (77%)	0.027	3.2	1034	0E+00
Median:	0.024 (2.4%)	0.81 (81%)	0.032	3.4	1126	0E+00
Mean:	0.028 (2.8%)	0.81 (81%)	0.034	3.4	1288	1E-190
3rd Qu:	0.031 (3.1%)	0.84 (84%)	0.039	3.6	1438	0E+00
Max:	0.057 (5.7%)	0.96 (96%)	0.076	5.4	2621	6E-189

La ventaja de minería de reglas de asociación es la especificidad y el gran detalle para encontrar los patrones de compra (qué productos se venden juntos en una transacción); sin embargo, ese es también su inconveniente si el objetivo es un análisis exploratorio.

Del razonamiento anterior podemos concluir que se presentan inconvenientes en los siguientes ámbitos: a) la combinatoria, b) la redundancia, c) el auto-contenido y d) la generalidad. Esto lo podemos verificar en la siguiente tabla resumen:

Tabla 34: Resumen de reglas de alta calidad

#	Lhs (antecedente)		Rhs (consecuente)	support	confidence	coverage	lift	count
[1]	{X_27, X_30}	⇒	{X_31}	0.045	0.848	0.052	3.48	2061
[2]	{X_11, X_27}	⇒	{X_31}	0.043	0.823	0.052	3.37	1995
[3]	{X_11, X_30}	⇒	{X_31}	0.057	0.748	0.076	3.07	2621
[34]	{X_27, X_30, X_46}	⇒	{X_31}	0.022	0.956	0.023	3.92	1028
[35]	{X_11, X_27, X_46}	⇒	{X_31}	0.022	0.941	0.023	3.86	997
[36]	{X_11, X_31, X_46}	⇒	{X_27}	0.022	0.714	0.030	5.31	997
[37]	{X_11, X_30, X_46}	⇒	{X_31}	0.023	0.947	0.025	3.88	1078
[38]	{X_11, X_31, X_46}	⇒	{X_30}	0.023	0.772	0.030	4.18	1078
[40]	{X_11, X_27, X_30}	⇒	{X_31}	0.030	0.927	0.032	3.80	1390

a) La combinatoria: a mayor cantidad de productos habrá un aumento considerable en la combinatoria de pares de productos que dificulta los cálculos.

b) La redundancia: vemos que las reglas [35, 36] son redundantes, dado que son los mismos 4 productos, pero con distinto soporte, confianza y lift, según el antecedente y el consecuente.

c) El autocontenido: podemos ver que las reglas [1, 2, 3] es un subconjunto y están autocontenidas en las reglas [34, 35, 36, 37, 38, 40].

d) La generalidad: podemos ver que las 52 reglas resultantes se forman con 18 productos en canastas de 5 productos como máximo. Por lo tanto, hay 37 productos que no generan reglas con el soporte y la confianza definida, esto no permite una visión amplia de lo que ocurre con todas las canastas, todos sus productos y todas sus transacciones.

En resumen, el análisis de canasta con minería de reglas de asociación MBA permite obtener patrones a partir de la selección de un pequeño conjunto de reglas con productos que se relacionan fuertemente en una transacción, compra, ticket o licitación.

El MBA se limita a transacciones de más de dos productos.

Por otro lado, el alto nivel de detalle que alcanza un MBA (alta precisión) puede ser poco práctico cuando se necesita un análisis exploratorio amplio y generalizado que permita obtener información referente de todas las transacciones, todos los productos y todas las canastas.

Sin embargo, la potencialidad del MBA tradicional está en que una o varias reglas pueden ser un segmento relevante que debe ser identificado en el proceso licitatorio, pero es posible que no sea detectado (*) por el análisis exploratorio de los modelos MCA o PCA y su posterior segmentación y clasificación.

(*) una regla obtenida mediante un MBA puede ser muy específica, absorbida por un segmento, o bien dividida en dos o más segmentos obtenidos a partir de un PCA o MCA y quede oculto junto con otras canastas.

Por ejemplo, puede ser relevante identificar el segmento formado por todos los productos contenidos en las reglas [34, 35, 36, 37, 38, 40] ya que es posible que la segmentación a partir de un MCA o PCA no sea capaz de encontrar esta combinación de productos, porque los separa en varias clases o son agrupados con otros productos que no forman una regla.

6.3 Anexo 3: Código UNSPSC a dos dígitos

A - Materias primas, productos químicos, papel, combustible
10000000 - Material
11000000 - Materiales de minerales y tejidos y de plantas y animales no comestibles
12000000 - Productos químicos incluyendo los bio-químicos y gases industriales
13000000 - Resina y colofonia y caucho y espuma y película y materiales elastoméricos
14000000 - Materiales y productos de papel
15000000 - Combustibles
B - Herramientas y equipos industriales
20000000 - Maquinaria de minería y perforación de pozos y accesorios
21000000 - Maquinaria y accesorios para agricultura
23000000 - Maquinaria y accesorios de fabricación y transformación industrial
24000000 - Maquinaria
26000000 - Maquinaria y accesorios para generación y distribución de energía
27000000 - Herramientas y maquinaria en general
C - Suministros y componentes
30000000 - Componentes y suministros de fabricación
31000000 - Componentes y suministros de fabricación
32000000 - Componentes y suministros electrónicos
39000000 - Suministros
D - Suministros y equipos de construcción, edificaciones y transportes
22000000 - Maquinaria y accesorios para construcción y edificación
25000000 - Vehículos comerciales
40000000 - Sistemas
95000000 - Terrenos, edificios, estructuras y vías
E - Productos farmacéuticos, y suministros y equipos de ensayo, de laboratorio y médicos
41000000 - Equipo de laboratorio
42000000 - Equipo
51000000 - Medicamentos y productos farmacéuticos
F - Suministros y equipos de servicios, limpieza y comida
47000000 - Equipo y suministros de limpieza
48000000 – Maquinaria
50000000 – Alimentos
G - Suministros y equipos tecnológicos, de comunicaciones y de negocios
43000000 - Telecomunicaciones y radiodifusión de tecnología de la información
44000000 – Equipo
45000000 - Equipo y suministros de imprenta
55000000 - Productos publicados
H - Suministros y equipos de defensa y seguridad
46000000 - Equipos y suministros de defensa
I - Suministros y equipos de consumo, domésticos y personales
49000000 – Equipos
52000000 – Muebles
53000000 – Ropa
54000000 - Productos para relojería

56000000 - Muebles y mobiliario
60000000 - Instrumentos musicales
J - Servicios
64000000 - Instrumentos financieros, productos, contratos y acuerdos
70000000 - Servicios de contratación agrícola
71000000 - Servicios de perforación de minería
72000000 - Servicios de construcción y mantenimiento
73000000 - Servicios de producción y fabricación industrial
76000000 - Servicios de limpieza industrial
77000000 - Servicios medioambientales
78000000 - Servicios de transporte
80000000 - Servicios de gestión
81000000 - Servicios basados en ingeniería
82000000 - Servicios editoriales
83000000 - Servicios públicos y servicios relacionados con el sector público
84000000 - Servicios financieros y de seguros
85000000 - Servicios sanitarios
86000000 - Servicios educativos y de formación
90000000 - Servicios de viajes
91000000 - Servicios personales y domésticos
92000000 - Servicios de defensa nacional
93000000 - Servicios políticos y de asuntos cívicos
94000000 - Organizaciones y clubes

6.4 Anexo 4: Divisiones del CPV de 2008 a dos dígitos

03000000-1 Productos de la agricultura, ganadería, pesca, silvicultura y productos afines
09000000-3 Derivados del petróleo, combustibles, electricidad y otras fuentes de energía
14000000-1 Productos de la minería, de metales de base y productos afines
15000000-8 Alimentos, bebidas, tabaco y productos afines
16000000-5 Maquinaria agrícola
18000000-9 Prendas de vestir, calzado, artículos de viaje y accesorios
19000000-6 Piel y textiles, materiales de plástico y caucho
22000000-0 Impresos y productos relacionados
24000000-4 Productos químicos
30000000-9 Máquinas, equipo y artículos de oficina y de informática, excepto mobiliario y paquetes de software
31000000-6 Máquinas, aparatos, equipo y productos consumibles eléctricos; iluminación
32000000-3 Equipos de radio, televisión, comunicaciones y telecomunicaciones y equipos conexos
33000000-0 Equipamiento y artículos médicos, farmacéuticos y de higiene personal
34000000-7 Equipos de transporte y productos auxiliares
35000000-4 Equipo de seguridad, extinción de incendios, policía y defensa
37000000-8 Instrumentos musicales, artículos deportivos, juegos, juguetes, artículos de artesanía, materiales artísticos y accesorios
38000000-5 Equipo de laboratorio, óptico y de precisión (excepto gafas)

39000000-2 Mobiliario (incluido el de oficina), complementos de mobiliario, aparatos electrodomésticos (excluida la iluminación) y productos de limpieza
41000000-9 Agua recogida y depurada
42000000-6 Maquinaria industrial
43000000-3 Maquinaria para la minería y la explotación de canteras y equipo de construcción
44000000-0 Estructuras y materiales de construcción; productos auxiliares para la construcción (excepto aparatos eléctricos)
45000000-7 Trabajos de construcción
48000000-8 Paquetes de software y sistemas de información
50000000-5 Servicios de reparación y mantenimiento
51000000-9 Servicios de instalación (excepto software)
55000000-0 Servicios comerciales al por menor de hostelería y restauración
60000000-8 Servicios de transporte (excluido el transporte de residuos)
63000000-9 Servicios de transporte complementarios y auxiliares; servicios de agencias de viajes
64000000-6 Servicios de correos y telecomunicaciones
65000000-3 Servicios públicos
66000000-0 Servicios financieros y de seguros
70000000-1 Servicios inmobiliarios
71000000-8 Servicios de arquitectura, construcción, ingeniería e inspección
72000000-5 Servicios TI: consultoría, desarrollo de software, Internet y apoyo
73000000-2 Servicios de investigación y desarrollo y servicios de consultoría conexos
75000000-6 Servicios de administración pública, defensa y servicios de seguridad social
76000000-3 Servicios relacionados con la industria del gas y del petróleo
77000000-0 Servicios agrícolas, forestales, hortícolas, acuícolas y apícolas
79000000-4 Servicios a empresas: legislación, mercadotecnia, asesoría, selección de personal, imprenta y seguridad
80000000-4 Servicios de enseñanza y formación
85000000-9 Servicios de salud y asistencia social
90000000-7 Servicios de alcantarillado, basura, limpieza y medio ambiente
92000000-1 Servicios de esparcimiento, culturales y deportivos
98000000-3 Otros servicios comunitarios, sociales o personales

CPV por concepto a nivel de Sección y Grupo

Sección A: Materiales
Grupo A: Metales y aleaciones
Grupo B: No metales
Sección B: Aspecto, forma, preparación y acondicionamiento
Grupo A: Aspecto
Grupo B: Forma
Grupo C: Preparación y acondicionamiento
Sección C: Materiales o productos con propiedades o modos de funcionamiento particulares
Grupo A: Materiales o productos con propiedades particulares
Grupo B: Modo de funcionamiento
Sección D: General, administración

Grupo A: Atributos generales y de administración
Sección E: Usuarios o beneficiarios
Grupo A: Usuarios o beneficiarios
Sección F: Uso específico
Grupo A: Uso educativo
Grupo B: Usos de seguridad
Grupo C: Uso para residuos
Grupo D: Uso estacional
Grupo E: Uso postal
Grupo F: Uso para limpieza
Grupo G: Otros usos
Sección G: Magnitudes y dimensiones
Grupo A: Indicación de dimensiones y potencia
Grupo B: Frecuencia
Grupo C: Otras indicaciones
Sección H: Otros atributos para alimentos, bebidas y comidas
Grupo A: Atributos para alimentos, bebidas y comidas
Sección I: Otros atributos para construcción/obras
Grupo A: Atributos para construcción/obras
Sección J: Otros atributos para la informática, las tecnologías de la información o la comunicación
Grupo A: Atributos para la informática, las tecnologías de la información o la comunicación
Sección K: Otros atributos para la distribución de energía y agua
Grupo A: Atributos para la distribución de energía y agua
Sección L: Otros atributos para medicina y laboratorios
Grupo A: Atributos para medicina y laboratorios
Sección M: Otros atributos para el transporte
Grupo A: Atributos para un tipo de vehículo determinado
Grupo B: Características del vehículo
Grupo D: Atributos para el transporte especial
Grupo E: Atributos para el transporte de mercancías especiales
Grupo F: Mediante uso de vehículos
Sección P: Servicios de alquiler
Grupo A: Servicios de alquiler o arrendamiento
Grupo B: Servicios de tripulación, conductor u operador
Sección Q: Otros atributos para servicios de publicidad y asesoría jurídica
Grupo A: Servicios de publicidad
Grupo B: Servicios de asesoría jurídica
Sección R: Otros atributos para servicios de investigación
Grupo A: Investigación médica
Grupo B: Servicios de investigación económica
Grupo C: Investigación tecnológica
Grupo D: Campos de investigación
Sección S: Otros atributos para servicios financieros
Grupo A: Servicios bancarios

Grupo B: servicios de seguros
Grupo C: Servicios de pensiones
Sección T: Otros atributos para servicios de impresión
Grupo A: Servicios de impresión
Sección U: Otros atributos para servicios comerciales al por menor
Grupo A: Servicios comerciales al por menor de productos alimenticios
Grupo B: Servicios comerciales al por menor de productos no alimenticios

6.5 Anexo 5: Análisis de canasta de mercado dinámico, cambios en el patrón de compra D-AR v/s MSPC-PCA y MSPC-MDA

En la propuesta de Kaur et al. (2016), el objetivo principal del MBA en marketing es proporcionar la información para comprender el comportamiento de compra del cliente, lo que puede ayudar al minorista (*retailer*) a tomar decisiones que mejoran el beneficio neto. Para realizar un MBA tradicional hay varios algoritmos, los cuales funcionan con datos estáticos y no capturan variaciones en los datos en función del tiempo.

En el artículo de Kaur et al. (2016) los autores proponen un análisis de canasta de mercado dinámico (D-AR). Es una técnica de minería de reglas de asociación dinámica con un algoritmo que puede ser útil para detectar variaciones del patrón de consumo. Se usa una lógica de reglas colaborativas, logrando una relación fuerte entre los atributos que mejoran los indicadores de calidad “confianza” y “soporte”.

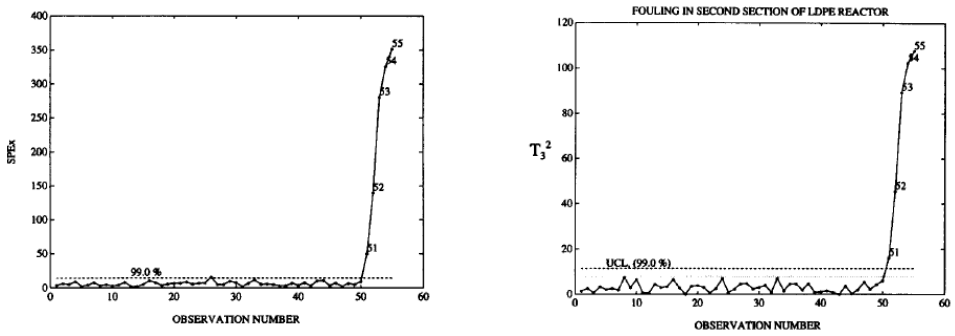
Una alternativa al algoritmo de D-AR es el que se analiza en Ferrer (2007), que propone un monitoreo estadístico de procesos multivariante del tipo MSPC-PCA (Multivariate Statistical Process Control) basado en variables latentes de un PCA, es usado para el monitoreo de variables de un proceso industrial continuo que registran datos de varios sensores e indicadores. La siguiente tabla muestra las diferencias del monitoreo.

Tabla 35: Comparación de tres técnicas de monitoreo estadístico de procesos multivariantes: D-MBA, MSPC-PCA y MSPC-MCA, para el monitoreo de trayectorias de canasta de licitaciones.

Modelo Estático (Análisis exploratorio de canasta)	**Modelo Dinámico** (Monitoreo de trayectoria)
MBA - AR → Variable Cualitativa: Binaria (Ordinal)	D-AR → Trayectoria del patrón
MBA - PCA → Variable Cuantitativa: Continua – Mixta	MSPC-PCA → Trayectoria del segmento
MBA - MCA → Variable Cualitativa: Categórica (Ordinal)	MSPC-MCA → Trayectoria del segmento

En el contexto de big data, a menudo los datos suelen tener alta correlación, deficiencia de rango, baja relación señal / ruido y valores faltantes. Ferrer (2007) discute estos temas y destaca el beneficio de usar MSPC-PCA, detectando cambios en el patrón y trayectoria de las variables del proceso en el instante que aparece un conjunto de datos nuevos con un patrón de consumo que se desvía de la distribución inicial.

El siguiente diagrama muestra el ciclo de un control de procesos MSPC-PCA, que verifica estadísticamente si un dato nuevo está o no dentro de los rangos del patrón del modelo, de modo que al identificarse una secuencia de datos fuera de rango, el control de proceso alertará que el patrón ha cambiado.



b a

Figura 29: Ejemplo MSPC-PCA MacGregor & Kourti (1995) (a) SPE: (sum of squared prediction errors) (b) The Hotelling's T^2.

A modo de ejemplo, los gráficos SPE y T^2, corresponden al proceso estudiado por MacGregor & Kourti (1995) que muestran el cambio del patrón cuando el proceso se sale de control en la observación 51, en que los puntos sobrepasan el límite del 99% definidos por en el modelo. Esto sería equivalente a decir que, en el caso de licitaciones, si un conjunto de licitaciones de un segmento determinado tiene en su canasta una nueva combinación de productos que no ha sido considerada por el modelo, habrá muchos puntos en los gráficos de control del proceso que saltarán sobre los límites definidos por el modelo y se detectará el cambio en el patrón.

Lo propuesto por Kaur, M. et al. (2016) indica que, si queremos detectar cambios en el patrón de compra, el control de procesos por la vía del modelo D-AR dinámico está restringido a transacciones de dos o más productos y a un nivel de detalle predeterminado para los parámetros de soporte y confianza, esto para obtener el subconjunto de las mejores reglas en un MBA tradicional con AR.

En cambio, los modelos MSPC-PCA o MSPC-MCA analizados por Ferrer, A. (2007) y propuesto por MacGregor & Kourti (1995), que están basados en variables latentes, son más amplios y genéricos porque consideran todas las observaciones y todas sus variables que caracterizan todos los segmentos. A consecuencia de ello, se logrará detectar cambios en el patrón de compra en cualquiera de sus variables (en el caso de las licitaciones o canastas serán productos), por lo tanto, el diseño del clasificador de licitaciones será más sensible a los cambios.

7 BIBLIOGRAFÍA

ABDI, H., & VALENTIN, D. (2007). Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, 2(4), 651–657. https://www.utdallas.edu/~Herve/Abdi-MCA2007-pretty.pdf.

ABDI, H., & WILLIAMS, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. https://doi.org/10.1002/wics.101.

AGRAWAL, R., IMIELIŃSKI, T., & SWAMI, A. (1993a). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207-216. https://doi.org/10.1145/170035.170072.

AGRAWAL, R., & SRIKANT, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases*, VLDB, 487-499.

AGRAWAL, R., SWAMI, A., & IMIELINSKI, T. (1993b). Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering, 5*(6), 914–925. https://doi.org/10.1109/69.250074.

BAYARDO, R. J., & AGRAWAL, R. (1999). Mining the most interesting rules. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 145–154. https://doi.org/10.1145/312129.312219.

BRADLOW, E., GANGWAR, M., KOPALLE, P., & VOLETI, S. (2017). The role of big data and predictive analytics in retailing*. Journal of Retailing*, 93(1), 79–95. https://doi.org/10.1016/j.jretai.2016.12.004.

CEPAL (2021). Documento metodológico para el aprovechamiento estadístico de registros administrativos económicos. Disponible online https://repositorio.cepal.org/handle/11362/47461 consultado en mayo 2022.

CHANG, H. J., HUNG, L. P., & HO, C. L. (2007). An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis. *Expert Systems with Applications*, 32(3), 753–764. https://doi.org/10.1016/J.ESWA.2006.01.049.

CHONG, E., STAROPOLI, C., & YVRANDE-BILLON, A. (2013). Auction versus negotiation in public procurement: Looking for empirical evidence. In *The Manufacturing of Markets: Legal, Political and Economic Dynamics*, 120–142. Cambridge University Press (Cambridge UK). https://doi.org/10.1017/CBO9781107284159.009.

DORN, M., JIANG, Z., HOU, W.-C., & WANG, C. F. (2008). An empirical study of qualities of association rules from a statistical view point. *Journal of Information Processing Systems*, 4(1), 27–32. https://doi.org/10.3745/JIPS.2008.4.1.027.

DURÁ JUEZ, P. (2003). *Teoría de subastas y reputación del vendedor*. Comisión Nacional del Mercado de Valores, 3 (Madrid, España).

EUROPEAN COMMISSION (2008). Manual del "Vocabulario Común de Contratos Públicos" (CPV): La contratación pública en la Unión Europea. Disponible online https://simap.ted.europa.eu/documents/10184/36234/cpv_2008_guide_es.pdf. (consultado en mayo 2022).

EUROPEAN COMMISSION (2017). Final report | Revision of CPV "Consultancy Services for Common Procurement Vocabulary expert group. Disponible online https://ec.europa.eu/docsroom/documents/27821 (consultado en mayo 2022).

FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37.

FERRER, A. (2007). Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process. *Quality Engineering*, *19*(4), 311-325.

GAUR, S. S., NARAYANAN, A., & AGRAWAL, R. (2013). Determining customer loyalty: Review and model. *The Marketing Review*, 12(3), 275–289. https://doi.org/10.1362/146934712X13420906885430.

GIL MARTÍNEZ, C. (2020). RPubs - Reglas de asociación. RPUBS.COM. Disponible online https://rpubs.com/Cristina_Gil/Reglas_Asociacion (consultado en mayo 2022).

GOBIERNO DE COLOMBIA (2013). Guía para la codificación de bienes y servicios de acuerdo con el código UNSPSC, V.14.080. disponible online https://colombiacompra.gov.co/sites/cce_public/files/cce_clasificador/manualclasificador.pdf (consultado en mayo 2022).

GREENACRE, M. (2008). *La práctica del análisis de correspondencias*. Fundación BBVA (Barcelona, España).

GREENACRE, M., & BLASIUS, J. (2006). *Multiple correspondence analysis and related methods* (1st Ed.). Chapman and Hall/CRC (London, UK).

GRIVA, A., BARDAKI, C., PRAMATARI, K., & PAPAKIRIAKOPOULOS, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, 100, 1–16. https://doi.org/10.1016/J.ESWA.2018.01.029.

KAUR, M., & KANG, S. (2016). Market Basket Analysis: Identify the changing trends of market data using Association Rule Mining. *Procedia Computer Science*, 85, 78–85. https://doi.org/10.1016/j.procs.2016.05.180.

KARATZOGLOU, A., SMOLA, A., HORNIK, K., & ZEILEIS, A. (2004). kernlab-an S4 package for kernel methods in R. *Journal of statistical software*, *11*(9), 1-20. doi.org/10.18637/jss.v011.i09.

Lalive, R., & Schmutzler, A. (2011). *Auctions vs Negotiations in Public Procurement Which Works Better?.* University of Zurich (No. 23). Department of Economics Working Paper Series, Working Paper. https://doi.org/10.2139/SSRN.1919531.

LÊ, S., JOSSE, J., & HUSSON, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of statistical software*, *2*5, 1-18.

MCAFEE, R. P., & MCMILLAN, J. (1987). Auctions and bidding. *Journal of Economic Literature*, 25(2), 399–738. https://www.jstor.org/stable/2726107.

MACGREGOR, J. F., & KOURTI, T. (1995). Statistical process control of multivariate processes. *Control engineering practice*, *3*(3), 403-414.

MILGROM, P. (2004). *Putting Auction Theory to Work*. Cambridge University Press (Cambridge UK). https://doi.org/10.1017/CBO9780511813825.

NISHISATO, S. (1980). Analysis of categorical data: Dual scaling and its applications. Toronto: *The University of Toronto Press* (Toronto Canada). https://doi.org/10.3138/9781487577995.

NISHISATO, S. (1994). Elements of Dual Scaling: An introduction to practical data analysis (1st ed.). *Psychology Press* (New York).

NISHISATO, S. (2014). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. In Elements of Dual Scaling (1st Ed.). Psychology Press (New York). doi.org/10.4324/9781315806907.

NISHISATO, S. (2022). *Optimal Quantification and Symmetry*. Springer (Singapore). doi.org/10.1007/978-981-16-9170-6.

OECD (2007a). *Competition in Bidding Markets*. Disponible online https://www.oecd.org/competition/abuse/38773965.pdf (consultado en mayo 2022).

OECD (2007b). *Public procurement: the role of competition authorities in promoting competition*. Disponible online https://one.oecd.org/document/DAF/COMP(2007)34/en/pdf (consultado en mayo 2022).

OECD (2015). *Recommendation of the Council on Public Procurement*, OECD/LEGAL/0411. Disponible en https://www.oecd.org/gov/public-procurement/OECD-Recommendation-on-Public-Procurement.pdf (consultado en mayo 2022).

OECD (2019). Reforming Public Procurement: Progress in Implementing the 2015 OECD Recommendation (Executive summary, Key findings), https://doi.org/10.1787/1de41738-en.

PEÑA, D. (2002). *Análisis de datos multivariantes*. McGraw-hill (Madrid, España).

PINHO L., J. (2010). *Métodos de Clasificación Basados en Asociación Aplicados a Sistemas de Recomendación*. Tesis doctoral. Universidad de Salamanca (Salamanca, España). http://hdl.handle.net/10366/83342.

WALLGREN, A., & WALLGREN, B. (2014). *Register-Based Statistics: Statistical Methods for Administrative Data* (2nd ed.). John Wiley & Sons (Chichester, UK).

WOLD, S., ESBENSEN, K., & GELADI, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

YIN, S., & KAYNAK, O. (2015). Big data for modern industry: challenges and trends (point of view). In *Proceedings of the IEEE*, 103(2), 143–146. doi.org/10.1109/JPROC.2015.2388958.