
VNIVERSITAT DE VALÈNCIA

Programa de doctorado en Estadística y Optimización

# Diseño de una Metodología de Análisis Cuantitativo-Cualitativo integrando Métodos Clásicos, fsQCA y Aprendizaje Automático

---

*Autor:*
Acácio Ernesto Dom Luís

*Directores:*
Dr. Rafael Benítez Suárez
Dra. María del Carmen Bas Cerdá

Junio 2025

*Diseño de una Metodología de Análisis Cuantitativo-Cualitativo integrando Métodos Clásicos, fsQ-CA y Aprendizaje Automático* ©
Junio, 2025

Autor:
Acácio Ernesto Dom Luís

Directores:
Dr. Rafael Benítez Suárez
Dra. María del Carmen Bas Cerdá

Universidad:
Universitat de València

# ÍNDICE GENERAL

# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# LISTA DE CÓDIGOS

# LISTA DE ALGORITMOS

# ABSTRACT

This thesis proposes an alternative and integrative methodology of configurational analysis that combines the logical rigor of Qualitative Comparative Analysis (Qualitative Comparative Analysis (QCA)) with the exploratory and computational efficiency of Association Rule Mining (Association Rules Mining (ARM)), incorporating fuzzy logic and multi-criteria ordering techniques. The work begins by recognizing that although QCA, particularly in its fuzzy-set variant (Fuzzy-set Qualitative Comparative Analysis (fsQCA)), has gained prominence in the social sciences for its ability to model complex causalities and multiple pathways to the same outcome (equifinality), it presents substantial limitations. These include high sensitivity to calibration, the forced conversion of fuzzy data into binary values, difficulty in scaling to large datasets, and the lack of rigorous mathematical formalization comparable to modern data analysis methods.

In response to these challenges, the research develops a new mathematical formalization of QCA, addressing both its crisp-set version (Crisp-sets Qualitative Comparative Analysis (csQCA)) and its fuzzy-set variant (fsQCA). It is theoretically demonstrated that solutions derived from csQCA are subsets of those obtained through binary-rule ARM, and that fsQCA, despite its linguistic approach, practically produces Boolean solutions since calibrated data are discretized during the construction of the truth table. Based on this observation, a unified logical-algorithmic framework is proposed that reinterprets the necessary and sufficient conditions of QCA as "IF–THEN" rules, typical of ARM logic.

Building on this new perspective, the thesis develops an alternative model based on fuzzy association rules that preserves the configurational logic of QCA (including consistency and coverage), but with greater semantic expressiveness and computational robustness. A rule minimization algorithm is introduced, based on an adaptation of the Quine–McCluskey method for fuzzy environments, enabling reduction of the combinatorial explosion without sacrificing interpretability. In addition, the multi-criteria method Unweighted Technique for Order of Preference by Similarity to Ideal Solution (uwTOPSIS) is incorporated, which facilitates the prioritization of relevant rules without requiring subjective weights, using metrics such as support, confidence, lift, and coverage.

The thesis also presents custom computational implementations in the `R` language, with reproducible and versatile algorithms applicable to various datasets. These tools were validated through multiple case studies of different nature and scale. Among them is an analysis of internet shutdowns in Africa (with small $N$), identifying causal configurations associated with authoritarian decisions in the face of protests and political tensions. The topic of immigration in international contexts (with large $N$) is also addressed, extracting rules that link social attitudes with factors such as educational level, national identity sentiment, and perception of economic threat. A case study on the fall of democracy in Europe

during the interwar period applied both fsQCA and quantitative rules. An example related to the COVID-19 pandemic demonstrated how uwTOPSIS can be applied.

Finally, the methodology was applied to the real educational context of Licungo University (Mozambique), exploring student satisfaction. Through fuzzy rules, clear patterns were detected: low satisfaction was associated with teacher unpunctuality, inappropriate treatment, and deficiencies in sanitation and infrastructure; while high satisfaction was linked to institutional support, teacher availability, and access to pedagogical resources. This analysis not only empirically validated the model but also provided useful evidence for improving internal policies.

The results show that fuzzy ARM overcomes the limitations of fsQCA, enabling a richer, more robust, and scalable modeling of causal configurations. The proposed mathematical formalization allows, for the first time, the integration of QCA into a coherent algorithmic framework aligned with modern artificial intelligence and data mining techniques. This contribution represents both a theoretical innovation and a practical tool applicable to diverse fields such as education, political science, public health, and social policy analysis. Altogether, the thesis offers a solid, flexible, and forward-looking methodological approach suited to the analytical challenges of the 21st century.

# RESUMEN

Esta tesis propone una metodología alternativa e integradora de análisis configuracional que combina el rigor lógico del Análisis Cualitativo Comparativo QCA con la eficiencia exploratoria y computacional de la Minería de Reglas de Asociación ARM, incorporando lógica difusa y técnicas de ordenación multicriterio. El trabajo parte del reconocimiento de que, si bien el QCA, especialmente en su versión difusa fsQCA, ha ganado notoriedad en las ciencias sociales por su capacidad para modelar causalidades complejas y trayectorias múltiples hacia un mismo resultado (equifinalidad), presenta limitaciones sustanciales. Entre ellas destacan la alta sensibilidad a la calibración, la conversión forzada de datos difusos a valores binarios, la dificultad de escalar a grandes volúmenes de datos y la ausencia de una formalización matemática rigurosa comparable a los métodos modernos de análisis de datos.

En respuesta a estos desafíos, la investigación desarrolla una nueva formalización matemática del QCA, abordando tanto su versión con conjuntos nítidos csQCA como su variante difusa fsQCA. Se demuestra teóricamente que las soluciones derivadas del csQCA son subconjuntos de las obtenidas mediante ARM con reglas binarias, y que el fsQCA, a pesar de su enfoque lingüístico, produce en la práctica soluciones booleanas, ya que los datos calibrados son discretizados en la construcción de la tabla de verdad. A partir de esta constatación, se propone un marco lógico-algorítmico unificado que reinterpreta las condiciones necesarias y suficientes del QCA como reglas del tipo "SI–ENTONCES", típicas de la lógica de la ARM.

Basándose en esta nueva perspectiva, la tesis desarrolla un modelo alternativo mediante reglas de asociación difusas que preserva la lógica configuracional del QCA (incluyendo consistencia y cobertura), pero con mayor expresividad semántica y robustez computacional. Se introduce un algoritmo de minimización de reglas basado en una adaptación del método de Quine–McCluskey para entornos difusos, permitiendo reducir la explosión combinatoria sin perder interpretabilidad. Asimismo, se incorpora el método multicriterio uwTOPSIS, que facilita la priorización de reglas relevantes sin la necesidad de pesos subjetivos, utilizando métricas como soporte, confianza, lift y cobertura.

La tesis también presenta implementaciones computacionales propias en lenguaje R, con algoritmos reproducibles y versátiles, aplicables a diversas bases de datos. Estas herramientas fueron validadas a través de múltiples estudios de caso de diferente naturaleza y escala. Entre ellos se incluye un análisis sobre los cortes de internet en África (con $N$ pequeño), donde se identificaron configuraciones causales asociadas a decisiones autoritarias frente a manifestaciones y contextos de tensión política. También se aborda el tema de la inmigración en contextos internacionales (con $N$ grande), extrayendo reglas que vinculan actitudes sociales con factores como el nivel educativo, el sentimiento de identidad nacional y la percepción de amenaza económica. Un caso de la caída de la democracia en Europa durante

el período de entreguerras aplicó se el fsQCA y reglas cuantitativas. Además, se utilizó un ejemplo relacionado con la pandemia de COVID-19, demostrando cómo uwTOPSIS pude ser aplicado.

Finalmente, se aplicó la metodología al contexto educativo real de la Universidad Licungo (Mozambique), explorando la satisfacción estudiantil. A través de reglas fuzzy, se detectaron patrones claros: baja satisfacción asociada a impuntualidad docente, trato inadecuado, carencias de saneamiento e infraestructura; y alta satisfacción vinculada a acogida institucional, disponibilidad docente y acceso a recursos pedagógicos. Este análisis no solo validó empíricamente el modelo, sino que también aportó evidencia útil para la mejora de políticas internas.

Los resultados muestran que la ARM difusa supera las limitaciones del fsQCA, permitiendo una modelización más rica, robusta y escalable de configuraciones causales. La formalización matemática propuesta permite, por primera vez, integrar el QCA en un marco algorítmico coherente, alineado con técnicas modernas de inteligencia artificial y minería de datos. Esta contribución representa tanto una innovación teórica como una herramienta práctica aplicable a diversos campos como la educación, la ciencia política, la salud pública y el análisis de políticas sociales. En conjunto, la tesis ofrece un enfoque metodológico sólido, flexible y orientado a los desafíos analíticos del siglo XXI.

# RESUM

Aquesta tesi proposa una metodologia alternativa i integradora d'anàlisi configuracional que combina el rigor lògic de l'Anàlisi Qualitativa Comparativa (QCA) amb l'eficiència exploratòria i computacional de la Mineria de Regles d'Associació (ARM), incorporant lògica difusa i tècniques d'ordenació multicriteri. El treball parteix del reconeixement que, si bé el QCA, especialment en la seua versió difusa (fsQCA), ha guanyat notorietat en les ciències socials per la seua capacitat per a modelitzar causalitats complexes i trajectòries múltiples cap a un mateix resultat (equifinalitat), presenta limitacions substancials. Entre aquestes destaquen l'alta sensibilitat a la calibració, la conversió forçada de dades difuses a valors binaris, la dificultat d'escalar a grans volums de dades i l'absència d'una formalització matemàtica rigorosa comparable als mètodes moderns d'anàlisi de dades.

Com a resposta a aquests desafiaments, la investigació desenvolupa una nova formalització matemàtica del QCA, abordant tant la seua versió de conjunts nítids (csQCA) com la seua variant difusa (fsQCA). Es demostra teòricament que les solucions derivades del csQCA són subconjunts de les obtingudes mitjançant ARM amb regles binàries, i que el fsQCA, malgrat el seu enfocament lingüístic, produeix en la pràctica solucions booleanes, ja que les dades calibrades es discretitzen en la construcció de la taula de veritat. A partir d'aquesta constatació, es proposa un marc lògic-algorítmic unificat que reinterpreta les condicions necessàries i suficients del QCA com a regles del tipus "SI–ALESHORES", típiques de la lògica de la ARM.

A partir d'aquesta nova perspectiva, la tesi desenvolupa un model alternatiu mitjançant regles d'associació difuses que preserva la lògica configuracional del QCA (incloent-hi la consistència i la cobertura), però amb major expressivitat semàntica i robustesa computacional. S'introdueix un algoritme de minimització de regles basat en una adaptació del mètode de Quine–McCluskey per a entorns difusos, permetent reduir l'explosió combinatòria sense perdre interpretabilitat. Així mateix, s'incorpora el mètode multicriteri uwTOPSIS, que facilita la priorització de regles rellevants sense necessitat de pesos subjectius, utilitzant mètriques com el suport, la confiança, el lift i la cobertura.

La tesi també presenta implementacions computacionals pròpies en llenguatge `R`, amb algoritmes reproduïbles i versàtils, aplicables a diverses bases de dades. Aquestes eines van ser validades mitjançant múltiples estudis de cas de diferent naturalesa i escala. Entre ells s'inclou una anàlisi sobre els talls d'internet a l'Àfrica (amb $N$ petit), on s'identificaren configuracions causals associades a decisions autoritàries davant manifestacions i contextos de tensió política. També s'aborda el tema de la immigració en contextos internacionals (amb $N$ gran), extraient regles que vinculen actituds socials amb factors com el nivell educatiu, el sentiment d'identitat nacional i la percepció d'amenaça econòmica. Un cas sobre la caiguda de la democràcia a Europa durant el període d'entreguerres va aplicar tant

el fsQCA com regles quantitatives. A més, es va utilitzar un exemple relacionat amb la pandèmia de la COVID-19, demostrant com uwTOPSIS pot ser aplicat.

Finalment, es va aplicar la metodologia al context educatiu real de la Universitat Licungo (Moçambic), explorant la satisfacció estudiantil. A través de regles difuses, es van detectar patrons clars: baixa satisfacció associada a la impuntualitat docent, tracte inadequat, mancances de sanejament i d'infraestructura; i alta satisfacció vinculada a l'acollida institucional, disponibilitat docent i accés a recursos pedagògics. Aquesta anàlisi no sols va validar empíricament el model, sinó que també va aportar evidència útil per a la millora de polítiques internes.

Els resultats mostren que la ARM difusa supera les limitacions del fsQCA, permetent una modelització més rica, robusta i escalable de configuracions causals. La formalització matemàtica proposada permet, per primera vegada, integrar el QCA en un marc algorítmic coherent, alineat amb tècniques modernes d'intel·ligència artificial i mineria de dades. Aquesta contribució representa tant una innovació teòrica com una eina pràctica aplicable a diversos camps com l'educació, la ciència política, la salut pública i l'anàlisi de polítiques socials. En conjunt, la tesi ofereix un enfocament metodològic sòlid, flexible i orientat als desafiaments analítics del segle XXI.

# ACRÓNIMOS

| | |
|---|---|
| **AR** | Association Rules |
| **ARM** | Association Rules Mining |
| **csQCA** | Crisp-sets Qualitative Comparative Analysis |
| **fsQCA** | Fuzzy-set Qualitative Comparative Analysis |
| **MCDA** | Multiple-Criteria Decision Analysis |
| **MDLP** | Minimum Description Length Principle |
| **NPAR** | Negative And Positive Association Rules |
| **QCA** | Qualitative Comparative Analysis |
| **QM** | Quiney-McCluskey |
| **uwTOPSIS** | Unweighted Technique For Order Of Preference By Similarity To Ideal Solution |

# 1

# INTRODUCCIÓN

La investigación en ciencias sociales hoy día debe enfrentar fenómenos que se presentan con una complejidad y un entrelazado crecientes. Frente a ello, el análisis cualitativo resulta esencial: nos brinda la posibilidad de asomarnos a la experiencia y la percepción de los propios actores, así como de captar detalles del entorno que suelen escapar a los métodos estrictamente numéricos. En sus orígenes, este enfoque emergió como respuesta al paradigma racionalista, precisamente para abordar cuestiones que las técnicas cuantitativas no logran esclarecer por completo. Al centrarse en la interpretación de narrativas, significados y procesos, el análisis cualitativo aporta una dimensión de profundidad que complementa de manera natural las evidencias estadísticas (Fernández, 2016).

Inspirado por esa visión configuracional, Charles Ragin presentó a finales de los años ochenta el Análisis Cualitativo Comparativo (QCA). En "The Comparative Method" (Ragin, 1987), Ragin incorporó el álgebra Booleana y la noción de conjuntos al estudio de casos sociales, proponiendo examinar grupos pequeños o medianos para descubrir qué combinaciones de condiciones resultan necesarias o suficientes para generar un determinado resultado. De este modo, el QCA se aleja de la búsqueda de promedios en grandes poblaciones y apunta, en cambio, a explicaciones precisas para cada caso, reconociendo la causalidad compleja y las múltiples rutas que pueden conducir a un mismo fenómeno.

La versión original de QCA es booleana (conjuntos nítidos o "crisp"), conocida como csQCA ("crisp-set QCA"), donde cada condición se codifica como presente (1) o ausente (0). Con el tiempo se desarrollaron extensiones: la mvQCA (multi-valor) para variables categóricas de más de dos categorías, y el fsQCA (fuzzy-set QCA), que permite grados intermedios de pertenencia. En fsQCA, cada caso puede tener una puntuación continua entre 0 (no pertenece) y 1 (pertenencia plena) a un conjunto de interés. Esta flexibilidad dota al analista de mayor matiz al capturar fenómenos sociales donde las condiciones no son simplemente binarias (por ejemplo, niveles de adopción de una política, intensidad de organización colectiva, grados de escolaridad, etc.).

Ambas variantes, QCA *crisp* y *fuzzy*, han mostrado su utilidad en ámbitos contemporáneos. Por ejemplo, en un estudio sobre la innovación y el emprendimiento (Kraus et al., 2018), compilaron 77 artículos que emplean fsQCA en gestión de negocios, revelando un uso creciente de esta técnica en áreas organizacionales. De manera similar, en ciencia política y administración pública se han utilizado métodos configuracionales para analizar la adopción de políticas sociales o la estabilidad institucional (ver p. ej. (Schneider y Wagemann, 2012)). En educación, se han aplicado QCA/fsQCA para comparar escuelas exitosas vs. menos exitosas bajo distintas combinaciones de prácticas pedagógicas y recursos dispo-

nibles (por ejemplo, (Carvalho et al., 2024)). En todos estos casos, estas metodologías han surgido para abordar el carácter múltiple y no lineal de las causas sociales, superando el binarismo rígido de enfoques más antiguos.

Los métodos QCA y fsQCA se basan en principios "cualitativo-comparativos": combinan la riqueza del análisis de casos individuales con una comparación formal entre múltiples casos. Metodológicamente vienen caracterizados por:

- **Teoría de conjuntos y lógica booleana**: QCA emplea la teoría de conjuntos y el álgebra Booleana. Cada caso se considera miembro (o no) de conjuntos definidos por cada condición causal. Esto se formaliza mediante tablas de verdad que muestran, para cada configuración de condiciones, la presencia o ausencia del resultado. De esta forma se identifican condiciones necesarias (siempre presentes cuando ocurre el resultado) y suficientes (cuyo conjunto de casos implica invariablemente el resultado). Como señalan Schneider y Wagemann (2012), estos métodos son "particularmente aptos para identificar lo mínimamente necesario y/o mínimamente suficiente" en conjuntos de condiciones que generan el fenómeno estudiado.

- **Análisis configuracional (equifinalidad)**: QCA se enfoca en configuraciones causales completas, es decir, combina dos o más condiciones para explicar un efecto. Esto reconoce explícitamente que puede haber "múltiples caminos" (equifinalidad) hacia el mismo resultado. Mientras el análisis multivariante clásico estudia el efecto neto de cada variable por separado, QCA pregunta cuáles combinaciones de variables (todas presentes) aparecen juntas en los casos con el resultado (Ragin, 2008). Es decir, el énfasis está en cómo interactúan las variables en cada configuración causal, más que en el peso estadístico de cada una aisladamente.

- **Comparación sistemática de casos**: A diferencia del estudio de caso tradicional puro, que a veces carece de reglas formales para comparar casos, el QCA exige criterios claros y homogéneos de calibración entre casos. Esto mejora la transparencia y replicabilidad del análisis. Por ejemplo, los investigadores deben definir umbrales de pertenencia y reglas para cada condición (qué significa "estar en un conjunto") antes de analizar los datos. De este modo se contrarresta la crítica típica hacia los estudios cualitativos narrativos, según la cual la selección de casos y la delimitación temporal pueden "depender del capricho del investigador". Al formalizar estas decisiones, QCA/fsQCA reduce la subjetividad inherente al análisis de casos únicos (Schneider y Wagemann, 2012).

Como hemos mencionado anteriormente, estas técnicas cualitativo-comparativas permiten revelar configuraciones causales complejas y sistémicas, aportando explicaciones más sólidas que las basadas únicamente en narrativas cualitativas o en correlaciones estadísticas simples. Sin embargo, el análisis cualitativo-comparativo no está exento de dificultades. Uno de los principales problemas metodológicos de QCA y fsQCA es su limitada escalabilidad. Estas técnicas se idearon originalmente para analizar un número reducido de casos (p. ej. N <50), lo que dificulta su aplicación a bases de datos grandes (Elgin et al., 2024). Al

aumentar el número de casos, se complica calibrar las condiciones y categorizar las variables: es más difícil mantener criterios consistentes de pertenencia y se pierde información detallada de cada caso. Finn (2022) señala que cuando crece N, "los investigadores usando QCA tienen menos capacidad de incorporar información detallada de cada caso", por lo cual extender QCA a estudios de large-N se vuelve un 'estiramiento' del método con resultados poco fiables. En la práctica, un alto número de variables incrementa exponencialmente las posibles configuraciones causales y hace inviable la construcción y validación de la tabla de verdad, limitando el análisis a conjuntos de datos de escala moderada.

Esta limitación contrasta con la complejidad inherente de los fenómenos sociales, que suelen depender de múltiples factores interrelacionados. La propia lógica de QCA se fundamenta en reconocer causalidades complejas y conjuntas: diferentes combinaciones de condiciones pueden conducir al mismo resultado (equifinalidad) y las relaciones pueden ser asimétricas. En tal sentido, QCA permite "explicar la compleja complejidad dentro de configuraciones" (Elgin et al., 2024) , capturando matices de causas necesarias y suficientes. Sin embargo, restringir el análisis a pocos casos o variables deja fuera combinaciones causales potencialmente relevantes en contextos amplios. En sistemas sociales reales, donde entran en juego condiciones económicas, políticas, culturales y personales simultáneamente, la escala reducida de QCA choca con la necesidad de modelar interacciones complejas entre muchas variables.

En este contexto, el desarrollo reciente del big data y de las técnicas de aprendizaje automático (machine learning) abre nuevas posibilidades para abordar esa complejidad. Estas herramientas permiten analizar grandes volúmenes de información y detectar patrones que serían imposibles de identificar manualmente. A diferencia de los métodos tradicionales, el aprendizaje automático no requiere asumir relaciones lineales entre variables, sino que es capaz de encontrar combinaciones complejas de factores que explican ciertos comportamientos o resultados. Además, en los últimos años han aparecido algoritmos de aprendizaje automático que permiten interpretar los resultados de forma similar a como lo hace el QCA, es decir, generando reglas o configuraciones causales que pueden ser comprendidas por los investigadores.

Existen ya ejemplos concretos de aplicación de estas técnicas en distintas disciplinas. En economía, se han aplicado algoritmos como los bosques aleatorios o el boosting para predecir indicadores macroeconómicos y detectar fraudes financieros, analizando grandes cantidades de datos (Desai, 2023). Y en el ámbito educativo, el aprendizaje automático se ha usado para predecir el abandono escolar combinando cientos de variables sobre el comportamiento del alumnado en plataformas virtuales. Por ejemplo, Rebelo Marcolino et al. (2025) emplearon CatBoost (https://catboost.ai/) sobre registros de actividad en la plataforma Moodle (número de accesos, tiempo de estudio, interacciones en foros, etc.) junto con datos demográficos de estudiantes, logrando predecir, con alta precisión, el abandono académico. Este caso concreto muestra cómo el ML puede combinar decenas de factores para identificar configuraciones causales del abandono escolar, en un escenario de datos masivos donde el QCA no podría aplicarse directamente.

Todos estos ejemplos muestran que es posible abordar fenómenos sociales complejos utilizando métodos que, aunque distintos en su forma, comparten con el QCA la idea de que no hay una única causa para cada efecto, sino múltiples combinaciones posibles. Por ello, resulta prometedor explorar formas de integrar el análisis cualitativo, el enfoque configuracional del QCA/fsQCA y las capacidades de análisis masivo que ofrece el aprendizaje automático.

De entre todas las herramientas de aprendizaje automático que podemos encontrar, una emerge en lo que parece ser una alternativa casi perfecta del análisis cualitativo comparativo. La **minería de reglas de asociación** (ARM por sus siglas en inglés) surge como un método práctico para descubrir patrones recurrentes en grandes bases de datos, ya sea entre categorías o, tras agrupar y discretizar valores, en variables numéricas. Cuando Agrawal y Srikant la presentaron en los años noventa (R. Agrawal et al., 1993), se centró en datos binarios, como esos análisis de "cesta de la compra", pero pronto se incorporaron técnicas que permiten trabajar directamente con rangos numéricos.

Al igual que en el QCA, esta aproximación basa sus hallazgos en reglas lógicas tipo "si…entonces…", lo que facilita entender por qué se dan ciertas relaciones. Dichas reglas suelen consistir en condiciones conectadas por operadores lógicos "AND" (en castellano "y") que implican un cierto resultado. Por ejemplo, una regla podría decir: "si alguien estudia a tiempo completo (A) y participa en actividades extracurriculares (B) , entonces suele obtener buenas notas (C)", un enunciado sencillo pero cargado de significado. En él, A y B son las condiciones causales y C representa esl resultado.

Esta similitud plantea la posibilidad de tratar problemas que normalmente se tratarían con la metodología y los algoritmos propios del QCA con reglas de asociación. Esto tendría la ventaja de permitir extender los casos de análisis a problemas más complejos, ya sea en cuanto a número de casos (completitud del muestreo) como, sobre todo, a número de variables (complejidad del modelo). Esto es así debido a que la minería de reglas de asociación conforma un conjunto de algoritmos que están ideados para poder funcionar en bases de datos mucho más grandes que las que se usan, de forma habitual, en el análisis cualitativo comparativo.

## 1.1 DESCRIPCIÓN DEL PROBLEMA

Una limitación importante en los métodos de análisis no configuracionales, tales como los métodos de regresión, radica en que se basan en la suposición de relaciones simétricas entre variables (Y. Liu et al., 2017). Esta suposición a menudo introduce sesgos en los resultados del análisis de datos, ya que las relaciones entre las variables son, en muchos casos, asimétricas. Como alternativa para superar esta limitación, tanto el QCA como el método de minería de datos mediante reglas de asociación (Association Rules (AR)) surgen como enfoques prometedores, debido a su capacidad configuracional para tratar con tales asimetrías en diversas áreas del conocimiento.

La teoría configuracional sostiene que diferentes combinaciones de variables pueden conducir al mismo resultado. Por lo tanto, la relación entre un resultado y sus condicio-

nes previas tiende a ser más asimétrica que simétrica (Fiss, 2007; Ragin, 2000; Woodside, 2013). Por ejemplo, distintos países pueden presentar diferentes condiciones que favorecen la emergencia y estabilidad de las democracias. Algunos países pueden considerarse democráticamente inestables hasta que se cumpla una condición específica, aunque esta condición por sí sola no sea suficiente para generar el resultado deseado. Tales relaciones asimétricas y las complejidades combinatorias asociadas no pueden modelarse adecuadamente con métodos convencionales como los modelos de regresión.

En los últimos años, en las ciencias sociales y en el campo de la computación, diversos debates sobre metodologías de análisis de datos han estado en el centro de atención de numerosos investigadores (por ejemplo, (Buxton et al., 2019; Changpetch y Lin, 2013; Ragin, Shulman et al., 2003; Seawright, 2005b; Vis, 2012; J. Zhao y Yan, 2020)). En estos estudios, tanto la minería de reglas de asociación como la metodología QCA son comúnmente señaladas como técnicas alternativas a las estadísticas cuantitativas tradicionales. Schneider y Wagemann (2012) destaca que el QCA es una metodología híbrida, que supuestamente combina lo mejor de dos mundos. Aunque el QCA representa una alternativa viable a los métodos estadísticos, presenta limitaciones al tratar con datos ruidosos o variables lingüísticas. De hecho este fue una de las principales críticas al csQCA original de Ragin (Ragin, 1987). Originalmente, en el QCA todas las variables eran binarias, sin embargo en muchos de los problemas que se intentaban abordar las variables no estaban tan perfectamente definidas de forma que estuviese claro si un caso determinado pertenecía o no a un tipo determinado, sino que la propia definición de la variable permitia *grados de pertenencia* diferentes al 0 o 1. Por ejemplo si una variable es "Corrupción en el gobierno", es lógico pensar que haya países con diferentes grados de corrupción. Por ello se planteo la extensión del QCA original al QCA con conjunto borrosos o fsQCA (fuzzy-set QCA) (Rihoux y Ragin, 2009). En esta extensión, uno de los primeros pasos es la conocida como **calibración**. En este paso se asigna un número entre 0 y 1 (grado de pertenencia) a cada caso en cada variable. Una vez calibrados los datos, se construye una tabla de verdad en el que a los valores superiores a 0.5 se les asigna 1 y a los menores, 0. Esta *de-fuzzyficación* plantea serios problemas en los casos ambiguos, es decir, aquellos que presentan valores cercanos al punto de corte y, por lo tanto, son difíciles de clasificar claramente como 0 o 1. Estos casos pueden generar soluciones menos claras o difíciles de interpretar. Además, la metodología QCA es vulnerable al sesgo en la definición de patrones de suficiencia y necesidad. La calibración de datos puede introducir sesgos si los puntos de corte no son definidos cuidadosamente o si hay sobrecarga de información en un extremo (0 o 1), afectando así la identificación de patrones causales. Otro desafío es la definición de consistencia: para que una solución en QCA sea válida, debe haber consistencia entre las variables calibradas, es decir, que las relaciones causales entre condiciones y resultado sean sostenibles. El QCA enfrenta serios problemas al trabajar con datos contradictorios (Ragin, 1987, 2000; Schneider y Wagemann, 2012).

Adicionalmente, un reto para la metodología fsQCA es la limitación impuesta por la discretización del espacio de propiedades. Cada vértice del espacio corresponde a una combinación específica de valores en la tabla de verdad. Por ejemplo, un espacio con tres

dimensiones tiene ocho vértices, es decir, al trabajar con tres variables, sólo hay ocho combinaciones posibles de condiciones causales. Como consecuencia, fsQCA presenta limitaciones para trabajar con variables lingüísticas, ya que no permite soluciones del tipo:"Si Educación [alta] y Desarrollo [medio], entonces Democracia [media]". Por tanto, la calidad de las soluciones crisp-set-QCA y fuzzy-set-QCA no difiere sustancialmente, lo que hace que la calibración sirva solo para calcular métricas. Cabe destacar que la versión fuzzy de QCA convierte los datos calibrados en valores binarios para construir la tabla de verdad, obteniendo luego condiciones causales crisp (Rihoux y Ragin, 2009).

En esta tesis, buscamos proponer un metodo alternativo para resolver las limitaciones del enfoque anterior de fuzzy-set-QCA propuesto por Rihoux y Ragin (2009) aplicado a problemas en ciencias sociales. Nuestro principal objetivo es demostrar, en primer lugar, que las reglas de asociación en su versión crisp pueden ajustarse para resolver un problema de crisp-QCA. En segundo lugar, proponemos una nueva metodología basada en un modelo fuzzy de minería de reglas de asociación, integrando un método de minimización y ordenamiento de reglas como alternativa al actual enfoque de fuzzy-set-QCA en ciencias sociales. Es decir, buscamos conservar los principios fundamentales del QCA como los conceptos de consistencia, cobertura, análisis de necesidad y suficiencia, al tiempo que resolvemos los problemas de complejidad computacional que presenta actualmente QCA. Para ello, nos centramos en los conceptos originales de fsQCA, por su alineación con el paradigma de las ciencias sociales (Rihoux y Ragin, 2009).

Nuestra hipótesis principal es que, al resolver un problema de QCA con más de dos términos lingüísticos por ejemplo, bajo, medio y alto en su versión fuzzy, las soluciones (conjuntos de condiciones causales) deben expresarse de manera más representativa de estos términos lingüísticos, superando así la ambigüedad de las soluciones generadas por crisp-set-QCA.

El objetivo central de esta tesis es realizar un análisis comparativo entre los métodos QCA en sus versiones crisp y fuzzy-set y el método de minería de reglas de asociación. En particular, discutimos sus similitudes y diferencias. Además, modificamos el algoritmo original de reglas de asociación propuesto por R. Agrawal et al. (1993), introduciendo un algoritmo de minimización de reglas para resolver el problema de explosión combinatoria en el número de reglas, y añadimos el método uwTOPSIS para ordenar las soluciones.

## 1.2 CONTEXTUALIZACIÓN Y JUSTIFICACIÓN

En las últimas décadas, el análisis de datos en las ciencias sociales ha estado dominado por métodos cuantitativos clásicos, como los modelos de regresión, que suponen relaciones simétricas y lineales entre variables. Sin embargo, esta perspectiva ha sido crecientemente cuestionada por su incapacidad para capturar la complejidad causal inherente a los fenómenos sociales. En respuesta a estas limitaciones, han emergido enfoques configuracionales como el Análisis Comparativo Cualitativo (QCA), y métodos alternativos de minería de datos, como las Reglas de Asociación (AR), que permiten una comprensión más

rica de las relaciones causales, especialmente cuando estas son asimétricas, no lineales y multidimensionales.

Este trabajo propone una metodología alternativa al QCA y busca presentar evidencias de que las soluciones de csQCA son subconjuntos de las soluciones derivadas de la minería de reglas de asociación en su versión crisp. Además, se pretende mostrar que la metodología QCA enfrenta dificultades relacionadas con la interpretabilidad en su versión fuzzy, con el desempeño computacional al trabajar con tamaños de muestra y números de variables más grandes, así como con la presencia de ruido en los datos.

Demostramos que las reglas de asociación constituyen una alternativa superior, con varias ventajas frente al QCA. Por ejemplo, QCA es muy sensible a pequeñas alteraciones en los puntos de anclaje utilizados para la calibración de los datos. Además, tiene limitaciones al generar soluciones con más de dos términos lingüísticos, ya que en su versión fuzzy produce soluciones binarias que no son propiamente difusas. En contraste, las reglas de asociación son interpretables y fáciles de aplicar e interactuar.

La metodología QCA, introducida por Ragin (1987), ha sido ampliamente empleada en estudios que buscan identificar combinaciones de condiciones suficientes y/o necesarias para un resultado. Aunque esta técnica permite estudiar la causalidad configuracional bajo los principios de equifinalidad y causalidad compleja, presenta desafíos considerables. Entre ellos destacan la sensibilidad a la calibración, las limitaciones para representar variables lingüísticas de forma expresiva, y la explosión combinatoria de configuraciones posibles al aumentar el número de condiciones.

Los principales aportes de esta tesis se fundamentan en cinco pilares:

1. Probamos que todas las soluciones generadas por las reglas de asociación son superconjuntos de las soluciones obtenidas mediante fsQCA.

2. Proponemos la minimización de reglas de asociación dando una alternativa al algoritmo de Quine-McCluskey para evitar la explosión combinatoria del número de reglas.

3. Introducimos el método uwTOPSIS para seleccionar y ordenar las reglas según distintos criterios.

La generación de reglas en fsQCA se basa en las condiciones de necesidad y suficiencia (Ragin, 2008). Esta lógica de "si y sólo si" implica un grado de determinismo que no es compatible con el alto nivel de heterogeneidad presente en datos microestructurados y en estudios con grandes muestras (large-N) (Chiu y Xu, 2023). Al intentar encontrar soluciones deterministas en presencia de errores probabilísticos, QCA tiende a descartar información relevante, ignorando casos con ruido. Con la adopción del concepto de "consistencia", QCA intenta aproximarse a la lógica determinista al recodificar los resultados, lo cual puede llevar a una pérdida de validez interpretativa.

Además, el algoritmo QCA busca inicialmente identificar un conjunto complejo de reglas que posteriormente se reduce a soluciones parsimoniosas. La parsimonia se refiere al equilibrio entre la complejidad de la regla (es decir, el número de condiciones) y la cantidad de casos cubiertos por ella (Duşa, 2019). Esta característica es fundamental para evitar

el sobreajuste y facilitar la interpretación (Tan et al., 2014). No obstante, a diferencia de las reglas de asociación, en QCA la parsimonia es una etapa secundaria, lo que se traduce en un alto costo computacional, incluso con un número moderado de condiciones.

Así, esta investigación se justifica tanto por su originalidad metodológica como por su contribución a la mejora del análisis configuracional aplicado a fenómenos sociales complejos. Al combinar lo mejor de ambos enfoques la lógica booleana del QCA y la eficiencia computacional de la minería de datos, proponemos una herramienta más robusta, flexible y expresiva para el análisis en las ciencias sociales.

## 1.3 OBJETIVOS

Este trabajo tiene como objetivo general desarrollar, implementar y validar una metodología alternativa al QCA, basada en minería de reglas de asociación, con el fin de superar las limitaciones metodológicas, computacionales y de interpretabilidad inherentes al enfoque tradicional, particularmente en su versión fuzzy.

Para alcanzar este objetivo general, se definen los siguientes objetivos específicos:

1. Demostrar que las soluciones obtenidas por csQCA constituyen subconjuntos de las soluciones derivadas por reglas de asociación.

2. Demostrar que las soluciones obtenidas mediante fsQCA son subconjuntos de las soluciones derivadas a partir de reglas de asociación cuantitativas discretizadas de forma binaria, lo que refuerza su base teórica. Asimismo, evidenciar que fsQCA no genera soluciones verdaderamente difusas, sino esencialmente binarias.

3. Demostrar que las reglas de asociación difusas constituyen una alternativa viable para identificar soluciones verdaderamente difusas en problemas de las ciencias sociales, los cuales la metodología QCA intenta abordar.

4. Desarrollar un algoritmo de minimización de reglas, que permita controlar la explosión combinatoria inherente a la minería de reglas, preservando la calidad explicativa del modelo.

5. Introducir el método uwTOPSIS como mecanismo de ordenación y selección óptima de reglas, utilizando métricas clásicas.

Este trabajo busca validar empíricamente la propuesta metodológica mediante el uso de diferentes conjuntos de datos provenientes de diversos campos del conocimiento, como la ciencia política, la educación, la salud y la informática, comparando el desempeño del nuevo algoritmo con la metodología QCA en términos de calidad de solución, tiempo de ejecución y robustez frente a contradicciones. Asimismo, se busca evaluar la eficiencia computacional del método propuesto utilizando métricas estándar de benchmarking, tales como *speedup*, *sizeup* y *scaleup*, con el fin de evidenciar su aplicabilidad en contextos de Big Data, donde fsQCA no resulta viable. Finalmente, se pretende poner a disposición pública el código fuente del algoritmo en un repositorio abierto (GitHub), promoviendo

la reproducibilidad científica y el acceso libre bajo la Licencia Pública General GNU. Estos objetivos tienen como finalidad ampliar las fronteras metodológicas del análisis configuracional, ofreciendo una alternativa robusta, eficiente y accesible que integra los principios de la lógica QCA con las capacidades exploratorias de la minería de datos.

## 1.4 ESTRUCTURA DE LA MEMORIA

La presente tesis está organizada de manera progresiva, partiendo de los fundamentos teóricos hasta llegar a las aplicaciones empíricas y propuestas metodológicas originales. La estructura se compone de siete capítulos, los cuales se describen a continuación.

En el Capítulo 2, se presenta la base teórica que sustenta esta investigación. Se inicia con una distinción entre métodos configuracionales, como el Análisis Cualitativo Comparativo (QCA), y métodos no configuracionales, como los modelos de regresión. Posteriormente, se profundiza en las variantes del QCA (csQCA y fsQCA) destacando su potencial analítico y limitaciones prácticas, especialmente en lo que respecta a la calibración de datos y su aplicabilidad en contextos a gran escala. A continuación, se introducen los fundamentos de la minería de datos, con énfasis en el algoritmo Apriori y sus extensiones basadas en lógica difusa. El capítulo concluye con un análisis comparativo entre estas metodologías y las técnicas estadísticas clásicas, finalizando con una reflexión sobre la utilidad de la lógica difusa en la modelización de la incertidumbre en las ciencias sociales.

El capítulo 3 establece un puente teórico y metodológico entre el csQCA y las reglas de asociación binarias. Se desarrolla una comparación formal entre ambos enfoques, presentando sus respectivas estructuras matemáticas. Se introduce un teorema de equivalencia que demuestra cómo las soluciones obtenidas mediante csQCA pueden ser interpretadas como subconjuntos de aquellas generadas por la minería de reglas. Además, se discute el concepto de reglas redundantes y se proponen estrategias para su reducción. La sección empírica incluye tres estudios de caso: bloqueos de internet durante elecciones, oposición a la inmigración en Europa y experimentos numéricos simulados, los cuales ilustran las ventajas computacionales e interpretativas del enfoque basado en reglas.

En el capítulo 4, se profundiza en la comparación entre fsQCA y las reglas de asociación cuantitativas. Se discuten técnicas de discretización supervisadas y no supervisadas, con énfasis en el criterio del Principio de Minimum Description Length Principle (MDLP). Este capítulo incluye un estudio empírico centrado en el declino de la democracia entre la Primera y la Segunda Guerra Mundial, mostrando la sensibilidad de los resultados a los métodos de discretización utilizados.

El capítulo 5 propone sustituir el enfoque fsQCA por una metodología basada en reglas de asociación difusas, implementada mediante un algoritmo fuzzy Apriori adaptado. Esta metodología se aplica al análisis de la participación electoral en los países de la OCDE. Tras la extracción de reglas difusas, se introduce un proceso de minimización inspirado en el método de Quine–McCluskey, empleando codificación binaria para simplificar las reglas y mejorar su interpretabilidad.

En el capítulo 6, se presenta un enfoque de clasificación de reglas de asociación mediante métodos de decisión multicriterio no ponderados. Se propone una extensión del método TOPSIS tradicional, denominada Unweighted TOPSIS, que no requiere la asignación previa de pesos a los criterios. Esta metodología se aplica al caso de la pandemia de COVID-19 en el estado brasileño de Espírito Santo, donde se extraen y clasifican reglas basadas en métricas como soporte, confianza, *lift* y simplicidad (longitud). El capítulo incluye además un análisis de sensibilidad, comparando los resultados obtenidos mediante uwTOPSIS con los del TOPSIS clásico, evidenciando la robustez del enfoque propuesto.

El capítulo 7 presenta una aplicación integrada de las reglas de asociación difusas y el método uwTOPSIS en el análisis de la satisfacción con el sistema educativo en Mozambique. Utilizando datos de la Universidad Licungo en la facultad de educación, se explora la percepción estudiantil mediante técnicas de minería difusa, con el objetivo de aportar evidencia empírica para el diseño de políticas educativas más sensibles y adaptativas.

Finalmente, en el capítulo 8, se sintetizan las principales contribuciones teóricas y metodológicas del trabajo, destacando el valor de la integración entre enfoques configuracionales y técnicas de minería de datos en el estudio de fenómenos sociales complejos. Asimismo, se plantean líneas futuras de investigación.

# 2

# ANTECEDENTES

La fundamentación teórica constituye uno de los pilares de cualquier investigación académica rigurosa, siendo esencial para contextualizar los fenómenos estudiados y justificar los enfoques metodológicos adoptados. Este capítulo cumple precisamente con ese propósito, al ofrecer una base conceptual sólida sobre los métodos configuracionales y los métodos cuantitativos de extracción de conocimiento a partir de datos, con especial énfasis en el QCA y en las Reglas de Asociación.

Inicialmente, se discuten las ventajas y limitaciones de los métodos configuracionales, centrándonos en el QCA, contrastándolos con las técnicas basadas en Reglas de Asociación y con los modelos tradicionales de regresión. A continuación, se analizan críticamente los puntos fuertes y las debilidades de cada uno de estos métodos, destacando sus contribuciones específicas y limitaciones contextuales en el tratamiento de datos sociales y conductuales.

El capítulo también presenta una breve retrospectiva histórica, trazando la trayectoria de diseminación de las metodologías QCA y de Reglas de Asociación en distintas áreas del conocimiento. Se busca evidenciar los contextos disciplinares en los cuales estas técnicas han sido más ampliamente empleadas por investigadores de diferentes campos.

Adicionalmente, se discute el papel de la lógica difusa (fuzzy logic) en la modelización de incertidumbres y ambigüedades típicas de las ciencias sociales, demostrando su valor heurístico y analítico en la construcción de inferencias basadas en datos imprecisos o subjetivos.

Finalmente, se presenta la lógica subyacente al algoritmo *Apriori*, uno de los métodos más reconocidos para la extracción de reglas de asociación, detallando su estructura conceptual y su importancia práctica en el contexto del análisis exploratorio de grandes volúmenes de datos.

## 2.1 PERSPECTIVA METODOLÓGICA GENERAL

### 2.1.1 *Métodos configuracionales vs métodos no configuracionales*

El QCA y la minería de reglas de asociación son herramientas metodológicas fundamentadas en la teoría de conjuntos. Una particularidad del QCA es el uso de la tabla de verdad (Schneider y Wagemann, 2012), además del análisis de necesidad y suficiencia (Ragin,

1987). Por otro lado, la metodología de extracción de reglas de asociación no utiliza la tabla de verdad.

Además, el QCA recurre a la minimización lógica basada en los principios del álgebra booleana, proceso mediante el cual las soluciones empíricas son expresadas de forma más parsimoniosa pero lógicamente equivalente, identificando similitudes y diferencias entre casos que comparten el mismo resultado (Duşa, 2019). En contraste, la minimización de reglas en la minería de asociaciones es aún escasa (ver Bedhouche et al. (2023)) y no constituye una etapa central del proceso.

La metodología QCA se basa en conceptos fundamentales de lógica, teoría de conjuntos y álgebra booleana, mientras que la minería de reglas de asociación se sustenta en los principios de la teoría de la probabilidad y la detección de patrones en bases de datos.

Las reglas de asociación pueden interpretarse como un problema QCA al fijar un resultado (*outcome*), aunque sin una interpretación causal explícita. Existen versiones de reglas de asociación que incorporan el tiempo como una variable analítica (Wang et al., 2018). Por otro lado, se han desarrollado esfuerzos para incluir indirectamente el tiempo en la metodología QCA como una dimensión causalmente relevante. Por ejemplo, Mahoney y Rueschemeyer (2003) sentaron las bases conceptuales para combinar explicaciones históricas con razonamientos basados en teoría de conjuntos, mientras que Caren y Panofsky (2005) propusieron modificaciones al algoritmo de QCA para considerar el orden de los eventos como relevante.

La Figura 2.1 presenta una visión general de las metodologías basadas en teoría de conjuntos. Cabe destacar que el término "métodos basados en teoría de conjuntos" abarca tanto enfoques prominentes como menos conocidos en las ciencias sociales y la informática, siendo QCA y las reglas de asociación solo algunos de ellos.

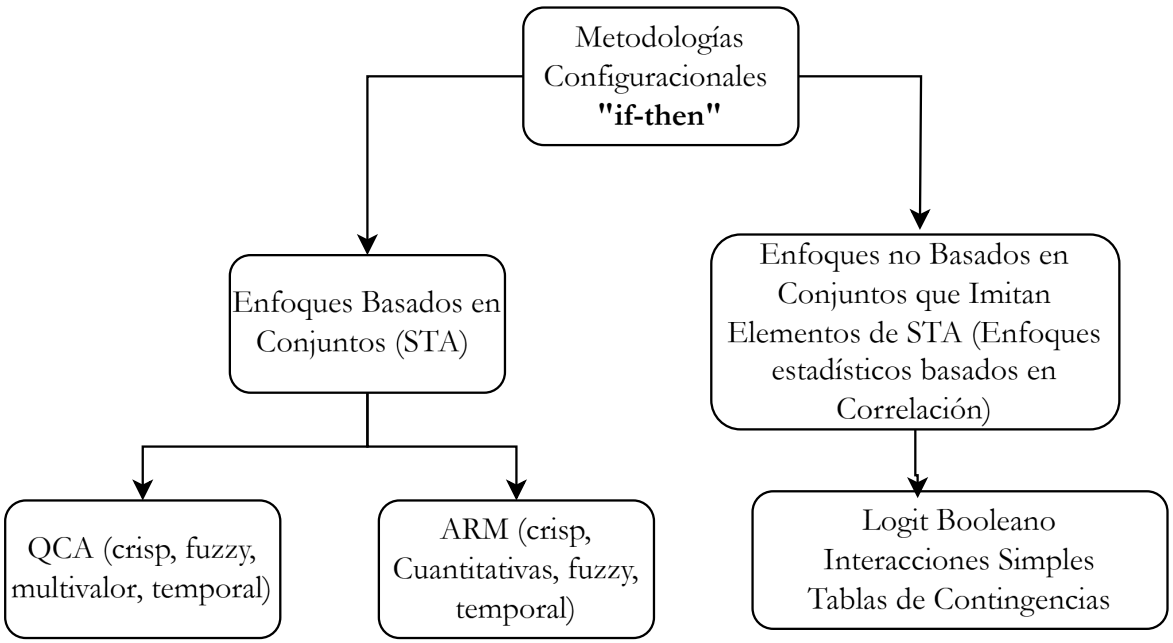


Figura 2.1: Método basado en teoría de conjuntos "If-Then"

En los últimos años, ha habido un intenso debate metodológico sobre el análisis de datos tanto en ciencias sociales como en informática (por ejemplo, Buxton et al. (2019), Changpetch y Lin (2013), Ragin, Shulman et al. (2003), Seawright (2005b), Vis (2012) y J. Zhao y Yan (2020)), donde QCA y minería de reglas de asociación son comúnmente discutidas como alternativas a las técnicas estadísticas convencionales. Según Schneider y Wagemann (2012), el QCA representa una metodología híbrida que intenta combinar lo mejor de los métodos cualitativos y cuantitativos. Sin embargo, esta presenta limitaciones cuando se enfrenta a datos ruidosos o conjuntos de datos con más de 50 casos. Por otro lado, algunos autores destacan algunas inconsistencias en el uso de QCA, como la alta sensibilidad de los resultados ante pequeñas variaciones en la calibración de los datos (ver Chiu y Xu (2023) y Mendel y Korjani (2013)).

Este problema está relacionado con el proceso de transformación de variables cualitativas en valores cuantitativos para el análisis. La calibración, tanto en QCA como en minería de reglas, es crítica: implica convertir variables categóricas u ordinales en valores entre 0 y 1. Entre los problemas comunes del QCA se incluyen:

- **Casos ambiguos:** algunos casos pueden situarse cerca del punto de corte y dificultar su clasificación como 0 o 1, lo que genera soluciones menos claras o más difíciles de interpretar.
- **Sesgo en los patrones de necesidad y suficiencia:** si los puntos de corte no se eligen cuidadosamente, pueden introducirse sesgos que afecten la identificación de condiciones necesarias o suficientes.
- **Problemas de consistencia:** para que una solución sea válida en QCA, debe existir consistencia entre las variables calibradas; esto significa que las relaciones causales deben ser sostenibles. QCA presenta serias dificultades cuando se trabaja con datos contradictorios.

### 2.1.2 *Conjunto de Reglas vs. Regresión*

Una de las principales utilidades de esta tesis radica en que las reglas de asociación ofrecen múltiples ventajas en comparación con el QCA. No obstante, tales ventajas resultarían irrelevantes si los métodos actualmente disponibles en la caja de herramientas de los investigadores que emplean QCA ya fueran capaces de desempeñar de manera más eficaz el rol propuesto por las RA. En esta sección, explicamos las principales limitaciones de los métodos clásicos, los cuales no logran cumplir con esta función, y destacamos las ventajas particulares de los métodos basados en reglas frente a otras aproximaciones.

Cuando realizamos un análisis estadístico tradicional, el objetivo principal es identificar el factor o variable que contribuye significativamente a explicar la variación en la variable dependiente, controlando el efecto de las demás variables. Para ello, se emplean modelos que buscan establecer relaciones causales lineales, basados en tres supuestos fundamentales: unifinalidad, aditividad y simetría (Ragin, 2008). La unifinalidad supone que existe una única trayectoria o camino causal que conduce al resultado estudiado. Es decir, todas

las variaciones de la variable dependiente pueden explicarse por un conjunto específico de factores, sin considerar que diferentes combinaciones puedan producir el mismo efecto.

La aditividad asume que los efectos de las variables independientes se suman de forma lineal e independiente. Esto implica que el impacto de cada variable se considera de forma aislada y luego se suma para explicar el resultado final, sin tener en cuenta posibles interacciones complejas entre ellas.

La simetría supone que los factores que explican la presencia de un determinado resultado también explican, de forma inversa, su ausencia. En otras palabras, lo que causa la aparición de un fenómeno también explica su no ocurrencia cuando dicho factor está ausente.

Sin embargo, muchos fenómenos sociales son complejos y no siguen estas reglas simples. Por ello, los métodos configuracionales proponen un enfoque diferente que rompe con estos supuestos tradicionales. Estos métodos asumen la multifinalidad, es decir, diferentes combinaciones de condiciones pueden conducir al mismo resultado, reconociendo múltiples trayectorias causales (Schneider y Wagemann, 2012; Tan et al., 2014).

También adoptan la configuracionalidad, al considerar que el efecto de una condición depende de la combinación en la que aparece, rechazando la idea de que los efectos de las variables se suman de manera simple.

Además, operan bajo el principio de asimetría, reconociendo que las causas que explican la ocurrencia de un fenómeno no son necesariamente las mismas que explican su ausencia (Ragin, 2000).

Desde el punto de vista algorítmico, tanto el QCA como las RA constituyen formas de aprender conjuntos de reglas a partir de los datos. Una ventaja fundamental de los métodos basados en reglas es su interpretabilidad, como destaca Chiu y Xu (2023). Los métodos que emplean condiciones "si-entonces" son fácilmente comprensibles e interpretables. Para cualquier observación específica, el investigador conoce exactamente la razón de su clasificación positiva o negativa y puede articular claramente dicho razonamiento para el lector. Además, muchas decisiones siguen una lógica basada en reglas. Por ejemplo, un juez puede usar la siguiente regla para decidir si negar la fianza: "Si el acusado tiene antecedentes de no comparecer ante el tribunal y es un riesgo de fuga, o ha cometido un delito violento y representa una amenaza pública, entonces negaré la fianza".

Esta interpretabilidad es una ventaja siempre que el objetivo sea comprender los datos.

Entre los métodos interpretables, los modelos de regresión son los más prominentes en la ciencia política. Estiman el "efecto" marginal (no necesariamente causal) de cada variable sobre el resultado, manteniendo constantes las demás variables. Sin embargo, una limitación común es que no modelan de manera natural la heterogeneidad, ya que asumen que el coeficiente de una variable no depende de los valores de otras variables (Duşa, 2019).

Existen al menos dos razones por las cuales los conjuntos de reglas son más adecuados que la regresión en escenarios donde la variable de respuesta $Y$ es dicotómica, las variables explicativas $X$ son discretas, y las relaciones son complejas. Primero, aunque la inclusión de términos de interacción puede aumentar la flexibilidad de un modelo de regresión, ello

conlleva una pérdida significativa en interpretabilidad en comparación con los conjuntos de reglas. Consideremos un modelo con interacción entre tres variables:

$$E[Y|X_1, X_2, X_3] = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \beta_{2,3} X_2 X_3 + \beta_{1,2,3} X_1 X_2 X_3.$$

Este modelo ya es complejo, aunque manejable si solo interesa una variable específica (lo cual es raro en la fase exploratoria). En cambio, un conjunto de reglas mantiene su claridad incluso con interacciones de orden superior, usando operadores lógicos intuitivos. Por ejemplo:

$$\text{“}\textit{Si } (X_1 = 1 \textit{ y } X_2 = 1) \textit{ o } (X_3 = 1), \textit{ entonces } Y = 1.\text{”}$$

Este modelo indica, entre otras cosas, que la relación entre $X_1$ y $Y$ depende de $X_2$, y que $X_3$ afecta $Y$ de manera independiente.

Modelar esta afirmación en un modelo de regresión requeriría que $\beta_{1,2} = \beta_3 = 1$, $\beta_1 = \beta_2 = \beta_{1,3} = \beta_{2,3} = 0$, y $\beta_{1,2,3} = -1$ para evitar superposición de efectos. Este tipo de complejidad se amplifica rápidamente. Por ejemplo:

$$\text{“}\textit{Si } X_1 = X_2 = X_3 = 1 \textit{ o } X_4 = X_5 = X_6 = 1 \textit{ entonces } Y = 1.\text{”}$$

Este patrón exige una interacción de sexto orden para corregir impactos duplicados, y con más reglas, serían necesarias interacciones de orden aún mayor.

De forma general, si existen $J$ variables, un modelo de regresión necesita hasta interacciones de orden $LM$ y $\sum_{m=1}^{M} \sum_{k=1}^{m} \binom{J}{k}$ términos no nulos para representar un conjunto de $M$ reglas, cada una con longitud máxima $L$. Esta explosión combinatoria genera complejidad y aumenta el riesgo de sobreajuste.

Además, al aumentar la dimensionalidad, el modelo enfrenta la maldición de la dimensionalidad: con $J$ variables e interacciones hasta orden $LM$, la dimensión sería $\sum_{k=1}^{LM} \binom{J}{k}$. En tales casos, es difícil seleccionar parámetros correctos sin regularización, y aumenta el riesgo de errores de estimación.

La segunda razón es más sutil: los conjuntos de reglas clasifican directamente el valor de $Y$ dado $X$, mientras que la regresión estima $E[Y|X]$. Estas no son equivalentes si $Y$ es binaria, ya que $Y$ es más restringida que su esperanza.

Considérese el siguiente ejemplo:

Tabla 2.1: Ejemplo de datos con variables binarias

| $X_1$ | $X_2$ | $E[Y]$ | $N$ |
|---|---|---|---|
| 0 | 0 | 0 | 50 |
| 1 | 0 | 1 | 50 |
| 0 | 1 | – | 50 |
| 1 | 1 | – | 0 |

Condicional a $X_1 = 1$, ya conocemos el valor de $Y$. Condicional a $X_1 = 0$, $X_2$ proporciona cierta información, pero no mejora sustancialmente la predicción. Por tanto, el conjunto de reglas es más parsimonioso que la regresión en este caso (Chiu y Xu, 2023).

Rihoux y Ragin (2009) argumentan que, epistemológicamente, los enfoques cuantitativos tradicionales como la regresión difieren de los métodos configuracionales. Sin embargo, Vis (2012) señala que estos métodos no compiten por el mismo espacio. Más aún, Ragin y Rihoux (2004), Rihoux (2006) y Seawright (2005a) sostienen que si realmente difieren en sus fundamentos epistemológicos, entonces no pueden responder correctamente a una misma pregunta de investigación. Aun así, un investigador pragmático podría considerar que esta diferencia es una ventaja, ya que ofrece visiones complementarias sobre el mismo fenómeno.

## 2.2 ANÁLISIS CUALITATIVO COMPARATIVO (QCA)

### 2.2.1 *Antecedentes y estado del arte*

Como mencionamos en el capítulo anterior, la metodología QCA fue diseñada para trabajar con muestras de pequeño $N$. Sin embargo, su aplicación en muestras grandes ha influido en la expansión de la metodología hacia diferentes campos.

El QCA se desarrolló basado en los cánones de John Stuart (Mill, 1843), específicamente en los métodos de concordancia y diferencia, para identificar relaciones causales en estudios comparativos. En lugar de depender de grandes muestras y estadísticas tradicionales, el QCA permite explorar combinaciones de condiciones que conducen a determinados resultados. Según el método de concordancia, si dos o más ejemplos del fenómeno investigado tienen solo una circunstancia en común, esta circunstancia común es la causa o el efecto del fenómeno en cuestión (Mill, 1843).

Por otro lado, según el método de diferencia, si existe una circunstancia presente cuando ocurre el fenómeno y otra cuando no ocurre, esta diferencia es la causa, el efecto o una parte necesaria del fenómeno. Al combinar ambos métodos, J. S. Mill creó lo que hoy se conoce como "método conjunto de concordancia y diferencia". Aunque menos elaborado que los métodos anteriores, este método avanzó en la aplicación de estas teorías a circunstancias reales y sentó las bases para el desarrollo del QCA (Vassinen, 2012).

Una de las limitaciones de los principios de Mill (1843) es la dificultad para encontrar un único factor que explique un fenómeno, ya que, en la realidad, muchos eventos resultan de diversas causas combinadas. Es decir, los métodos de Mill asumen que existe solo una causa principal, sin tener en cuenta que diferentes factores pueden contribuir simultáneamente al mismo efecto. Además, el método de concordancia solo funciona cuando existen ejemplos positivos del fenómeno, es decir, cuando realmente ocurre, lo que lo hace poco útil para analizar situaciones en las que el evento no ocurre.

Aunque el QCA ha sido tradicionalmente utilizado en estudios con pocas o medianas observaciones, Emmenegger y Klemmensen (2013) argumentan que también puede aplicarse en investigaciones basadas en encuestas. Sin embargo, para superar los desafíos que

surgen con este tipo de muestras, los autores sugieren evaluar la robustez de los resultados y presentan un enfoque específico para garantizar esta confiabilidad. Fiss et al. (2013) sugieren combinar el enfoque del QCA con el análisis de regresión para minimizar posibles limitaciones, como la omisión de variables importantes.

El análisis QCA se basa en el análisis de suficiencia y necesidad para producir un resultado. Una condición se considera necesaria cuando está presente en todas las instancias del resultado y suficiente cuando, siempre que esa condición está presente, el resultado también ocurre (Duşa, 2019; Rihoux y Ragin, 2009; Schneider y Wagemann, 2012). La suficiencia puede expresarse como $X \rightarrow Y$, lo que significa que "X es un subconjunto de Y", es decir, X es suficiente para Y. Sin embargo, para que se considere suficiente, debe observarse que no existen casos en los que X ocurra sin Y. Los casos en los que X no ocurre son irrelevantes para la suficiencia de X, ya que las relaciones de conjunto son asimétricas, es decir, la presencia de X no implica la ausencia de Y.

Además, al afirmar que $X$ es suficiente para $Y$, deben cumplirse tres puntos: primero, existen casos en los que $X$ y $Y$ ocurren juntos; segundo, no debe haber casos en los que $X$ ocurra sin $Y$; y tercero, no podemos afirmar nada sobre $Y$ cuando $X$ no está presente. Cabe destacar que pueden existir múltiples causas (soluciones) suficientes para un mismo resultado, lo que amplía la comprensión de las relaciones causales.

Dos conceptos fundamentales del QCA son la consistencia y la cobertura. La consistencia indica la proporción de casos en los que una determinada configuración causal conduce al mismo resultado. Cuanto mayor es la consistencia, más fuerte es la regla empírica para esa relación. Las reglas con baja consistencia son menos fiables, ya que existen muchos casos en los que la misma combinación de factores no resulta en el mismo resultado. Por otro lado, la cobertura mide la cantidad de casos explicados por una regla específica. A diferencia de la consistencia, una baja cobertura no significa que la regla sea irrelevante. En situaciones donde el mismo resultado puede alcanzarse mediante diferentes reglas causales, una regla puede tener baja cobertura y aún así ser esencial para comprender una de las combinaciones que conducen al resultado (Ragin, 1987, 2000; Woodside y M. Zhang, 2012).

Para organizar las diferentes combinaciones de variables o condiciones que conducen a un resultado, el QCA utiliza tablas de verdad (Fiss et al., 2013; Rihoux y Ragin, 2009). En estas tablas, cada fila representa una configuración específica de condiciones causales asociadas al resultado en análisis. Si hay $k$ condiciones causales, el número total de filas en la tabla de verdad es $2^k$. Sobre estas filas, además, se definirán umbrales de frecuencia para determinar qué configuraciones son relevantes y deben ser consideradas en el análisis.

El QCA tiene dos variantes principales: fuzzy-set QCA (fsQCA) y crisp-set QCA (csQCA).

### 2.2.2 *Tipos de QCA: csQCA y fsQCA*

#### 2.2.2.1 *crisp-Set QCA*

El csQCA fue la primera variante desarrollada de la metodología QCA, creada por Charles Ragin y Kriss Drass en 1987. Su propósito era simplificar el análisis de configuraciones complejas mediante el uso de la lógica booleana, permitiendo identificar múltiples y distintas combinaciones causales. El csQCA, también conocido como QCA-CRISP, se basa en la lógica booleana y en la teoría de conjuntos para examinar las relaciones entre variables cualitativas (Oana et al., 2021).

El método considera las variables independientes como conjuntos de condiciones y busca identificar combinaciones específicas de estas condiciones que sean suficientes o necesarias para la ocurrencia de un determinado resultado. En el csQCA, cada condición se representa de forma binaria, es decir, con uno de dos posibles valores: 1, cuando la condición está presente o fuertemente asociada al fenómeno analizado, y 0, cuando la condición está ausente.

La cantidad total de configuraciones posibles en el csQCA está dada por $2^k$, donde $k$ representa el número de condiciones consideradas en el estudio. El proceso central en el análisis csQCA es la minimización booleana, un procedimiento cuyo objetivo es eliminar condiciones innecesarias para hacer la explicación más concisa. Si dos expresiones son idénticas, excepto por una única condición cuya presencia o ausencia no altera el resultado final, dicha condición se considera irrelevante y puede ser descartada.

A pesar de su utilidad, el csQCA se enfrenta a desafíos al aplicarse en contextos del mundo real, especialmente porque trabaja con valores dicotómicos, mientras que los fenómenos sociales con frecuencia no pueden reducirse a categorías estrictamente binarias. Para superar esta limitación y permitir el análisis de condiciones graduadas, se desarrolló el enfoque fuzzy-set QCA por Ragin (2000).

#### 2.2.2.2 *Fuzzy-set QCA (fsQCA)*

Entre las tres variantes del QCA, el enfoque fsQCA se ha consolidado como el más utilizado en investigaciones. Según datos del proyecto Compasss (https://compasss.org), que recopila publicaciones sobre el método QCA, el número de estudios que emplean conjuntos difusos ha crecido significativamente en comparación con las versiones anteriores (csQCA y mvQCA). Aunque inicialmente fue desarrollado para analizar países, el fsQCA ha expandido su aplicación a diversas áreas del conocimiento, siendo más común en las ciencias sociales, como se abordará en la Sección 3.

La primera formulación del fsQCA, presentada en el Fuzzy Set Social Science por Ragin (2000) no fue concebida directamente como una variante del QCA. Sin embargo, la versión más reciente (Ragin, 2008) incorpora los conjuntos difusos al método QCA, permitiendo la conversión de información cualitativa en valores cuantitativos sin perder sus diferencias esenciales.

Además, el fsQCA permite integrar múltiples afirmaciones dentro de un único modelo analítico. Otro aspecto relevante es su capacidad para manejar relaciones causales asimétricas. Esto significa que, si una condición A conduce a un resultado B, la inversa no necesariamente es verdadera; es decir, la relación entre las condiciones puede no ser bidireccional (Vassinen, 2012).

Siguiendo a Rihoux y Ragin (2009), el fsQCA asigna valores de pertenencia a las condiciones en una escala de 0.0 (no pertenece al conjunto) a 1.0 (pertenece completamente al conjunto), con 0.5 como punto de cruce o máxima ambigüedad. Esta asignación se conoce como calibración del conjunto. Cada configuración de condiciones causales puede representarse como una esquina en un espacio gráfico. Así, una configuración con valores altos para todas sus condiciones, en una tabla con $2^2$ filas, ocupará un punto cercano a la esquina superior derecha de un gráfico de cuatro cuadrantes. No obstante, en una representación con múltiples configuraciones, algunas esquinas pueden estar vacías, ya que ciertas combinaciones de condiciones causales pueden no ocurrir en la realidad empírica.

Una de las tareas clave antes de establecer las configuraciones causales es la definición de las condiciones. Los investigadores deben asegurarse de que dichas condiciones sean relevantes para el análisis. El estudio de un resultado específico debe basarse en teorías existentes y en investigaciones previas dentro del campo especializado. Asimismo, las condiciones seleccionadas deben ser válidas para toda o, al menos, la mayor parte de la población investigada, y los criterios de evaluación deben ser aplicables a dicho grupo.

Una característica del csQCA que lo distingue del fsQCA es su capacidad para identificar contradicciones en las configuraciones causales. Estas contradicciones pueden servir como una línea de investigación independiente. Cuando surge una contradicción, se señala al investigador la necesidad de revisar los casos analizados y la teoría utilizada, a fin de detectar posibles errores o inconsistencias en el modelo.

A diferencia del csQCA, el fsQCA no fue diseñado originalmente para detectar contradicciones. En el csQCA, las condiciones con bajos niveles de consistencia o cobertura son descartadas automáticamente, ya que su valor siempre es 0 (Ragin, 1987, 2000; Woodside y M. Zhang, 2012). En el fsQCA, sin embargo, estas condiciones aún pueden ser consideradas, aunque tengan baja relevancia, lo que puede conducir a soluciones basadas en condiciones irrelevantes. Mientras que en el csQCA incluir condiciones irrelevantes no afecta los cálculos de consistencia y cobertura, en el fsQCA esas condiciones pueden tener valores distintos de 0 (aunque inferiores a 0.5), lo que sí puede afectar los resultados analíticos.

Para afrontar este problema, Rubinson (2013) sugiere el uso del software Kirq, que ayuda a identificar contradicciones. También describe métodos para detectarlas manualmente, sin necesidad de herramientas especializadas.

El problema de las condiciones irrelevantes puede dar lugar a falsos positivos en el análisis. Según Schwellnus (2013), cuando los valores de consistencia son muy bajos pero similares entre sí, pueden ocupar la primera posición en la tabla de verdad del fsQCA. En cambio, en el csQCA, esa misma configuración tendría consistencia igual a 0 y sería

eliminada. Además, como señala Rubinson (2013), establecer un umbral fijo de consistencia puede conducir a conclusiones erróneas.

### 2.2.3 *Revisión bibliográfica del QCA*

En el repositorio de Scopus, se identificaron más de 6.703 artículos que emplean la metodología QCA en sus tres variantes: fsQCA, csQCA y mvQCA. En los últimos 10 años, el número de publicaciones alcanzó 4.337 artículos. Según Roig-Tierno et al. (2017), hasta 2017, las ciencias políticas eran el área con mayor número de aplicaciones de QCA, seguidas por las ciencias de negocios y la economía. Sin embargo, esta tendencia ha cambiado drásticamente en los últimos años.

De acuerdo con los datos actualizados de Scopus, consultados hasta el 7 de abril de 2025, las áreas con mayor número de publicaciones utilizando QCA son actualmente: Ingeniería (1.290), Ciencias de la Computación (1.239), Ciencias Sociales (1.066), Medicina (674) y Ciencias de Negocios, Gestión y Contabilidad (625). En la actualidad, Ingeniería y Ciencias de la Computación juntas representan más del 55 % de los artículos publicados que emplean la metodología QCA, evidenciando un cambio significativo en el perfil de las disciplinas que más utilizan este enfoque.

El QCA ha demostrado una amplia versatilidad en el análisis de fenómenos en diversas áreas del conocimiento, trascendiendo su origen en las ciencias políticas y consolidándose como una herramienta metodológica importante en campos como la economía, la gestión, la salud y el comportamiento del consumidor.

En las ciencias políticas y económicas, el QCA se ha aplicado para evaluar la eficacia de diferentes instrumentos de políticas públicas. Un ejemplo destacado es el estudio de Albers et al. (2013), que investigó cómo distintas combinaciones de políticas de apoyo a la innovación en organizaciones mexicanas influyen en sus niveles de inversión en actividades innovadoras. De manera similar, Becerril-Elías y Merritt (2021) utilizaron el QCA para explorar la relación entre la eficiencia de los mercados financieros y la protección de los derechos de propiedad intelectual en distintos países, contribuyendo a la comprensión de las condiciones que favorecen ambientes institucionales más eficaces.

En el área de gestión de recursos humanos, el estudio de Zavyalova et al. (2020) comparó arquitecturas organizacionales de empresas ágiles y tradicionales, identificando mediante QCA combinaciones específicas de prácticas de gestión que conducen a un alto rendimiento. Este estudio destacó cómo múltiples causas interactúan para generar los mismos resultados, una de las premisas centrales de la lógica del QCA.

En el ámbito de la responsabilidad social corporativa (RSC), Sinthupundaja et al. (2019) aplicaron el QCA para entender cómo las relaciones interorganizacionales impactan el desempeño financiero de hospitales en contextos marcados por prácticas de RSC. La metodología permitió aislar condiciones causales distintas que explican el éxito financiero, considerando tanto factores estructurales como relacionales.

En el campo del comportamiento del consumidor digital, Y. Zhao et al. (2020) utilizaron el QCA para analizar los factores que influyen en el uso de boca a boca electrónico

(e-WOM) en servicios de viaje. El estudio combinó distintas dimensiones de la calidad de la información con características individuales de los consumidores para explicar los patrones de adopción de este tipo de comunicación.

En el dominio de la psicología de la salud, Valero-Moreno et al. (2020) aplicaron el QCA para investigar el bienestar de pacientes pediátricos con disquinesia ciliar primaria (PCD), en comparación con niños saludables. El enfoque permitió identificar configuraciones psicológicas que contribuyen al bienestar, considerando múltiples variables interdependientes.

Estos ejemplos refuerzan la utilidad del QCA para abordar problemas en diversas áreas del conocimiento.

A pesar de su origen en la ciencia política, el QCA ha ganado terreno en otras disciplinas. Sin embargo, áreas como las ciencias ambientales, la agricultura y la farmacología siguen privilegiando otras metodologías, a menudo de naturaleza más experimental. En campos como la educación y los estudios jurídicos, el QCA sigue siendo escasa. En áreas como la economía y la sociología, donde la disponibilidad de datos estadísticos es mayor, el QCA se ha incorporado de forma complementaria, ofreciendo una lente analítica alternativa para comprender las condiciones causales.

El primer artículo que aplicó la variante fsQCA fue publicado por Pennings en el año 2000 (Pennings, 2000). Esta publicación representó un hito en la adopción empírica de la metodología QCA, especialmente en lo relativo al uso de conjuntos difusos. No obstante, presentó ciertas limitaciones conceptuales y técnicas, ya que la formalización más robusta de fsQCA fue desarrollada posteriormente por Ragin en 2008 (Ragin, 2008). Cabe destacar que, desde 2010, más de 6738 de los 6778 artículos publicados que utilizan QCA han empleado las variantes csQCA y fsQCA.

Históricamente, entre 1987 y 2001, el número de publicaciones utilizando el QCA como método de análisis era relativamente bajo. Durante este período, el método estaba en fase de desarrollo y consolidación. A partir de 2001, sin embargo, se observó un crecimiento exponencial en el uso de la metodología. El número de publicaciones aumentó de cinco a casi veinte en solo tres años, alcanzando más de setenta artículos en 2015. Entre 2014 y el primer semestre de 2015, se contabilizaron 134 artículos publicados basados en este enfoque.

En la actualidad, el fsQCA es la variante más ampliamente utilizada dentro de la metodología QCA, a pesar de que el csQCA fue la primera versión introducida. El csQCA surgió en 1987, cuando no existían métodos similares disponibles, siendo pionero en el análisis de causalidad configuracional. El fsQCA, por su parte, fue desarrollado en 2000 como una alternativa que buscaba superar algunas de las limitaciones del csQCA, ofreciendo mayor flexibilidad analítica mediante el uso de conjuntos difusos.

La tercera variante, el mvQCA, presenta un nivel de adopción significativamente menor. A pesar de haber sido propuesta como alternativa al csQCA, introduciendo mejoras en el tratamiento de variables con múltiples valores, no logró resolver satisfactoriamente algunas de las limitaciones estructurales del csQCA, lo que restringió su popularidad (Roig-Tierno et al., 2017).

El análisis comparativo de la distribución de publicaciones muestra que el mvQCA representa una proporción marginal, con solo un 3 % de los artículos. En contraste, el fsQCA representa aproximadamente el 65 % de las publicaciones, mientras que el csQCA representa el 32 %. Aunque el csQCA fue introducido mucho antes, los datos evidencian un claro predominio de la aplicación del fsQCA entre los investigadores, reflejando una preferencia por enfoques más flexibles y adaptados a la complejidad empírica de los estudios contemporáneos.

### 2.2.4 *Calibración de datos en QCA*

La calibración de los datos en QCA es el proceso de asignación de puntajes de pertinencia a los conjuntos para los casos específicos. Duşa (2019) destaca aspectos importantes a considerar en este proceso:

1. Una definición cuidadosa de la población relevante de casos;
2. Una definición precisa del significado de todos los conceptos (tanto las condiciones como el resultado) utilizados en el análisis;
3. Una decisión sobre la ubicación del punto de máxima indiferencia entre pertenencia y no pertenencia (representado por los anclajes de 0,5 en los conjuntos fuzzy y por el umbral en los conjuntos crisp);
4. Una decisión sobre la definición de pertenencia plena (1) y de no pertenencia plena (0);
5. Una decisión sobre la gradación de pertenencia entre los anclajes cualitativos.

Ragin (2000) destaca que la asignación de valores de pertinencia a los conjuntos se basa en un equilibrio entre la calibración mecánica, el conocimiento teórico y las evidencias empíricas. Por lo tanto, es de entera responsabilidad del investigador encontrar un procedimiento racional para asignar valores de pertinencia a los conjuntos. El investigador debe utilizar conocimientos externos a los propios datos (Ragin, 2008). Estos conocimientos provienen de diversas fuentes, como hechos evidentes o conocimientos ampliamente aceptados. Por ejemplo, es un hecho bien establecido que la finalización del duodécimo año escolar en España resulta en la obtención de un diploma de educación secundaria. Al calibrar el conjunto "ciudadanos con educación secundaria", existe una diferencia cualitativa entre completar el undécimo año y completar el duodécimo año. Además, hay nociones ampliamente aceptadas en las ciencias sociales, así como el conocimiento específico del investigador sobre el campo de estudio o casos específicos. Esto requiere un trabajo de campo extenso y un análisis cuidadoso de fuentes primarias y secundarias antes de la calibración propiamente dicha. Así, entrevistas, cuestionarios, datos obtenidos por observación participante, grupos focales, análisis organizacional, análisis de contenido cualitativo y cuantitativo, entre otros, pueden proporcionar información útil para el proceso de calibración de conjuntos (Schneider y Wagemann, 2012).

En el caso de los conjuntos nítidos (crisp), al usar solo un umbral, el procedimiento produce los llamados conjuntos "binarios", dividiendo los casos en dos grupos. Si se utilizan dos o más umbrales, el procedimiento genera conjuntos "multi-valor". En todas las situaciones, el número de grupos obtenidos es igual al número de umbrales más 1 (Duşa, 2019). En el contexto del QCA, el concepto de calibración se asocia más frecuentemente con conjuntos fuzzy, lo cual es correcto, ya que el proceso de calibración para conjuntos nítidos es, esencialmente, un proceso de recodificación de datos. Como el resultado final es "crisp", para conjuntos binarios, el objetivo es recodificar todos los valores por debajo de un cierto límite a 0, y todos los valores por encima de ese límite a 1. Para conjuntos nítidos de múltiples valores, es posible agregar nuevos valores para los intervalos entre dos umbrales consecutivos.

### 2.2.5 *Aplicación de fsQCA en el ámbito educativo*

Para ilustrar el enfoque y su aplicación en el ámbito educativo, presentamos un ejemplo concreto de fsQCA, en el cual utilizamos los diferentes cursos de formación docente como casos de análisis. Este estudio fue realizado en la Universidad Licungo en el año 2024 y se encuentra actualmente en evaluación por la revista científica Rect@ (Dom Luís, Benítez, M. d. C. Bas y Cuamba, 2025, mayo).

El objetivo del estudio fue analizar las condiciones causales asociadas a la (in)satisfacción global de los estudiantes con el sistema educativo ofrecido por dicha institución de educación superior. Se consideraron múltiples factores que influyen en la experiencia académica, tales como la calidad de la enseñanza, los recursos disponibles, y el compromiso docente, con el fin de comprender cómo estos afectan la percepción general del alumnado.

En esta sección, presentamos un subconjunto de ese estudio, en el cual trabajamos con tres condiciones causales: Metodología de Enseñanza y Calidad Académica (MEQA), Accesibilidad y Recursos Tecnológicos (ART), y Puntualidad y Compromiso Docente (PCD) y la variable explicativa, que corresponde a la Satisfacción Global (SG) de los estudiantes.

La recolección de datos se llevó a cabo mediante un cuestionario basadas en una escala de Likert de 1 a 10. Los datos recopilados fueron organizados por curso, calculándose las medias de satisfacción para cada una de las dimensiones evaluadas (ver Tabla 2.2).

La descripción completa del proceso de recolección de datos se encuentra desarrollada en el Capítulo 7 de esta tesis (7).

Tabla 2.2: Promedio de puntuaciones de satisfacción estudiantil por dimensión.

| Carreras | MEQA | ART | PCD | SG |
|---|---|---|---|---|
| Inglés | 7.04761 | 2.66667 | 8.17857 | 7.73810 |
| Geografía | 7.25000 | 4.00000 | 7.00000 | 7.60000 |
| Química | 5.14815 | 3.40741 | 7.22222 | 6.48148 |
| Biología | 7.40790 | 3.73684 | 7.84211 | 7.89474 |
| Psicología | 7.70833 | 3.16667 | 7.08333 | 7.50000 |
| Matemática | 6.65278 | 3.16667 | 7.69444 | 7.61111 |
| Física | 6.18182 | 3.42424 | 6.54546 | 6.81818 |
| Portugués | 7.54436 | 3.35484 | 8.04032 | 8.16129 |
| Francés | 5.67647 | 3.05882 | 5.94118 | 6.11765 |
| Educación Visual | 5.65000 | 3.66667 | 6.16667 | 6.13333 |

**Nota:** MEQA = Metodología de Enseñanza y Calidad Académica; ART = Accesibilidad y Recursos Tecnológicos; PCD = Puntualidad y Compromiso Docente; SG = Satisfacción Global.

El proceso de calibración en la metodología fsQCA se basa en la transformación de los datos numéricos originales en puntuaciones difusas que reflejan los niveles de pertenencia a los conjuntos analizados, siendo una etapa fundamental para garantizar la validez y la interpretabilidad de las fases posteriores del análisis.

La calibración es un procedimiento que convierte las puntuaciones numéricas brutas en valores entre 0 y 1, donde 1 representa la condición de estar "completamente dentro" del conjunto (*fuzzy* = 1.0) y 0 indica estar "completamente fuera" del conjunto (*fuzzy* = 0.0). Este proceso se fundamenta en umbrales tanto empíricos como teóricos, asegurando que las puntuaciones reflejen niveles claros y significativos de pertenencia, como señalan Ragin (2008) y Schneider y Wagemann (2012).

En este estudio, se empleó el método de calibración directa propuesto por Rihoux y Ragin (2009), utilizando tres anclas definidas a partir de umbrales cualitativos y evidencia empírica. Los datos originales consistían en puntuaciones ordinales categóricas que oscilaban entre 1 (Muy insatisfecho) y 10 (Muy satisfecho).

Para la definición de las anclas de calibración, se siguió la recomendación de Duşa (2019), quien sugiere el uso de tres puntos de corte que dan lugar a una función de pertenencia con forma sigmoide (en "S"), representada por $\mu(x)$, como se ilustra en la Figura 2.2.

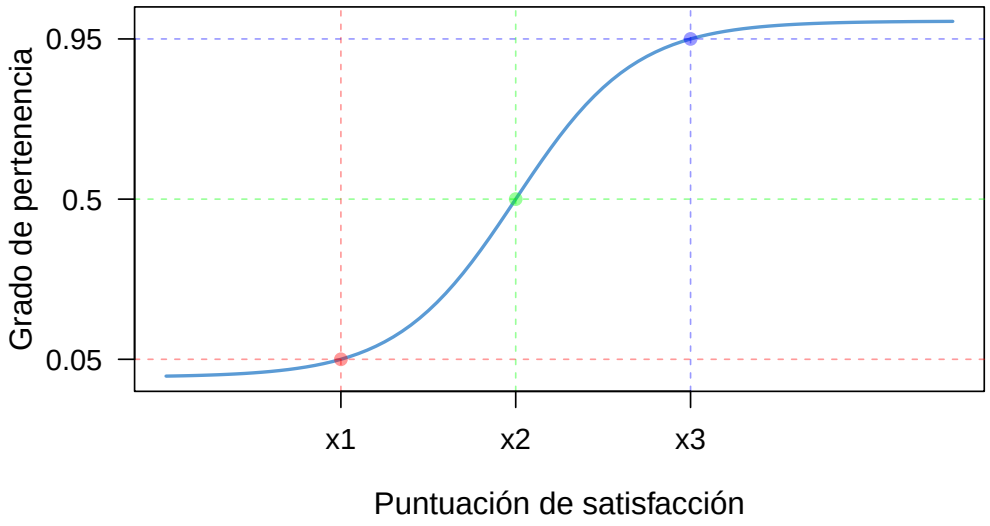


Figura 2.2: Función logística de calibración directa en fsQCA, con los umbrales $x_1$, $x_2$ y $x_3$.

Ragin (2008) propone estimar la función de pertenencia con forma sigmoide utilizando el logaritmo de las probabilidades (*log-odds*):

$$\log\text{-}odds(x) = \log\left(\frac{\mu(x)}{1-\mu(x)}\right), \tag{2.1}$$

donde $\mu(x)$ representa el grado de pertenencia a un conjunto difuso. Las probabilidades de pertenencia corresponden a la razón entre la pertenencia al conjunto y la no pertenencia al mismo:

$$odds\text{-}\mu(x) = \frac{\mu(x)}{1-\mu(x)}.$$

Al aplicar la función logística de Ragin, los puntos de corte $x_1$ (umbral de no pertenencia total) y $x_3$ (umbral de pertenencia total) se establecen, respectivamente, en los niveles de pertenencia 0,05 y 0,95. El punto de cruce (*crossover point*), representado por $x_2$, se define con un valor de pertenencia igual a 0,5 (Remo-Diez et al., 2024).

Una vez definidos estos tres puntos críticos, es posible calibrar el grado de pertenencia a las condiciones o al resultado mediante la siguiente fórmula:

$$\mu(x) = \begin{cases} \dfrac{\exp\left(-\log\text{-}odds(x_1) \cdot \dfrac{(x - x_2)}{x_2 - x_1}\right)}{1 + \exp\left(-\log\text{-}odds(x_1) \cdot \dfrac{(x - x_2)}{x_2 - x_1}\right)}, & \text{si } x < x_2 \\ \dfrac{\exp\left(\log\text{-}odds(x_3) \cdot \dfrac{(x - x_2)}{x_3 - x_1}\right)}{1 + \exp\left(\log\text{-}odds(x_3) \cdot \dfrac{(x - x_2)}{x_3 - x_1}\right)}, & \text{si } x \geq x_2 \end{cases} \tag{2.2}$$

En el presente estudio, se realizó la calibración utilizando la función logística de Ragin, adoptando como puntos de corte los percentiles P5, P50 y P95. No obstante, es importante señalar que también pueden emplearse otras funciones de pertenencia, tanto lineales como no lineales, dependiendo de la naturaleza teórica y empírica del fenómeno bajo análisis (Mendel y Korjani, 2018).

En la metodología *fsQCA*, es crucial evitar valores de calibración exactamente iguales a 0.5, ya que en la construcción de la tabla de verdad los datos calibrados se binarizan para definir claramente la pertenencia o no pertenencia a un conjunto (0 = fuera, 1 = dentro). Un valor de 0.5 representa el máximo grado de ambigüedad, situando al caso justo en el umbral que divide estas dos categorías. En consecuencia, la Tabla 2.3 presenta los datos calibrados derivados de la Tabla 2.2, asegurando que no haya valores con pertenencia igual a 0.5. (Rihoux y Ragin, 2009; Schneider y Wagemann, 2012).

Tabla 2.3: Datos calibrados para los cursos analizados

| **Curso** | **MEQA** | **ATR** | **PCD** | **SG** |
|---|---|---|---|---|
| Inglés | 0.4781 | 0.0200 | 0.9590 | 0.7558 |
| Geografía | 0.7052 | 0.9746 | 0.4018 | 0.5745 |
| Química | 0.4786 | 0.5402 | 0.5551 | 0.0997 |
| Biología | 0.8498 | 0.8910 | 0.8930 | 0.8881 |
| Psicología | 0.9668 | 0.2380 | 0.4559 | 0.4742 |
| Matemática | 0.2889 | 0.2380 | 0.8415 | 0.5907 |
| Física | 0.1335 | 0.5647 | 0.1674 | 0.1814 |
| Portugués | 0.9225 | 0.4657 | 0.9388 | 0.9752 |
| Francés | 0.0516 | 0.1478 | 0.0389 | 0.0497 |
| Educación Visual | 0.0490 | 0.8439 | 0.0685 | 0.0513 |

Después de la calibración de los datos, la metodología fsQCA aplica una secuencia de tres etapas fundamentales para convertir una matriz de datos calibrados en una tabla de verdad. La primera etapa consiste en la construcción de la tabla de verdad, que enumera todas las combinaciones posibles de las condiciones causales. En segundo lugar, se asignan los casos a las filas correspondientes de dicha tabla en función de sus puntuaciones de pertenencia. Finalmente, se determina el valor del resultado para cada combinación de condiciones, considerando los valores del resultado de los casos asociados a cada fila.

En el presente estudio se consideran diez casos (carreras) con tres condiciones causales y un resultado, según la matriz de datos fuzzy presentada en la Tabla 2.3.

Aunque la lógica fuzzy permite grados de pertenencia continuos entre 0 y 1, tanto en fsQCA como en csQCA (crisp-set QCA), la construcción de la tabla de verdad se basa en la binarización de los valores fuzzy a partir de un umbral cualitativo, usualmente 0.5. Los valores superiores a este punto de corte se interpretan como una pertenencia significativa al conjunto (codificados como 1), mientras que los valores inferiores se consideran como no pertenencia (codificados como 0). Por lo tanto, cada condición puede asumir dos estados, y el número total de combinaciones posibles (filas de la tabla de verdad) se determina mediante la fórmula $2^k$, donde $k$ representa el número de condiciones causales.

En fsQCA, los casos con puntuaciones fuzzy de pertenencia no siempre coinciden exactamente con una fila específica de la tabla de verdad, a menos que se binaricen según el criterio antes mencionado. Por ejemplo, considérese el caso de la carrera de Inglés, cuya vector de pertenencias en la Tabla 2.3 es: MEQA = 0.4781, ATR = 0.0200 y PCD = 0.9590. Para determinar a qué fila de la Tabla 2.4 debe asignarse este caso, se convierte cada valor fuzzy a binario. En este caso, tanto MEQA como ATR son menores que 0.5, por lo que se codifican como 0 (ausencia), mientras que PCD es mayor que 0.5, por lo que se codifica como 1 (presencia). La configuración resultante es (MEQA = 0, ATR = 0, PCD = 1), representada simbólicamente como *meqa*atr*PCD*. Esta configuración corresponde a la séptima fila de la Tabla 2.4.

El uso de letras minúsculas indica la negación de una condición. Por ejemplo, "meqa" representa la ausencia de la condición MEQA. Este procedimiento de binarización y asignación se aplica de forma sistemática a todos los casos de la matriz de datos fuzzy, garantizando la coherencia lógica entre los datos empíricos calibrados y la estructura de la tabla de verdad utilizada en los análisis posteriores de fsQCA.

Tabla 2.4: Tabla de verdad a partir de la matriz de datos fuzzy

| Orden | Combinación | MEQA | ATR | PCD | SG | N | Casos |
|---|---|---|---|---|---|---|---|
| 1 | MEQA*ATR*PCD | 1 | 1 | 1 | 1 | 1 | Biología |
| 2 | MEQA*ATR*pcd | 1 | 1 | 0 | 1 | 1 | Geografía |
| 3 | MEQA*atr*PCD | 1 | 0 | 1 | 1 | 1 | Portugués |
| 4 | MEQA*atr*pcd | 1 | 0 | 0 | 0 | 1 | Psicología |
| 5 | meqa*ATR*PCD | 0 | 1 | 1 | 0 | 1 | Química |
| 6 | meqa*ATR*pcd | 0 | 1 | 0 | 0 | 2 | Física, Educación Visual |
| 7 | meqa*atr*PCD | 0 | 0 | 1 | 1 | 2 | Inglés, Matemática |
| 8 | meqa*atr*pcd | 0 | 0 | 0 | 0 | 1 | Francés |

Como comentamos anteriormente, la tabla de verdad constituye el elemento central de la metodología QCA. Para cada una de las combinaciones lógicamente posibles de las condiciones causales, la tabla de verdad muestra los caminos que conducen a la presencia o

ausencia del resultado, en este caso, la (in)satisfacción global de los estudiantes, basándose en los diez casos analizados.

Se observa que todas las ocho posibles condiciones causales están representadas en el conjunto de las diez carreras analizadas. Esto significa que no hay combinaciones lógicas posibles sin instancias empíricas asociadas, es decir, no existen restos lógicos (*logical remainders*) en la tabla. De lo contrario, estaríamos ante un escenario de *diversidad empírica limitada*.

Se considera que una configuración conduce a la presencia del resultado cuando el valor de la variable de salida (satisfacción global, $SG$) es igual a 1. Por ejemplo, la configuración correspondiente a la carrera de Geografía (segunda fila de la Tabla 2.4) está dada por $MEQA * ATR * pcd$ y se asocia con la presencia del resultado ($SG = 1$). Esto se interpreta de la siguiente manera: estudiantes satisfechos con la metodología de enseñanza, Calidad Academemica Y con la accesibilidad y los recursos tecnológicos Y altamente insatisfechos con la puntualidad y el compromiso docente, tienden a presentar una satisfacción global positiva con el sistema educativo.

En nuestro estudio, la presencia del resultado ($SG = 1$) se observó en las carreras de Biología, Geografía, Portugués, Inglés y Matemáticas. Las configuraciones causales que conducen al resultado son las siguientes:

- $MEQA * ATR * PCD$ (Biología),
- $MEQA * ATR * pcd$ (Geografía),
- $MEQA * atr * PCD$ (Portugués),
- $meqa * atr * PCD$ (Inglés y Matemáticas).

Aplicando las reglas del álgebra booleana, podemos identificar redundancias y simplificar estas expresiones. Por ejemplo:

- Las configuraciones $MEQA * ATR * PCD$ y $MEQA * ATR * pcd$ (Biología y Geografía) pueden reducirse a $MEQA * ATR$, ya que la condición $PCD$ (puntualidad y compromiso docente) no altera el resultado.
- De forma análoga, $MEQA * ATR * PCD$ (Biología) y $MEQA * atr * PCD$ (Portugués) pueden reducirse a $MEQA * PCD$, indicando que la presencia de MEQA y PCD es suficiente para generar el resultado, siendo la condición ATR irrelevante en este contexto.
- Por último, las configuraciones $MEQA * atr * PCD$ (Portugués) y $meqa * atr * PCD$ (Inglés y Matemáticas) comparten las condiciones $atr * PCD$, por lo que la condición MEQA se vuelve dispensable. Estas configuraciones pueden, por tanto, reducirse a $atr * PCD$.

Por lo tanto, el análisis realizado mediante la técnica fsQCA permitió identificar tres trayectorias causales distintas y consistentes que conducen a la presencia del resultado ($SG = 1$) entre los estudiantes de la Facultad de Educación de la Universidad Licungo:

$$SG = MEQA * ATR + MEQA * PCD + atr * PCD$$

Cada una de estas soluciones puede interpretarse de la siguiente manera:

$MEQA * ATR$**:** Esta solución implica que cuando los estudiantes están satisfechos con la metodología de enseñanza (**MEQA**) y con la accesibilidad y los recursos tecnológicos (**ATR**), tienden a reportar una satisfacción global positiva con el sistema de enseñanza, independientemente de su percepción sobre la puntualidad y el compromiso docente. En otras palabras, esta combinación es suficiente por sí sola para generar el resultado deseado.

$MEQA * PCD$**:** Esta combinación muestra que la elevada satisfacción con la metodología de enseñanza (**MEQA**) junto con una percepción positiva sobre la puntualidad y el compromiso docente (**PCD**) también conduce a una alta satisfacción global. Aquí, la confianza en la enseñanza y el compromiso del profesorado compensa otras posibles carencias.

$atr * PCD$**:** Esta tercera combinación revela que una combinación entre condiciones deficientes en metodología de enseñanza (**meqa**, es decir, baja satisfacción) y una percepción positiva del profesorado (**PCD**) puede igualmente generar satisfacción global.

Es importante destacar que, en QCA, no importa la frecuencia con la que una configuración aparece en los datos empíricos. Incluso si se observa una sola vez, puede considerarse causal si es lógicamente consistente y empíricamente válida.

En cuanto al análisis de condiciones necesarias y suficientes, se concluye que, en el conjunto de casos analizados, ninguna condición individual es necesaria para la ocurrencia del resultado. No obstante, las conjunciones descritas constituyen trayectorias suficientes que explican de forma consistente la satisfacción positiva de los estudiantes con el sistema educativo.

Al tratar con el análisis de suficiencia en conjuntos fuzzy, la consistencia de una condición suficiente se define como el grado en que la pertenencia de cada caso en la condición $X$ es menor o igual a su pertenencia en el resultado $Y$. En otras palabras, se evalúa hasta qué punto la condición $X$ conduce lógicamente al resultado $Y$, basándose en los grados de pertenencia difusa observados.

Esta lógica puede expresarse mediante la siguiente fórmula, según Rihoux y Ragin (2009):

$$\text{Consistencia}_{X \Rightarrow Y} = \frac{\sum_{i=1}^{I} \text{mín}(X_i, Y_i)}{\sum_{i=1}^{I} X_i} \tag{2.3}$$

Si, para todos los casos, los valores de $X$ son menores o iguales a los valores de $Y$, entonces todos los puntos se ubicarán por debajo o sobre la diagonal principal en el gráfico de dispersión $XY$, y la fórmula toma el valor de consistencia igual a 1, ya que el mínimo entre $X$ y $Y$ será, en todos los casos, el valor de $X$.

Cuantos más casos presenten un valor de pertenencia en $X$ que excede su pertenencia en $Y$ (y cuanto mayor sea esta diferencia), más casos se situarán por encima de la diagonal (y más alejados de ella), lo que reduce el valor de la consistencia, ya que el numerador acumula valores menores que el denominador.

Además de la consistencia, es habitual calcular la cobertura, que evalúa el grado en que la condición $X$ explica o cubre el resultado $Y$. La cobertura es una medida de relevancia empírica, similar a la noción de varianza explicada en modelos estadísticos.

La fórmula de la cobertura es:

$$\text{Cobertura}_{X \Rightarrow Y} = \frac{\sum_{i=1}^{I} \text{mín}(X_i, Y_i)}{\sum_{i=1}^{I} Y_i} \tag{2.4}$$

La cobertura indica qué proporción del resultado $Y$ está explicada por la condición suficiente $X$. Incluso una condición con alta consistencia puede presentar una cobertura baja, lo que sugiere que, aunque la condición conduzca lógicamente al resultado, no se observa con la frecuencia suficiente como para explicar una parte significativa de los casos en los que el resultado ocurre.

Para nuestro ejemplo, la Tabla 2.5 presenta el proceso de cálculo de las métricas de las tres configuraciones solución (consistencia y cobertura).

Tabla 2.5: Tabla de Soluciones con Consistencia y Cobertura

| **Solución** | $X_i$ | $Y_i$ | $\sum \text{mín}(X_i)$ | $\sum \text{mín}(Y_i)$ | $\sum \text{mín}(X_i, Y_i)$ | **Const** | **Cob** |
|---|---|---|---|---|---|---|---|
| $MEQA * ATR \rightarrow SG$ | MEQA*ATR | $SG$ | 3.23 | 4.64 | 2.72 | 0.84 | 0.58 |
| $MEQA * PCD \rightarrow SG$ | MEQA*PCD | $SG$ | 4.10 | 4.64 | 3.72 | 0.91 | 0.80 |
| $atr * PCD \rightarrow SG$ | atr*PCD | $SG$ | 3.58 | 4.64 | 2.83 | 0.79 | 0.60 |

Interpretación racional de las métricas:

$MEQA * ATR \rightarrow SG$**:** Esta combinación presenta una consistencia de **0,84**, lo que indica que en el 84 % de los casos en que se observa la configuración, también se produce el resultado. Su cobertura es de **0,58**, lo que significa que esta combinación explica el 58 % de todos los casos en que se alcanza la satisfacción global. Es una condición lógicamente consistente, pero con cobertura moderada.

$MEQA * PCD \rightarrow SG$**:** Con una consistencia de **0,91**, esta configuración es la más fuerte lógicamente entre las tres, indicando una relación casi perfecta entre condición y resultado. Su cobertura de **0,80** sugiere que también tiene una alta relevancia empírica, cubriendo el 80 % de los casos en que se presenta el resultado.

$atr * PCD \rightarrow SG$**:** Esta configuración tiene una consistencia de **0,79**, lo que representa una relación lógica aceptable, aunque más débil en comparación con las otras combinaciones.

La cobertura es de **0,60**, lo que indica que explica el 60 % de los casos con satisfacción global, lo que le confiere una relevancia empírica moderada.

## 2.3 MINERÍA DE DATOS Y REGLAS DE ASOCIACIÓN

### 2.3.1 *Minería de Datos*

La minería de datos es el proceso de descubrir nueva información en grandes cantidades de datos, utilizando patrones o reglas. Esta nueva información resulta ser útil en todo el proceso de toma de decisiones sobre actividades presentes y futuras. Para que este proceso sea eficaz, debe llevarse a cabo de manera eficiente en grandes archivos o bases de datos (Elmasri y Navathe, 2017).

El descubrimiento de conocimiento en grandes volúmenes de datos es un desafío, y la minería de datos se considera un campo interdisciplinario. Esto se debe a que abarca conceptos y técnicas de otras áreas, lo que permite su aplicabilidad (Fayyad et al., 1996). Entre las áreas involucradas, destacan el aprendizaje automático, las redes neuronales, los algoritmos genéticos y la estadística (Tan et al., 2014). La ciencia de la extracción de conocimiento es un proceso no trivial que identifica patrones válidos, nuevos, potencialmente útiles y comprensibles en los datos. Este proceso de descubrimiento de conocimiento se divide en varias fases, que se describen a continuación:

1. **Selección**: En esta etapa, se selecciona una base de datos relevante para buscar patrones y conocimientos útiles.

2. **Preprocesamiento**: Tras el proceso de selección, los datos se organizan y se tratan con el objetivo de depurar datos faltantes e inconsistentes.

3. **Transformación**: Los datos se ajustan al formato apropiado para continuar con el proceso de minería.

4. **Minería de datos**: En esta etapa, se aplican métodos inteligentes para extraer patrones o reglas relevantes.

5. **Interpretación y Evaluación**: Los resultados de la minería se analizan, consolidan y se presentan a los usuarios y tomadores de decisiones.

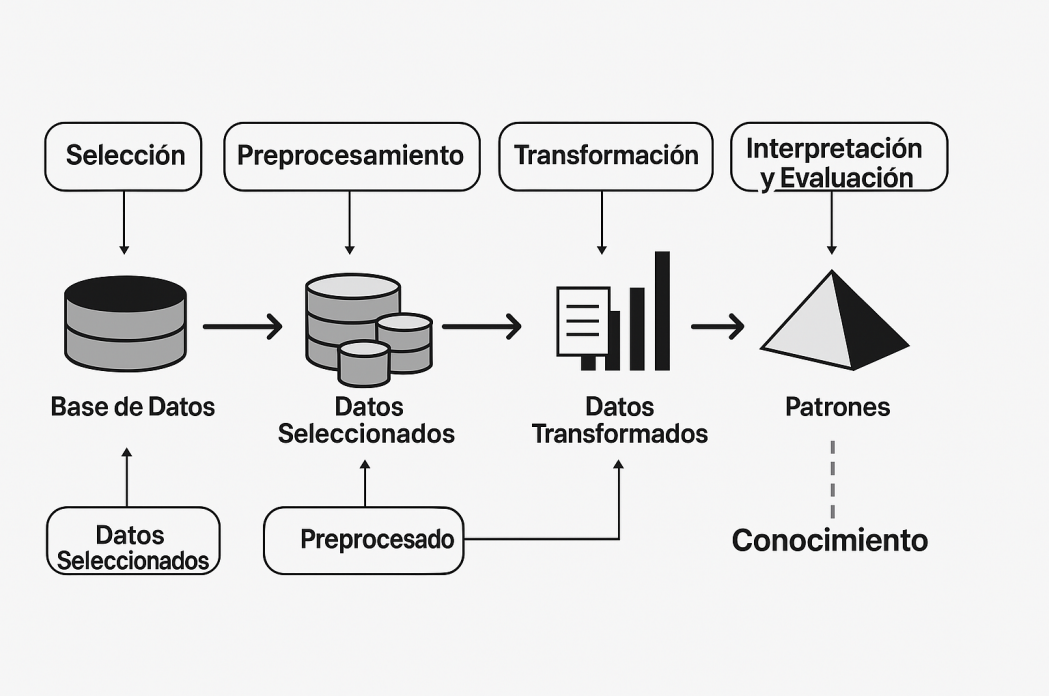


Figura 2.3: Fases del proceso de Minería

La Figura 2.3 ilustra todas las etapas del proceso de descubrimiento de conocimiento en secuencia. En el proceso de descubrimiento de minería de datos, existen diferentes técnicas para cada tipo de problema que se busca resolver. Es decir, diferentes métodos son adecuados para diferentes propósitos. Cada uno de estos métodos ofrece ventajas y desventajas. Por lo tanto, es importante que el investigador se familiarice con las técnicas para elegir la más adecuada para el problema que desea resolver (Witten y Frank, 2005). Las técnicas más prominentes en la minería de datos son: reglas de asociación, clasificación, agrupamiento (clustering) y regresión. La extracción de reglas de asociación es una de las más utilizadas y también uno de los modelos centrales de esta tesis.

### 2.3.2 *Reglas de Asociación*

Generalmente, en todas las actividades humanas, existe una acumulación constante de grandes volúmenes de datos generados por las operaciones diarias. Un ejemplo claro de esto son los datos sobre las compras de clientes, los cuales se registran diariamente en las cajas de los supermercados. Estos datos, cuando se analizan adecuadamente, pueden revelar patrones y asociaciones valiosas.

Investigadores de diversas áreas del conocimiento han mostrado interés en comprender el conjunto de variables que ocurren simultáneamente y que pueden estar relacionadas con ciertos fenómenos. En medicina, por ejemplo, los profesionales buscan identificar las condiciones que, en conjunto, contribuyen al desarrollo de enfermedades como la diabetes. En el comercio, los minoristas analizan el comportamiento de compra de los consumidores con el objetivo de comprender mejor sus preferencias y necesidades.

Cuando se extraen e interpretan de manera efectiva, esta información puede aplicarse en diversas áreas, como la salud, la educación y los negocios. Además, puede orientar decisiones estratégicas relacionadas con promociones de marketing, gestión de inventarios y relaciones con los clientes, contribuyendo a una mayor eficiencia y mejores resultados en diferentes contextos.

La extracción de reglas de asociación es una de las técnicas más utilizadas en el área de minería de datos. Este enfoque fue inicialmente propuesto por R. Agrawal et al. (1993), con el objetivo de identificar relaciones entre artículos adquiridos conjuntamente en transacciones comerciales, particularmente en supermercados. Este tipo de análisis se conoció ampliamente como "análisis de la cesta de la compra", tal como también lo discuten Tan et al. (2014).

Las reglas de asociación permiten identificar relaciones relevantes entre variables en una base de datos. En el contexto de la gestión estratégica de marketing, por ejemplo, los gestores pueden utilizar estas reglas para identificar oportunidades de venta cruzada (cross-selling), optimizando sus estrategias de promoción y fidelización de clientes.

Aunque la aplicación inicial del análisis de reglas de asociación fue en marketing, su uso ha crecido significativamente en los últimos años. Actualmente, esta técnica también se aplica en áreas como bioinformática, diagnóstico médico, minería de datos en la web y análisis de datos científicos. En el campo de las ciencias de la Tierra, por ejemplo, la identificación de patrones de asociación ha permitido revelar conexiones relevantes entre procesos oceánicos, terrestres y atmosféricos. Esta información es fundamental para que los científicos comprendan mejor cómo los diferentes componentes del sistema terrestre interactúan entre sí (Tan et al., 2014).

La minería de reglas de asociación es una técnica que identifica asociaciones significativas entre un gran conjunto de artículos en los datos. Las reglas de asociación revelan condiciones de valores de atributos que ocurren frecuentemente de forma conjunta en un determinado conjunto de datos, y se expresan en un formato "si-entonces". Estas reglas son extraídas de los datos, y su estructura consta de dos partes: el antecedente (parte "si") y el consecuente (parte "entonces"). Además de estas dos partes, una regla de asociación incluye dos parámetros que expresan el grado de incertidumbre sobre su validez.

En el escenario original, las reglas de asociación se extraen de bases de datos transaccionales compuestas por un conjunto de artículos $I = \{i_1, i_2, \ldots, i_n\}$, donde cada artículo $i$ es un atributo binario, y por un conjunto de transacciones $D = \{t_1, t_2, \ldots, t_m\}$, en el cual cada transacción $t_k$ es un subconjunto de $I$. En términos generales, una regla de asociación es un par de conjuntos de artículos $(X, Y)$, que se representa frecuentemente como una implicación $X \Rightarrow Y$, donde $X$ es el antecedente (o premisa) y $Y$ es el consecuente (o conclusión) de la regla. Además, $X$ y $Y$ son conjuntos disjuntos, es decir, $X \cap Y = \emptyset$, lo que significa que los artículos de $X$ y $Y$ no pueden solaparse.

Para evaluar la relevancia y la fuerza de una regla de asociación, se utilizan tres métricas fundamentales: el soporte, la confianza y el lift.

El soporte de un conjunto de artículos $X$ determina la frecuencia con que ese conjunto de artículos aparece en una base de datos transaccional. Para una regla de asociación $X \Rightarrow Y$,

el soporte se define como el porcentaje de transacciones que contienen tanto los artículos de $X$ como los de $Y$, en relación con el número total de transacciones presentes en la base de datos. Matemáticamente, esto se expresa con la fórmula:

$$\text{Soporte}(X \Rightarrow Y) = \frac{|X \cup Y|}{|D|} \tag{2.5}$$

donde $|X \cup Y|$ es el número de transacciones que contienen tanto los conjuntos $X$ como $Y$, y $|D|$ es el número total de transacciones en la base de datos.

La confianza de una regla de asociación $X \Rightarrow Y$ evalúa con qué frecuencia los artículos de $Y$ aparecen en las transacciones que contienen los artículos de $X$. En otras palabras, la confianza determina la probabilidad de que el consecuente $Y$ ocurra en las transacciones que ya contienen el antecedente $X$. La fórmula para calcular la confianza de una regla es la siguiente:

$$\text{Confianza}(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|} \tag{2.6}$$

donde $|X \cup Y|$ es el número de transacciones que contienen tanto $X$ como $Y$, y $|X|$ es el número total de transacciones que contienen $X$.

Finalmente, el lift de una regla de asociación $X \Rightarrow Y$ es una medida que evalúa la fuerza de la asociación entre $X$ y $Y$, considerando la independencia de los dos conjuntos. El lift mide el aumento de la probabilidad de ocurrencia de $Y$ cuando $X$ ocurre, en comparación con la probabilidad de $Y$ ocurra de forma independiente. Si el lift es mayor que 1, esto indica que $X$ y $Y$ aparecen juntos más frecuentemente de lo esperado, lo que implica una asociación positiva. Si es igual a 1, significa que $X$ y $Y$ son independientes, y si es menor que 1, esto indica una asociación negativa entre $X$ y $Y$. El lift se calcula con la siguiente fórmula:

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confianza}(X \Rightarrow Y)}{\text{Soporte}(Y)} \tag{2.7}$$

Las tres métricas, soporte, confianza y lift, son esenciales para el análisis de reglas de asociación, ya que el soporte permite identificar la frecuencia de ocurrencia de los artículos juntos en la base de datos, la confianza refleja la dependencia entre los artículos, ayudando a determinar la probabilidad de ocurrencia del consecuente, dado el antecedente, y el lift ayuda a evaluar la fuerza de la asociación entre $X$ y $Y$, considerando la independencia entre los artículos.

### 2.3.3 *El Algoritmo Apriori*

El Algoritmo Apriori es uno de los métodos más conocidos y utilizados para la minería de datos en la búsqueda de reglas de asociación. Representó un avance significativo en comparación con otros algoritmos, especialmente en términos de rendimiento y estrategia para resolver el problema de extracción de reglas. Es considerado un algoritmo pionero,

y a partir de él se han desarrollado muchos otros algoritmos que conforman la llamada Familia Apriori.

La estrategia principal del algoritmo Apriori consiste en dos pasos:

1. **Identificación de Conjuntos Frecuentes**: Detectar los conjuntos de ítems cuyo soporte (frecuencia de aparición en la base de datos) sea mayor o igual a un umbral mínimo de soporte, denotado como min.sup.

2. **Generación de Reglas de Asociación**: A partir de los conjuntos frecuentes, se construyen reglas que cumplan con un umbral mínimo de confianza, min.conf.

#### 2.3.3.1 *Principio Apriori*

La principal innovación del Apriori radica en el siguiente principio:

**Teorema 1** (Principio Apriori)**.** *Si un conjunto de ítems es frecuente, entonces todos sus subconjuntos también deben ser frecuentes.*

Gracias a este principio, es posible optimizar la búsqueda de conjuntos frecuentes eliminando de forma anticipada los superconjuntos de ítems infrecuentes. Esta técnica se conoce como poda basada en soporte, véase la Figura 2.4.

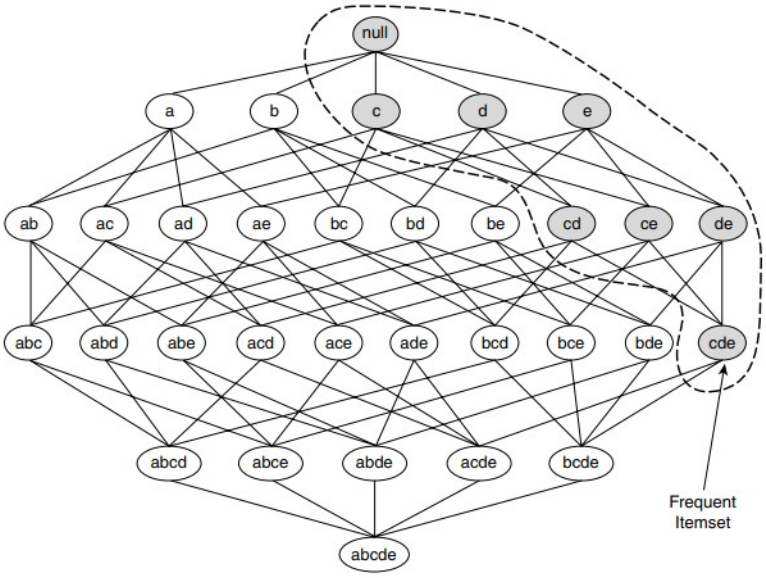


Figura 2.4: Ilustración del Principio Apriori: si {c, d, e} es frecuente, entonces todos sus subconjuntos también lo son.

Por el contrario, si un conjunto de ítems es infrecuente, todos sus superconjuntos también serán infrecuentes. Esta estrategia reduce significativamente el espacio de búsqueda, véase en la Figura 2.5.

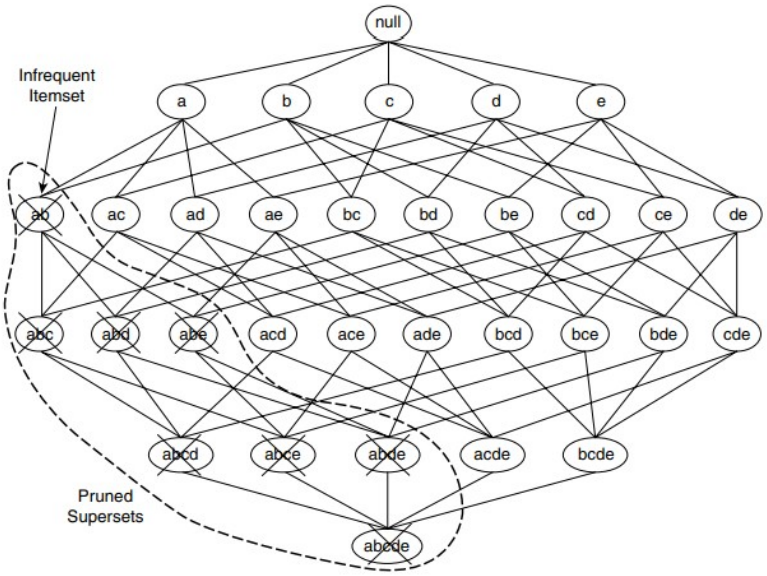


Figura 2.5: Poda basada en soporte: si {a, b} es infrecuente, todos sus superconjuntos también lo serán.

**Definición 1** (Propiedad de Monotonicidad)**.** *Sea $J$ el conjunto potencia de un conjunto $A$. Una medida $f$ es monótona (o cerrada hacia arriba) si para todos $X, Y \in J$, tal que $X \subseteq Y \Rightarrow f(X) \leq f(Y)$.*

**Definición 2** (Propiedad de Antimonotonicidad)**.** *Una medida $f$ es antimonótona (o cerrada hacia abajo) si para todos $X, Y \in J$, tal que $X \subseteq Y \Rightarrow f(Y) \leq f(X)$.*

Estas propiedades son fundamentales para la eficiencia del algoritmo, ya que permiten podar el espacio de búsqueda de forma sistemática.

Para un conjunto de datos con $k$ ítems, se pueden generar hasta $2^k - 1$ conjuntos (excluyendo el conjunto vacío). Este crecimiento exponencial justifica la importancia de la poda.

El cálculo del soporte para cada conjunto candidato implica recorrer la base de datos y verificar la presencia del conjunto en cada transacción. Para mejorar el rendimiento, se puede usar una estructura llamada *árbol hash* (*hash-tree*) que facilita esta verificación de forma eficiente.

El algoritmo Apriori para la generación de patrones frecuentes funciona mediante la creación iterativa de conjuntos de ítems candidatos y el cálculo de su soporte. En cada iteración, se generan los conjuntos candidatos de tamaño $k$ a partir de los ítems frecuentes encontrados en la iteración anterior ($k-1$) mediante la función `apriori-gen`. Posteriormente, se calcula el soporte de cada candidato, contando cuántas veces aparece en la base de datos. Solo los conjuntos cuyo soporte alcanza o supera el umbral mínimo establecido (*min_sup*) son considerados frecuentes y retenidos para la siguiente iteración.

Se puede observar que el algoritmo utiliza dos funciones principales: *apriori-gen* y *subset*, cada una con operaciones específicas. La función *apriori-gen* recibe como argumento el conjunto $L_{k-1}$ (compuesto por todos los $(k-1)$-*itemsets* frecuentes) y retorna el conjunto $C_k$ (compuesto por todos los *k-itemsets* candidatos). Esta función realiza dos etapas clave:

**Algoritmo 2.1** Algoritmo Apriori Principal

**Require:** $D$: base de datos, $min\_sup$: soporte mínimo
**Ensure:** $L$: conjunto de todos los itemsets frecuentes
1: $L_1 \leftarrow \{$itemsets frecuentes de tamaño 1 en $D\}$
2: $k \leftarrow 2$
3: **while** $L_{k-1} \neq \emptyset$ **do**
4: $C_k \leftarrow$ apriori-gen$(L_{k-1})$
5: **for** cada transacción $t \in D$ **do**
6: $C_t \leftarrow$ subset$(C_k, t)$
7: **for** cada candidato $c \in C_t$ **do**
8: $c$.count $\leftarrow c$.count $+ 1$
9: **end for**
10: **end for**
11: $L_k \leftarrow \{c \in C_k \mid c.\text{count} \geq min\_sup\}$
12: $k \leftarrow k + 1$
13: **end while**
14: **return** $\bigcup_k L_k$

- **Join (Unión):** Genera combinaciones al unir $L_{k-1}$ consigo mismo.
- **Prune (Poda):** Elimina los *itemsets* candidatos que contengan subconjuntos no frecuentes. Formalmente, se remueve todo *k-itemset* $c \in C_k$ si existe un $(k-1)$-*itemset* $s$ tal que $s \subset c$ y $s \notin L_{k-1}$.

Estas etapas se detallan en el pseudocódigo siguiente:

La función *subset* recibe como argumento el conjunto $C_k$ con los *itemsets* ya podados y una transacción $t$ de la base de datos, y retorna el conjunto $C_t$ compuesto por los *itemsets* presentes en $t$. Esta función utiliza una estructura de hash para encontrar los *itemsets* de $C_k$ en la transacción $t$. Con esto, el soporte de cada *itemset* en $C_k$ se determina con el propósito de eliminar los *itemsets* que no son frecuentes en la base de datos de entrada.

Después de generar el conjunto $L$ que contiene todos los *itemsets* frecuentes, puede comenzar la fase de extracción de reglas de asociación. Para esto, el Algoritmo Apriori utiliza la función *qp-gerardes*, mostrada a continuación.

En esta función, el conjunto $L$ se recibe como argumento junto con el valor *min_conf* (valor de confianza mínima requerida para las reglas). A partir de cada *itemset* presente en $L$, se extraen todos los subconjuntos de ítems posibles, y para cada subconjunto, se calcula la confianza de la regla a generar. Si el valor de la confianza satisface el mínimo requerido, la regla se extrae y se especifica como $Y \rightarrow (X - Y)$. Así, todas las posibles reglas de asociación se generan a partir del conjunto de *itemsets* frecuentes y se muestran al final de la ejecución del algoritmo.

### 2.3.4 *Estado del Arte de las Reglas de Asociación*

Las primeras reglas de asociación, propuestas por Agrawal en 1993, estaban basadas en datos booleanos, es decir, los ítems se representaban de manera binaria: presente o ausente.

**Algoritmo 2.2** Función apriori-gen para generación de candidatos

**Require:** $L_{k-1}$: Conjunto de $(k-1)$-itemsets frecuentes
**Ensure:** $C_k$: Conjunto de $k$-itemsets candidatos

**// Paso 1 (join)**
**for** cada itemset $l_1 \in L_{k-1}$ **do**
  **for** cada itemset $l_2 \in L_{k-1}$ **do**
    **if** $l_1[1] = l_2[1] \wedge \cdots \wedge l_1[k-1] < l_2[k-1]$ **then**
      $c \leftarrow l_1[1] \cdot l_1[2] \cdots l_1[k-2] \cdot l_1[k-1] \cdot l_2[k-1]$
      $C_k \leftarrow C_k \cup \{c\}$
    **end if**
  **end for**
**end for**
**// Paso 2 (prune)**
**for** cada candidato $c \in C_k$ **do**
  **for** cada $(k-1)$-subconjunto $s$ de $c$ **do**
    **if** $s \notin L_{k-1}$ **then**
      $C_k \leftarrow C_k \setminus \{c\}$
      **break**
    **end if**
  **end for**
**end for**
**return** $C_k$

**Algoritmo 2.3** Generación de reglas de asociación (qp-gerardes)

**Require:** $L$: conjunto de itemsets frecuentes, $min\_conf$: confianza mínima
**Ensure:** $R$: conjunto de reglas de asociación

**for** cada $k$-itemset $l \in L$ **do**
  **for** $i = k$ **to** 1 **step** $-1$ **do**
    **for** cada $i$-subconjunto $s$ de $l$ **do**
      conf $\leftarrow \frac{\text{soporte}(l)}{\text{soporte}(s)}$
      **if** conf $\geq min\_conf$ **then**
        $R \leftarrow R \cup \{s \Rightarrow (l \setminus s)\}$
      **end if**
    **end for**
  **end for**
**end for**
**return** $R$

Estas reglas son conocidas como RAB (Reglas de Asociación Booleanas) y utilizan transacciones en las que los ítems se representan como 1 (verdadero) o 0 (falso). Por ejemplo, una regla como {Pan} $\Rightarrow$ {Mantequilla} indicaría que, si un cliente compra pan, hay una alta probabilidad de que también compre mantequilla.

Con la evolución de las RAB, surgió la necesidad de trabajar con datos numéricos y categóricos, además de los binarios. Las reglas de asociación numéricas permiten el análisis de datos que incluyen salarios, edades y otros atributos cuantitativos o cualitativos. Estas reglas representan una evolución significativa de las reglas booleanas tradicionales, al permitir la inclusión de variables que no se limitan a valores binarios. Un ejemplo de regla numérica podría ser:

$$\text{Edad} \in [21, 35] \wedge \text{Género} : [\text{Masculino}] \Rightarrow \text{Salario} \in [2000, 3000]$$

lo que indicaría que los empleados con edades entre 21 y 35 años y que son hombres ganan entre 2.000 y 3.000. Las reglas numéricas permiten un análisis más detallado, ya que pueden involucrar variables continuas y categóricas (Aggarwal y Yu, 1998).

Por otro lado, las reglas difusas (fuzzy) representan una evolución de las reglas binarias y numéricas. En lugar de tratar con datos precisos y bien definidos, las reglas difusas operan con conjuntos difusos. En esta metodología, tanto el antecedente como el consecuente pueden representarse como valores difusos, y la interpretación de los datos se realiza en términos de grados de pertenencia, como se discutió anteriormente.

Las reglas de asociación han ganado mayor protagonismo en la nueva era, en la que personas y máquinas están cada vez más equipadas con sensores que generan datos de monitoreo o alerta. Estas reglas son fundamentales por su capacidad de procesar y extraer información útil de estos datos, con el objetivo de mejorar el proceso de toma de decisiones y, así, reducir la frecuencia de decisiones equivocadas y sus costos asociados.

Las reglas de asociación, especialmente en su forma tradicional (como el algoritmo Apriori) y sus variantes difusas, han sido ampliamente aplicadas por diversos autores en distintas áreas, con el objetivo de extraer patrones ocultos y relevantes a partir de grandes volúmenes de datos. Estas reglas permiten descubrir relaciones frecuentes entre conjuntos de variables, lo que las convierte en una herramienta poderosa de análisis exploratorio y apoyo a la decisión. En 2025, varios estudios destacaron su utilidad práctica en contextos como la salud, la ortopedia y el comercio electrónico, cada uno explorando sus ventajas específicas.

En el artículo de Liao (2025), las reglas de asociación se aplican en el contexto del comercio electrónico con transmisión en vivo (*live streaming*). El autor propone un sistema de soporte a la decisión basado en minería de reglas de asociación difusas, que busca mejorar la visibilidad de productos y la eficacia de las promociones. En este sistema, se definen diferentes niveles de rendimiento con base en la tasa de acceso de los consumidores y la cantidad de ventas. A través de la aplicación de reglas difusas, el sistema puede adaptar las promociones de manera dinámica, considerando caídas en los indicadores de rendimiento para ajustar las estrategias de exposición de productos. Una de las principales ventajas de este modelo es su capacidad para lidiar con incertidumbres y variabilidades en el comportamiento del consumidor, algo que los métodos clásicos difícilmente capturan con precisión. Además, el modelo difuso permite una mayor flexibilidad y sensibilidad en la interpretación de datos, lo cual es crucial en entornos de mercadotecnia digital altamente competitivos y volátiles.

En el campo de la salud, las reglas de asociación se aplican para descubrir relaciones entre marcadores clínicos y desenlaces terapéuticos. Por ejemplo, en un estudio de S.-C. Chen et al. (2025) sobre el carcinoma hepatocelular (HCC) en pacientes que no secretan la proteína AFP, los autores utilizan estas reglas para demostrar una fuerte asociación entre la respuesta del marcador PIVKA-II y la progresión de la enfermedad, así como con la supervivencia libre de progresión y la supervivencia global. La principal ventaja aquí

es que las reglas de asociación permiten identificar biomarcadores sustitutos de manera objetiva, con base en la frecuencia y fuerza estadística de las correlaciones. Esto tiene gran valor clínico, ya que ofrece una forma no invasiva y más accesible de monitorear el progreso del tratamiento, ayudando a los médicos a tomar decisiones más informadas sin depender exclusivamente de exámenes radiológicos, que son más costosos y no siempre concluyentes.

En el área de ortopedia, específicamente en el estudio de Dissaneewate et al. (2025), se empleó un enfoque basado en reglas de asociación para revisar y validar los criterios de Lafontaine, tradicionalmente utilizados para predecir la inestabilidad de fracturas del radio distal. Utilizando análisis retrospectivo y técnicas de regresión, los autores identificaron factores con fuerte asociación a la inestabilidad, como el acortamiento radial y la presencia de fractura cubital asociada. Aunque el enfoque principal del estudio fue la regresión logística, el espíritu de las reglas de asociación está presente en la búsqueda de patrones recurrentes en grandes conjuntos de datos clínicos, con el objetivo de proponer criterios más robustos y con mayor poder predictivo. La ventaja aquí es la mejora de la precisión diagnóstica y la confiabilidad de los criterios clínicos, lo que puede reducir intervenciones quirúrgicas innecesarias o fallos en tratamientos conservadores.

De forma general, las principales ventajas de las reglas de asociación incluyen:

- La capacidad de revelar relaciones ocultas y no triviales entre variables en grandes volúmenes de datos;
- La simplicidad en la interpretación de los resultados, lo que las hace accesibles incluso para usuarios sin formación estadística avanzada;
- La flexibilidad de aplicación en contextos diversos, desde el marketing digital hasta la medicina personalizada;
- La posibilidad de integración con otras técnicas, como algoritmos difusos o modelos de regresión, para aumentar su robustez y adaptabilidad.

Las áreas más comunes de aplicación de estas reglas incluyen:

1. Salud y Medicina – para la identificación de factores de riesgo, biomarcadores, perfiles clínicos y patrones de respuesta al tratamiento;
2. Comercio Electrónico y Marketing – para el análisis del comportamiento del consumidor, personalización de ofertas y optimización de promociones;
3. Diagnóstico Clínico y Ortopedia – para la construcción o revisión de criterios diagnósticos y predictivos basados en patrones empíricos;
4. Ciencias Sociales y Psicología – en estudios sobre comportamiento, hábitos de consumo o patrones de opinión en grandes poblaciones;
5. Educación y Tecnologías de Aprendizaje – para la detección de patrones de rendimiento estudiantil, recomendación de recursos didácticos y análisis de interacciones en plataformas digitales.

### 2.3.5 *Sistemas Fuzzy*

La lógica difusa fue propuesta por Lotfi A. (Zadeh, 1965). Los sistemas difusos permiten interpretar información sensorial imprecisa e incompleta proporcionada por los respectivos órganos sensoriales. Esta lógica ofrece soporte para sistemas de control y procesos de toma de decisiones.

La teoría de conjuntos difusos proporciona un método sistemático para manejar información expresada lingüísticamente, realizando cálculos numéricos con base en etiquetas lingüísticas definidas por funciones de pertenencia. Además, la selección de reglas difusas del tipo SI-ENTONCES constituye el componente clave de un sistema de inferencia difusa, capaz de modelar eficazmente la experiencia humana en aplicaciones específicas.

La aplicación de los sistemas difusos es bastante amplia, siendo útil para convertir información subjetiva, normalmente expresada a través del lenguaje natural, en representaciones matemáticas. Por ejemplo, si un investigador es capaz de describir las causas de la ocurrencia de un fenómeno utilizando reglas del tipo SI-ENTONCES, es posible crear un patrón que puede ser implementado computacionalmente y descubrir las relaciones existentes.

A diferencia de la lógica booleana tradicional, que opera con valores binarios (0 o 1), la lógica difusa permite un conjunto continuo de valores, representando diferentes grados de verdad.

El lenguaje natural es el medio de comunicación más eficaz utilizado por los seres humanos, especialmente en contextos que requieren razonamiento y toma de decisiones. Por esta razón, la lógica difusa se adapta bien al lenguaje natural.

Los sistemas difusos crean estructuras que representan el conocimiento, al integrar expresiones lingüísticas y datos numéricos, combinando restricciones y mecanismos de inferencia de una manera más flexible y cercana al razonamiento humano.

La lógica difusa es un enfoque eficaz para asociar un conjunto de entradas a un conjunto de salidas, funcionando de manera similar a una "caja negra" que realiza un procesamiento interno para generar una respuesta, es decir, la salida del sistema.

Para ilustrar lo que significa mapear el espacio de entrada al espacio de salida, consideremos el siguiente ejemplo: un investigador estudia las condiciones de oposición a la inmigración en los EE. UU. Basado en este análisis, expresa de manera difusa como "muy fuerte" o "razonable", el sistema difuso puede sugerir un valor apropiado para la ideología (ver Figura 2.6). Entre el conjunto de entradas y el conjunto de salidas, existe un mecanismo denominado "caja negra", responsable de procesar la información y generar una respuesta.


Figura 2.6: Ejemplo de mapeo del espacio de entrada y salida

La elección de la lógica difusa para clasificar diferentes sistemas cotidianos se justifica por su simplicidad de interpretación, bajo tiempo de ejecución, rapidez en el procesamiento y costos de implementación significativamente inferiores en comparación con los métodos alternativos.

La principal metodología utilizada para relacionar los espacios de entrada con el espacio de salida, a través de la caja negra, se basa en conjuntos de reglas o condiciones del tipo SI-ENTONCES. Para construir un sistema capaz de interpretar estas reglas, es necesario definir previamente todos los atributos o elementos que se utilizarán, así como las características o términos lingüísticos que los describen.

Volviendo al ejemplo de la oposición a la inmigración en los EE. UU., es esencial establecer categorías que representen diferentes niveles de oposición a los inmigrantes, permitiendo así evaluar la ideología subyacente.

### 2.3.6 *Conjuntos Fuzzy*

El *sharp boundary problem* (SBP) es un problema de limitación asociado a la eficiencia de los sistemas expertos. Ocurre cuando existe una división rígida entre categorías, lo que provoca una subestimación o sobrestimación de elementos que se encuentran cerca de los límites de estas divisiones. Tal situación se debe a la forma en que se particionan los atributos cuantitativos, impactando directamente en la precisión de los sistemas expertos (Verlinde et al., 2006).

En la lógica clásica, la asociación de un elemento a un conjunto se trata de forma binaria: o el elemento pertenece al conjunto, o no pertenece. Los límites entre los conjuntos se definen de manera exacta (*crisp*), sin espacio para interpretaciones intermedias.

Así, la idea de pertenencia está bien delimitada: dado un conjunto $A$ dentro de un universo $X$, cada elemento de $X$ o está en $A$, o está fuera de él. Este concepto se representa por una función característica $f_A(x)$, que asume el valor 1 cuando el elemento pertenece al conjunto y 0 cuando no pertenece.

En la lógica clásica, la pertenencia de un elemento $x$ a un conjunto $A$ se define por una función característica $f_A(x)$ de la siguiente manera:

$$f_A(x) = \begin{cases} 1, & \text{si } x \in A \\ 0, & \text{si } x \notin A \end{cases} \tag{3}$$

Aunque los conjuntos de la lógica clásica son adecuados para diversas aplicaciones y han demostrado ser una herramienta importante para las matemáticas y la informática, muchas veces no reflejan la naturaleza de los conceptos y pensamientos humanos, que tienden a ser abstractos e imprecisos. Por ejemplo, matemáticamente podemos expresar el conjunto de personas altas como el conjunto de personas cuya altura es superior a 170 cm. Sin embargo, si tenemos dos personas, una con 170.1 cm y otra con 169.9 cm de altura, la lógica *crisp* clasificaría a la primera persona como alta y a la segunda como baja. Esta distinción es intuitivamente irracional, pues el fallo viene de la transición brusca entre la inclusión y la exclusión de un conjunto. No tiene sentido considerar una diferencia de solo 0.2 cm como un criterio tan decisivo, ya que el razonamiento humano no funciona de esa manera. No sería lógico decir que una persona con 170.1 cm es alta mientras que otra con 169.9 cm es baja.

En el conjunto *fuzzy*, la transición de "pertenecer a un conjunto" a "no pertenecer a un conjunto" es gradual. Esta transición suave se caracteriza por funciones de pertenencia, que confieren a los conjuntos difusos flexibilidad en la modelización de expresiones lingüísticas comúnmente utilizadas, como "el agua está caliente" o "la temperatura es elevada". Como Zadeh (1965) señaló en su artículo titulado *Fuzzy Sets*, tales conjuntos o clases definidos de forma imprecisa desempeñan un papel importante en el pensamiento humano, particularmente en los dominios del reconocimiento de patrones, la comunicación de información, la abstracción y la investigación académica.

Sin embargo, Zadeh (1965) propuso una generalización de estos conceptos, permitiendo que la función de pertenencia asumiera cualquier valor real en el intervalo $[0, 1]$. Así, un conjunto *fuzzy* $A$ en un universo $X$ se define por una función de pertenencia:

$$\mu_A(x) : X \to [0, 1]$$

Este conjunto *fuzzy* puede representarse como un conjunto de pares ordenados:

$$A = \{(x, \mu_A(x)) \mid x \in X\} \tag{4}$$

En esta definición, $\mu_A(x)$ expresa el grado de pertenencia de $x$ al conjunto difuso $A$. Esto significa que un mismo elemento $x$ puede pertenecer simultáneamente a diferentes conjuntos *fuzzy*, con distintos niveles de compatibilidad.

Normalmente, $X$ se denomina universo del discurso, o simplemente universo, y puede consistir en objetos discretos o espacios continuos.

Por ejemplo, sea

$$X = \{1.55,\ 1.60,\ 1.65,\ 1.70,\ 1.80,\ 1.90\}$$

el conjunto de estaturas que un individuo puede tener. El conjunto *fuzzy* $A =$ "la altura de categoría media de un individuo" puede describirse de la siguiente forma:

$$A = \{(1.55,\ 0.1),\ (1.60,\ 0.2),\ (1.65,\ 0.5),\ (1.70,\ 0.9),\ (1.80,\ 0.4),\ (1.90,\ 0.1)\}.$$

Por lo tanto, tenemos un universo ordenado de variables continuas $X$. La función de pertenencia (*membership function*, MF) para el conjunto *fuzzy* $A$ está representada en color verde en la Figura 2.7. Se observa que los grados de pertenencia del conjunto de alturas medias son medidos de forma difusa.

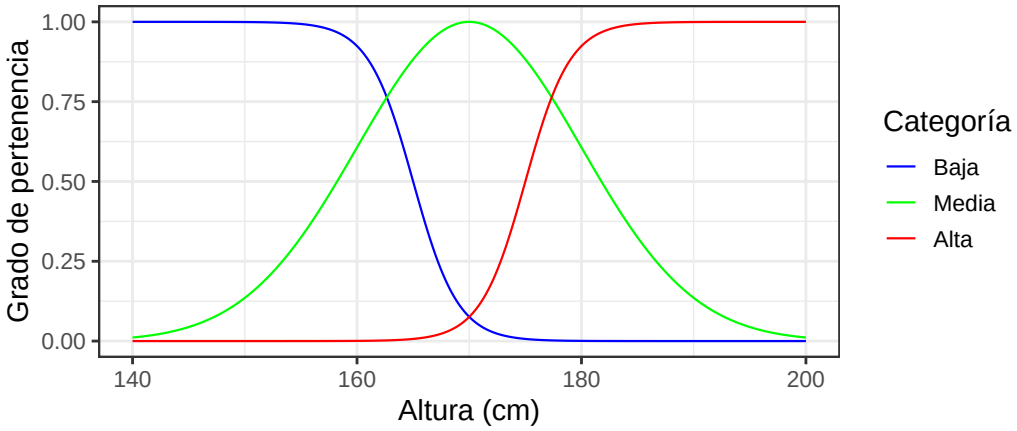


Figura 2.7: Función de pertenencia *fuzzy*

La construcción de un conjunto *fuzzy* depende de dos elementos: (i) la identificación de un universo del discurso o de un término lingüístico adecuado; y (ii) la identificación de una función de pertenencia apropiada. La especificación de una función de pertenencia es subjetiva, lo que significa que diferentes funciones pueden definirse para un mismo concepto (por ejemplo, "individuos con altura media").

Los adjetivos que aparecen en nuestro uso lingüístico en fenómenos de las ciencias sociales, como bajo, medio o alto, se llaman valores lingüísticos o etiquetas lingüísticas. Así, el universo del discurso $X$ se designa frecuentemente como una variable lingüística.

Una variable lingüística es la caracterización de valores mediante palabras, en lugar de números. Los términos lingüísticos reflejan conceptos de experiencias cotidianas, como "caliente", "frío", "cerca" o "muy cerca". Como se vio en el ejemplo de altura de un conjunto de personas, los valores pueden definirse en "baja", "media" y "alta", pero también podrían definirse en "muy baja", "baja", "media", "alta" y "muy alta". Estas variables pueden representarse mediante conjuntos *crisp* o *fuzzy*, con base en una función de pertenencia.

En el ejemplo de alturas, el intervalo de 150 a 200 cm fue segmentado en tres categorías: baja, media y alta.

Una variable lingüística permite la descripción tanto de valores cualitativos como cuantitativos, siendo caracterizada por una quíntupla $(N, T(N), X, G, M)$, donde:

- N: nombre de la variable.
- T(N): conjunto de nombres de los valores lingüísticos asociados a N.
- X: universo del discurso, es decir, el intervalo de valores posibles.
- G: regla sintáctica que describe la generación de los valores de N como una combinación de términos de T(N), conectivos lógicos, modificadores y delimitadores.

- M: regla semántica que asocia a cada valor generado por G un conjunto *fuzzy* en el universo del discurso X.

En el ejemplo de la variable "altura" en el contexto del conjunto de alturas de personas, tenemos:

- N: altura.
- T(N): {baja, media, alta}.
- X: intervalo de 150 a 200cm.
- G: por ejemplo, "no baja y no muy alta".
- M: asocia ese valor a un conjunto *fuzzy* cuya función de pertenencia expresa su significado en el contexto de la altura.

Por otro lado, una función de pertenencia es una curva que mapea cada punto del espacio de entrada a un valor de pertenencia o grado de pertenencia, variando entre 0 y 1. Este valor se denomina $\mu$. En el ejemplo del conjunto de alturas de dos individuos con 169.9cm y 170.1cm, la diferencia entre ellos es pequeña, lo que dificulta establecer una distinción clara. La función de pertenencia es responsable por establecer esa transición de "no tan alto" a "alto", asignando a ambas personas un grado $\mu$ que indica que ambas son altas, aunque una sea algo más baja.

La figura 2.8 ilustra las funciones de pertenencia asociadas a la variable altura. Se nota que, para estaturas de hasta 1.55m, el grado de pertenencia en el conjunto de bajas es casi 1 (color azul). A medida que la altura aumenta, el grado de pertenencia disminuye progresivamente. Para 1.75m, el valor es totalmente compatible con el conjunto de alturas medias (color verde), mientras que alturas superiores a 1.80m presentan grado de pertenencia diferente de cero en el conjunto de personas altas (color rojo). Para alturas superiores a 1.90m, la clasificación como "alta" es inequívoca. En alturas cercanas a 1.75m, el grado de pertenencia diferente de cero aparece solo en el conjunto de personas medianas.

La construcción de funciones de pertenencia puede variar según el grupo de personas, ya que depende de las percepciones y experiencias individuales. Por ejemplo, el concepto de un ciudadano "alto" para el alistamiento militar en Guatemala puede diferir del adoptado por el ejército español, desde la perspectiva de la función de pertenencia usada.

Diversos tipos de funciones de pertenencia son comunes en la modelización de problemas *fuzzy*. Lo fundamental es que estas funciones varíen de 0 a 1 y puedan asumir distintas formas de curvas. El contexto del problema a resolver es el que determina el tipo más adecuado. En general, cuanto más simple, conveniente, rápida y eficiente sea la función, mejor será su desempeño.

A continuación, ilustramos algunos tipos de funciones de pertenencia *fuzzy*, construidas a partir de funciones básicas. Entre ellas, funciones lineales por tramos, funciones de distribución gaussiana, curvas sigmoides, curvas polinomiales cuadráticas y cúbicas. Las funciones más simples son las de línea recta, que tienen la ventaja de la simplicidad.

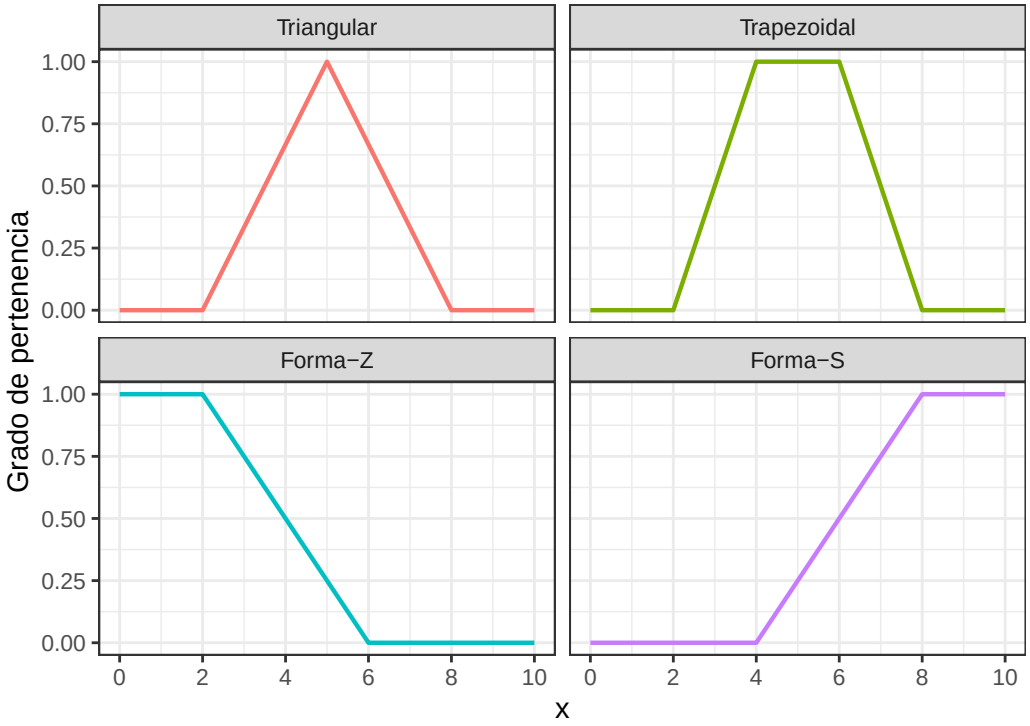


Figura 2.8: Funciones de pertenencia

La mayoría de los investigadores suele preferir las funciones de pertenencia del tipo triangular o trapezoidal, pues han mostrado buenos resultados en la mayoría de aplicaciones, como señaló (Jang y Gulley, 1995).

Además, existen otras opciones, como las funciones basadas en curvas gaussianas. Entre ellas se destacan: la función gaussiana simple (`gaussmf`), que utiliza una sola curva en forma de campana; la función gaussiana compuesta (`gauss2mf`), que combina dos curvas gaussianas; y la función campana generalizada (*generalized bell*), similar a la gaussiana pero con un parámetro adicional que ofrece más flexibilidad en la forma de la curva.

Estas funciones gaussianas y la *generalized bell* son populares en determinados conjuntos *fuzzy* por presentar transiciones suaves (sin rupturas bruscas) y por no alcanzar nunca el valor cero en ningún punto del dominio, como se ilustra en la figura 2.9.

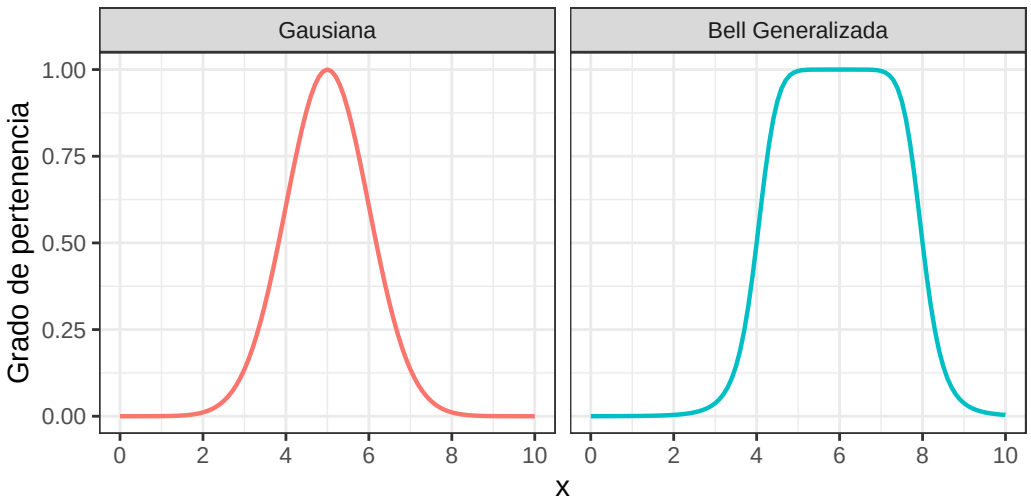


Figura 2.9: Funciones gaussianas

### 2.3.7 *Reglas Fuzzy SI-ENTONCES*

Las reglas difusas son expresiones condicionales del tipo "si-entonces" que utilizan números o términos lingüísticos difusos para representar relaciones imprecisas y graduales entre variables. Estas reglas permiten modelar situaciones complejas, siendo especialmente útiles en la toma de decisiones en contextos marcados por la incertidumbre.

Una regla difusa típica asume la forma: **Si** $x$ **es** $A$, **entonces** $y$ **es** $B$. Aquí, $A$ y $B$ son términos lingüísticos definidos por conjuntos difusos en los universos del discurso $X$ y $Y$, respectivamente. La parte "si $x$ es $A$" se denomina antecedente o premisa, mientras que "entonces $y$ es $B$" se denomina consecuente o conclusión.

Por ejemplo, en el contexto de la oposición a la inmigración en los Estados Unidos, una posible regla difusa podría ser: *Si la oposición a la inmigración es alta, entonces el régimen es Republicano*.

El antecedente representa una interpretación que devuelve un valor numérico o un término difuso, mientras que el consecuente es una asignación que asocia todo el conjunto difuso $B$ a la variable de salida $y$.

La interpretación de una regla difusa "si-entonces" implica dos etapas principales: primero, la evaluación del antecedente, que requiere la difusificación de la entrada y la aplicación de operadores difusos, si es necesario; y luego, la aplicación del consecuente, proceso conocido como implicación.

El antecedente puede incluir múltiples condiciones. Por ejemplo, una regla con más de un antecedente podría ser: *Si (el país tiene un historial de violencia después de las elecciones) y (el gobierno tiene control sobre las operadoras móviles) y (el régimen es una dictadura), entonces (hay corte de internet)*.

La palabra "es" ("is" en inglés) asume roles distintos dependiendo de su posición en la regla: en el antecedente, actúa como una condición que debe ser evaluada; en el consecuente, expresa una asignación a la variable de salida.

Cuando existen múltiples antecedentes, es necesario combinar sus grados de pertenencia para obtener un valor resultante. En los casos en que los antecedentes están conectados por "y" (conjunción), debe utilizarse un método que preserve el menor grado de pertenencia entre ellos, como el mínimo. En cambio, cuando los antecedentes están conectados por "o" (disyunción), se emplea un método que preserve el mayor grado de pertenencia, como el máximo.

Estos métodos de combinación son fundamentales para la lógica difusa, según lo descrito por Jang y Gulley (1995), ya que garantizan coherencia e interpretabilidad en la modelación de reglas en ambientes inciertos.

## 2.4 COMPARACIÓN METODOLÓGICA ENTRE QCA, ARM Y TÉCNICAS ESTADÍSTICAS CLÁSICAS

En la búsqueda de enfoques metodológicos más robustos para el análisis de datos sociales complejos, diversos estudios han explorado la comparación y la combinación de métodos configuracionales, como el Análisis Cualitativo Comparativo (QCA), con técnicas estadísticas tradicionales, tales como los Modelos de Regresión Logística (LRM), las Reglas de Asociación (AR) y los Modelos de Ecuaciones Estructurales (SEM). Esta diversidad de investigaciones refleja un creciente interés por integrar herramientas que permitan capturar tanto la complejidad causal como la generalización estadística.

La Tabla 2.6 presenta estudios realizados por diversos autores que buscan comparar o integrar diferentes metodologías, aclarando los paralelismos, distinciones, fortalezas y debilidades de las técnicas estadísticas clásicas en comparación con los métodos configuracionales emergentes. Ejemplos notables dentro de estos estudios incluyen el examen del QCA frente a modelos de regresión lineal (LRM) (Ragin, Shulman et al., 2003; Seawright, 2005b; Vis, 2012), AR frente a LRM (Buxton et al., 2019; Changpetch y Lin, 2013), QCA frente a modelos de ecuaciones estructurales (SEM) (J. Zhao y Yan, 2020), y AR frente a SEM (Mguirris et al., 2015). Esta tabla ofrece un resumen conciso de las investigaciones más significativas sobre la comparación entre métodos configuracionales y enfoques estadísticos clásicos.

A pesar de la abundancia de estudios que aplican tanto la metodología QCA como la minería de reglas de asociación, y de las similitudes evidentes entre ambas técnicas, hasta donde sabemos, no existe un estudio que intente relacionar de manera precisa csQCA y ARM. En la siguiente sección, estableceremos un vínculo entre ambas técnicas mediante una formalización matemática precisa del csQCA.

Tabla 2.6: Resumen de estudios seleccionados que comparan o combinan QCA, Reglas de Asociación (AR), Modelos de Regresión Logística (LRM) y Modelos de Ecuaciones Estructurales (SEM).

| **Referencia** | **Área** | **Tamaño de muestra** | **Objetivo** | **QCA** | **AR** | **LRM** | **SEM** | **Principales hallazgos** |
|---|---|---|---|---|---|---|---|---|
| Ragin, Shulman et al. (2003) | Sociología | 41 aldeas en India | Analizar fortalezas y debilidades del QCA comparado con regresión. | ✓ | | ✓ | | QCA capta complejidad; regresión mide efectos netos. Son complementarios. |
| Seawright (2005b) | Ciencia Política | ∼80–100 elecciones | Comparar inferencia causal en QCA y regresión. | ✓ | | ✓ | | QCA requiere supuestos tan fuertes como la regresión. |
| Buxton et al. (2019) | Biomedicina | 65.4K pacientes | Descubrir correlaciones usando AR y regresión. | | ✓ | ✓ | | AR halla patrones; LRM elimina asociaciones falsas. |
| Changpetch y Lin (2013) | Computación | 432 robots | Seleccionar modelos óptimos con AR y regresión. | | ✓ | ✓ | | AR + LRM mejora ajuste e interpretabilidad. |
| J. Zhao y Yan (2020) | Computación | 229 usuarios de Internet | Analizar efectos individuales y combinados. | ✓ | | | ✓ | QCA + SEM mejora robustez interpretativa. |
| Vis (2012) | Ciencia Política | 53 gobiernos | Evaluar fortalezas y límites de QCA y regresión. | ✓ | | ✓ | | Ambos útiles para muestras medianas. |

# 3

# CONECTANDO EL ANÁLISIS CUALITATIVO COMPARATIVO DE CONJUNTOS NÍTIDOS Y LA MINERÍA DE REGLAS DE ASOCIACIÓN

En este capítulo presentaremos uno de los principales resultados de esta tesis, publicado recientemente en Dom Luís, Benítez y M. d. C. Bas (2025). En él, discutimos las similitudes y diferencias entre el Análisis Cualitativo Comparativo de Conjuntos Crisp (csQCA) y la Minería de Reglas de Asociación (ARM), dos metodologías comúnmente utilizadas para identificar patrones lógicos en conjuntos de datos binarios. Para ello, en primer lugar, establecemos por primera vez en la literatura una formulación matemática rigurosa de lo que es un modelo de csQCA. A continuación, damos la definición estándar de un problema de minería de reglas de asociación binarias, para poder establecer una equivalencia matemática formal entre las configuraciones de csQCA y un subconjunto específico de reglas de asociación, que incluye tanto reglas positivas como negativas (Negative and Positive Association Rules (NPAR)). Además, a partir de la definición de reglas de asociación interesantes, enunciamos y demostramos un teorema que nos permitirá proponer un algoritmo de minimización para reglas de asociación que refleja el método de reducción de Quine-McCluskey, utilizado en csQCA.

Terminaremos el capítulo ilustrando dos ejemplos, uno con un pequeño número de casos (*"small-N"*) sobre cortes de internet en África Subsahariana en periodo de elecciones y otro con un número de casos grande *"large-N"* sobre actitudes hacia la inmigración en Europa. Con ellos mostramos que ambas metodologías producen resultados consistentes, aunque ARM ofrece mayor escalabilidad y robustez en contextos de alta dimensionalidad. Asimismo, completaremos el estudio realizando una serie de experimentos numéricos para comparar cuantitativamente la escalabilidad de los algoritmos propios del csQCA y los de la ARM.

## 3.1 FORMALIZACIÓN MATEMÁTICA DE CSQCA

En esta sección, presentamos una descripción matemática de la metodología csQCA.

**Definición 3.1.** Un modelo csQCA es una **terna** $(\mathcal{V}, R, M)$, donde $\mathcal{V} = \{V_1, \ldots, V_n\}$ es un conjunto de $n$ variables, $R$ es una variable resultado o de respuesta (es decir, otra variable tal que $R \notin \mathcal{V}$), y $M$ es el conjunto de datos muestrales, que es una matriz de $m \times (n+1)$ de la forma

$$M = \begin{array}{c|cccc|c} & V_1 & V_2 & \cdots & V_n & R \\ \hline \mathbf{c}_1 & x_{11} & x_{12} & \cdots & x_{1n} & y_1 \\ \mathbf{c}_2 & x_{21} & x_{22} & \cdots & x_{2n} & y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{c}_m & x_{m1} & x_{m2} & \cdots & x_{mn} & y_m \end{array}, \tag{3.1}$$

donde $x_{ij}$, $y_i \in \{0,1\}$, para todo $i \in \{1,\dots m\}$, $j \in \{1,\dots n\}$, donde 1 indica la presencia, y 0 indica la ausencia de cada condición/variable y del resultado.

Cada fila de la matriz $M$ representa un **caso** muestral y se denotará por

$$\mathbf{c_i} = (\mathbf{x}_i; y_i), \quad i = 1,\dots,m,$$

siendo $\mathbf{x}_i = (x_{i1},\dots,x_{in})$ el vector antecedente, y $y_i$ el valor consecuente (o resultado).

**Definición 3.2.** Decimos que un conjunto de datos csQCA $M$ es **consistente** si para cualesquiera dos casos $\mathbf{c_{i_1}} = (\mathbf{x}_{i_1}; y_{i_1})$, y $\mathbf{c_{i_2}} = (\mathbf{x}_{i_2}; y_{i_2})$, si $\mathbf{x}_{i_1} = \mathbf{x}_{i_2}$, entonces, necesariamente, $y_{i_1} = y_{i_2}$. Esto significa que no hay contradicciones entre los casos muestrales.

**Definición 3.3.** Una **configuración** de orden $k$ ($k \leq n$) es un par $\mathcal{S}_k = (\mathcal{V}_k, \mu)$, donde $\mathcal{V}_k \subset \mathcal{V}$ con $|\mathcal{V}_k| = k$, y $\mu : \mathcal{V}_k \to \{0,1\}$.

**Definición 3.4.** Un caso $\mathbf{c}_i = (\mathbf{x}_i; y_i)$ es **compatible** con la configuración $\mathcal{S}_k = (\mathcal{V}_k, \mu)$ si

$$\mu(V_{j_l}) = x_{ij_l}, \quad \forall l \in \{1,\dots,k\}$$

**Definición 3.5.** Una configuración $\mathcal{S}_k$ es **conducente** al resultado $R$ si existe al menos un caso $\mathbf{c}_i = (\mathbf{x}_i; y_i)$, compatible con $\mathcal{S}_k$, para el cual $y_i = 1$.

*Observación* 3.1. Nótese que todas las posibles $n$-configuraciones (conducentes o no conducentes al resultado R) pueden escribirse como una tabla de verdad con $2^{n+1}$ filas. Cada fila de la tabla de verdad corresponde a una combinación específica de condiciones, y colectivamente, todas las filas muestran las posibles combinaciones lógicas de condiciones usando la operación lógica AND. Estas configuraciones se observan en uno o más casos (filas de la matriz $M$) y la suficiencia se determina por la asociación de estos casos con el resultado $R$ (Oana et al., 2021).

**Definición 3.6.** Dos configuraciones de orden $k$, $\mathcal{S}_k^{(1)} = \left(\mathcal{V}_k^{(1)}, \mu^{(1)}\right)$, y $\mathcal{S}_k^{(2)} = \left(\mathcal{V}_k^{(2)}, \mu^{(2)}\right)$ pueden **minimizarse** (o reducirse) a una configuración de orden $k-1$ si

- Ambas son conducentes al resultado $R$,
- $\mathcal{V}_k^{(1)} = \mathcal{V}_k^{(2)} = \{V_{j_1},\dots,V_{j_k}\}$, y
- existe un $p \in \{1,\dots,k\}$ tal que

$$|(\mu_1(V_{j_1}), \mu_1(V_{j_2}),\dots,\mu_1(V_{j_k})) - (\mu_2(V_{j_1}), \mu_2(V_{j_2}),\dots,\mu_2(V_{j_k}))| = \mathbf{e}_p, \tag{3.2}$$

donde $\mathbf{e}_p \in \mathbb{R}^k$ es el $p$-ésimo vector de la base canónica (es decir, $\mathbf{e}_p = (0,\ldots,0,1,0,\ldots,0)$)

En ese caso, $\mathcal{S}_k^{(1)}$ y $\mathcal{S}_k^{(2)}$ pueden reducirse a la $(k-1)$-configuración $\mathcal{S}_{k-1} = (\mathcal{V}_{k-1}, \mu)$, donde

$$\mathcal{V}_{k-1} = \mathcal{V}_k^{(1)} \setminus \{V_{j_p}\} = \{V_{j_1},\ldots,V_{j_{p-1}},V_{j_{p+1}},\ldots,V_{j_k}\},$$

y

$$\mu = \mu^{(1)}\Big|_{\mathcal{V}_{k-1}} = \mu^{(2)}\Big|_{\mathcal{V}_{k-1}}.$$

Nótese que de (3.2), se infiere que para todo $V \in \mathcal{V}_{k-1}$, $\mu^{(1)}(V) = \mu^{(2)}(V)$.

Si dos configuraciones no pueden minimizarse, diremos que son **irreducibles**.

El objetivo en csQCA es obtener todas las configuraciones irreducibles que conducen a un cierto resultado. Es decir, podemos definir el conjunto solución de un problema csQCA como

**Definición 3.7.** Las **soluciones** del modelo csQCA $(\mathcal{V}, R, M)$ vienen dadas por

$$\mathcal{C} = \bigcup_{k=1}^{n} \mathbf{C}_k$$

donde cada conjunto $\mathbf{C}_k$ es la familia de todas las configuraciones irreducibles de orden $k$ ($\mathbf{C}_k$ puede ser vacío) conducentes al resultado $R$, es decir, $\mathbf{C}_k$ viene dado por

$$\mathbf{C}_k = \bigcup_{p=1}^{n_k} \mathcal{S}_k^{(p)},$$

donde

- $\mathcal{S}_k^{(p)}$ es conducente al resultado $R$ para todo $p \in \{1,\ldots,n_k\}$
- $\mathcal{S}_k^{(p_1)}$ y $\mathcal{S}_k^{(p_2)}$ son irreducibles $\forall p_1, p_2 \in \{1,\ldots,n_k\}$ con $p_1 \neq p_2$.

En csQCA, evaluar la fiabilidad y relevancia de configuraciones causales requiere el uso de métricas clásicas, concretamente **consistencia** y **cobertura**. Estas métricas ayudan a determinar la fuerza y poder explicativo de los patrones identificados. La consistencia mide el grado en que los casos que apoyan una configuración dada también exhiben el resultado, asegurando que la solución es lógicamente coherente con el conjunto de datos. Es decir,

**Definición 3.8.** Dada una configuración $\mathcal{S}_k = (\mathcal{V}_k, \mu)$, con $\mathcal{V}_k = \{V_{j_1}, V_{j_2}, \ldots, V_{j_k}\}$, entonces la **consistencia** de la configuración se define como

$$\text{cons}(\mathcal{S}_k) = \frac{n_{xy}}{n_x}, \tag{3.3}$$

donde

$$n_x = |\{\mathbf{c} = (\mathbf{x}, y) : \mathbf{c} \text{ es compatible con } \mathcal{S}_k\}| \tag{3.4}$$
$$n_{xy} = |\{\mathbf{c} = (\mathbf{x}, y) : \mathbf{c} \text{ es compatible con } \mathcal{S}_k \text{ y } y = 1\}|$$

Es decir, $n_x$ representa el número de casos en el conjunto de datos muestrales compatibles con la configuración $\mathcal{S}_k$, independientemente del resultado del caso, y, por otro lado, $n_{xy}$ es el número de casos en el conjunto de datos muestrales compatibles con $\mathcal{S}_k$ para los cuales el resultado es positivo.

Sin embargo, la cobertura evalúa cuánto del resultado observado es explicado por una configuración específica, proporcionando información sobre la generalizabilidad de los hallazgos, como se describe en la siguiente definición.

**Definición 3.9.** La **cobertura** de una configuración causal se define como la proporción de casos con resultado positivo compatibles con la configuración dada. Es decir,

$$\text{cov}(\mathcal{S}_k) = \frac{n_{xy}}{n_y}, \tag{3.5}$$

donde $n_{xy}$ se define como antes y $n_y$ es el número de casos en el conjunto de datos muestrales para los cuales el resultado es 1, es decir

$$n_y = |\{\mathbf{c} = (\mathbf{x}, y) : y = 1\}| \tag{3.6}$$

Una configuración con alta consistencia indica un fuerte vínculo causal, mientras que una alta cobertura sugiere que la configuración explica una porción significativa de los casos con el resultado.

Una configuración $\mathcal{S}_k$ es compatible con el conjunto de datos muestrales $M$ si es compatible con al menos un caso del conjunto de datos muestrales. En caso contrario, diremos que la configuración y el conjunto de datos muestrales son incompatibles.

Obviamente tendremos que, dada una configuración $\mathcal{S}_k$:

1. Si $\mathcal{S}_k$ es compatible con $M$, entonces $\text{cons}(\mathcal{S}_k) > 0$.
2. Si $\mathcal{S}_k$ es conducente al resultado $R$, y el conjunto de datos muestrales $M$ es consistente (no hay contradicciones), entonces $\text{cons}(\mathcal{S}_k) = 1$.

El procedimiento para encontrar el conjunto solución $\mathcal{C}$ se basa en el algoritmo de minimización booleana desarrollado en la década de 1950 por Quine y McCluskey en el contexto de la optimización de circuitos electrónicos (McCluskey, 1956; Quine, 1952, 1955). Este procedimiento es un algoritmo de arriba hacia abajo que comienza con las configuraciones maximales, es decir, las configuraciones de orden $n$ (siendo $n$ el número de variables en el conjunto $\mathcal{V}$), y luego para cada tamaño $k$, encuentra las configuraciones de orden $k$ reducibles a configuraciones de orden $k-1$. El procedimiento se detiene una vez que se alcanzan las configuraciones de orden 1, o en un paso determinado, las configuraciones de orden $k$ son el conjunto vacío.

El cálculo de las soluciones csQCA se basa en el análisis de tablas de verdad, que enumeran sistemáticamente todas las configuraciones posibles de condiciones causales. Aunque este enfoque garantiza una evaluación exhaustiva de las relaciones causales, también presenta un desafío computacional significativo. Dado que el número de configuraciones posibles crece exponencialmente con el número de variables (específicamente, como $2^n$ para $n$ variables binarias), el análisis de grandes conjuntos de datos con numerosas condiciones causales se vuelve cada vez más inviable. Esta explosión combinatoria puede llevar a tiempos de procesamiento excesivos y limitaciones computacionales, especialmente en casos donde el conjunto de datos incluye docenas de variables.

## 3.2 DESCRIPCIÓN MATEMÁTICAS DE LA MINERÍA DE REGLAS DE ASOCIACIÓN

Matemáticamente, ARM opera sobre una base de datos transaccional $T$, donde cada transacción se representa mediante un vector binario que codifica la presencia (1) o ausencia (0) de ítems de un conjunto universal $I$. Más precisamente, sea $I = \{I_1, I_2, \ldots, I_m\}$ un conjunto universal de **ítems**, y sea $T = \{\tau_1, \tau_2, \ldots, \tau_n\}$ una base de datos transaccional donde cada transacción $\tau_i \subseteq I$. Entonces tenemos lo siguiente:

**Definición 3.10.** Una **regla de asociación** (R. Agrawal et al., 1993) es una implicación de la forma:

$$X \Rightarrow Y \quad \text{donde} \quad X \subseteq I,\ Y \subseteq I,\ X \cap Y = \emptyset,\ \text{y}\ X, Y \neq \emptyset.$$

Esta regla indica que la ocurrencia del conjunto de ítems $X$ en una transacción implica la ocurrencia del conjunto de ítems $Y$ en la misma transacción, es decir, para toda transacción $\tau \in T$,

$$X \subseteq \tau \longrightarrow Y \subseteq \tau.$$

Análogamente a los problemas csQCA, la extracción de reglas de asociación significativas se basa en criterios cuantitativos para evaluar su significancia estadística y relevancia práctica. Dos métricas fundamentales sirven como filtros críticos para distinguir patrones robustos de correlaciones espurias: **soporte** y **confianza**. El soporte asegura que las reglas reflejen co-ocurrencias frecuentemente observadas, mientras que la confianza mide la fiabilidad de la implicación. Matemáticamente:

**Definición 3.11.** Dada una regla de asociación $X \Rightarrow Y$, definimos:

- **Soporte**: La proporción de transacciones en $T$ que contienen tanto $X$ como $Y$:

$$\text{supp}(X \Rightarrow Y) = \frac{|\{\tau \in T \mid X \cup Y \subseteq \tau\}|}{|T|}.$$

- **Confianza**: La probabilidad condicional de que una transacción que contiene $X$ también contenga $Y$:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{|\{\tau \in T \mid X \cup Y \subseteq \tau\}|}{|\{\tau \in T \mid X \subseteq \tau\}|}.$$

Estas métricas permiten a los investigadores priorizar reglas que son tanto prevalentes en el conjunto de datos como probabilísticamente robustas, equilibrando así generalidad y fuerza predictiva. Al establecer umbrales para estos valores, los profesionales pueden descartar sistemáticamente asociaciones triviales o coincidentales, centrándose en cambio en insights accionables con validez empírica.

En una regla de asociación $X \Rightarrow Y$, $X$ es el **antecedente** (lado izquierdo, LHS) e $Y$ es el **consecuente** (lado derecho, RHS).

**Definición 3.12.** La regla $X \Rightarrow Y$ se considera **fuerte** o **interesante** si satisface umbrales definidos por el usuario tanto para soporte (frecuencia) como para confianza (fiabilidad), comúnmente denominados *soporte mínimo* y *confianza mínima*.

En la práctica, la identificación de reglas interesantes (aquellas que satisfacen los umbrales mínimos de soporte y confianza) se basa en algoritmos eficientes capaces de manejar grandes conjuntos de datos. Uno de los métodos más utilizados para este propósito es el algoritmo Apriori, que descubre sistemáticamente conjuntos de ítems frecuentes que sirven como base para generar dichas reglas.

El algoritmo Apriori es una técnica fundamental en Minería de Reglas de Asociación (ARM) utilizada para descubrir conjuntos de ítems frecuentes y generar reglas de asociación a partir de grandes conjuntos de datos. Opera identificando sistemáticamente conjuntos de ítems cuya frecuencia supera un umbral de soporte mínimo predefinido. A partir de estos conjuntos de ítems frecuentes, se generan reglas de asociación, siempre que su confianza supere un nivel mínimo especificado por el usuario, garantizando así tanto frecuencia como fiabilidad.

El algoritmo procede iterativamente: comienza con 1-ítemsets y se expande a conjuntos de ítems más grandes uniendo conjuntos de ítems frecuentes previamente descubiertos. Para mantener la eficiencia computacional, Apriori aprovecha la **propiedad de clausura hacia abajo**, que establece que si un conjunto de ítems es frecuente, entonces todos sus subconjuntos también deben ser frecuentes. Recíprocamente, si un conjunto de ítems se considera infrecuente, todos sus superconjuntos pueden podarse de manera segura sin consideración adicional. A pesar de su efectividad, el algoritmo Apriori puede volverse computacionalmente costoso con grandes conjuntos de datos debido a la generación de numerosos conjuntos de ítems candidatos. Para abordar esto, se han desarrollado varios algoritmos alternativos para mejorar la eficiencia y escalabilidad. Alternativas notables incluyen:

- FP-Growth (Frequent Pattern Growth), que elimina la necesidad de generación de candidatos mediante el uso de una estructura de árbol compacta para almacenar transacciones (Han et al., 2000).
- ECLAT (Equivalence Class Clustering and bottom-up Lattice Traversal), que se basa en una estrategia de búsqueda en profundidad y operaciones de intersección para encontrar conjuntos de ítems frecuentes (M. Zaki, 2000).
- CHARM (Closed Association Rule Mining), que se centra en descubrir conjuntos de ítems cerrados para reducir redundancia (M. J. Zaki y Hsiao, 2002).

- RARM (Rapid Association Rule Mining), optimizado para conjuntos de datos de alta dimensionalidad mediante el aprovechamiento de estructuras de datos especializadas (Amitabha Das et al., 2001).

Estos algoritmos avanzados ofrecen mejoras de rendimiento sobre Apriori, particularmente en el manejo de escenarios con datos a gran escala y de alta dimensionalidad.

## 3.3 EL TEOREMA DE EQUIVALENCIA

Consideremos un modelo csQCA $(\mathcal{V}, R, M)$, con $\mathcal{V} = \{V_1, \ldots, V_n\}$, y $M = (\mathbf{x}_{ij}; y_i)$, para $i \in \{1, \ldots, m\}$, y $j \in \{1, \ldots, n\}$. Entonces definimos el siguiente problema AR:

- El conjunto de ítems será $I = \{V_1, v_1, V_2, v_2, \ldots, V_n, v_n, R, r\}$, con $2n + 2$ ítems.
- La base de datos transaccional será $T = \{\tau_1, \ldots, \tau_m\}$, donde la transacción $\tau_i$ está relacionada con el caso csQCA $\mathbf{c}_i = (\mathbf{x_i}, y_i)$ de la siguiente manera: $\tau_i = \{\phi_{i1}, \phi_{i2}, \ldots, \phi_{in}, \eta_i\}$, siendo

$$\phi_{ij} = \begin{cases} V_j & \text{si } x_{ij} = 1 \\ v_j & \text{si } x_{ij} = 0, \end{cases} \qquad \eta_i = \begin{cases} R & \text{si } y_i = 1 \\ r & \text{si } y_i = 0, \end{cases}$$

  Así, cada transacción $\tau_i$ puede considerarse como una colección de $n$ elementos que son las variables $V_j$ o sus negaciones $v_j$, junto con la presencia del resultado $R$, o su ausencia $r$.

*Observación* 3.2. Es importante notar que, aunque hay $2n + 2$ ítems en $I$, las condiciones impuestas en el conjunto de transacciones $T$ limitan enormemente el rango de posibilidades: primero, todas las transacciones tendrán longitud $n + 1$; segundo, ninguna transacción puede contener una variable y su negación ($V_j$ y $v_j$, o $R$ y $r$).

Ahora necesitamos establecer la relación entre las $k$-configuraciones en el modelo csQCA y las reglas de asociación en el modelo AR.

Sea $S_k = (\mathcal{V}_k, \mu)$ una $k$-configuración conducente al resultado $R$. Asumamos que el conjunto $\mathcal{V}_k = \{V_{j1}, \ldots, V_{jk}\}$, entonces $S_k$ es equivalente a la regla de asociación $X \Rightarrow R$ donde

$$X = \{\psi_1, \ldots, \psi_k\},$$

es

$$\psi_l = \begin{cases} V_{jl} & \text{si } \mu(V_{jl}) = 1 \\ v_{jl} & \text{si } \mu(V_{jl}) = 0. \end{cases}$$

*Observación* 3.3. Denotaremos por $\mathcal{T}$ esta correspondencia entre las $k$-configuraciones csQCA y las reglas de asociación, es decir,

$$\mathcal{T}(S_k) = X \Rightarrow R.$$

Similarmente, abusando de la notación, hablaremos de $\mathcal{T}(\mathcal{C})$ como la transformación elemento por elemento de las configuraciones que definen $\mathcal{C}$ (ver Definición 3.7)

Con esta relación entre $k$-configuraciones y reglas de asociación, es obvio que hemos establecido una correspondencia biunívoca entre los casos descritos en la matriz de datos $M$ y el conjunto de transacciones $T$. Además, verificar que si el caso $\mathbf{c}_i$ es compatible con la configuración $S_k$ entonces $X \subset \tau_i$ es directo. Es decir, el antecedente de la regla de asociación asociada con $S_k$ está incluido en la transacción $\tau_i$ asociada con el caso $\mathbf{c}_i$.

Sin embargo, nótese que en un problema csQCA, es suficiente que exista un caso en $M$ compatible con una $k$-configuración dada para que esta se incorpore al conjunto solución. Aún así, en un problema de reglas de asociación, solo las reglas interesantes se consideran en el conjunto solución, y estas dependen de los puntajes de soporte y confianza. Por lo tanto, necesitamos comparar las métricas en QCA y ARM.

De las definiciones de consistencia y cobertura (ecuaciones (3.3) y (3.5)), y la definición anterior de los conjuntos de ítems y transacciones, es claro que

$$n_x = |\{\tau \in T \mid X \subseteq \tau\}|, \qquad n_y = |\{\tau \in T \mid Y \subseteq \tau\}|, \qquad n_{xy} = |\{\tau \in T \mid X \cup Y \subseteq \tau\}|,$$

y por lo tanto, dada una $k$-configuración $S_k$ y su regla de asociación equivalente $X \Rightarrow R$, primero vemos que la cantidad $n_{xy}$ está directamente relacionada con el soporte de la regla $X \Rightarrow R$. En particular,

$$supp(X \Rightarrow R) = \frac{n_{xy}}{m}, \tag{3.7}$$

donde $m$ denota el número de filas (casos) en la matriz de datos muestrales $M$. Además, existe una relación directa entre la consistencia de la configuración y la confianza de la regla equivalente, es decir,

$$cons(S_k) = conf(X \Rightarrow R) \tag{3.8}$$

Por otro lado, la consistencia puede considerarse como la confianza de la condición necesaria para el resultado, es decir,

$$cov(S_k) = conf(R \Rightarrow X) \tag{3.9}$$

Con estas consideraciones en mente, podemos enunciar el siguiente teorema que relaciona las soluciones de un problema csQCA con la minería de un cierto tipo de reglas de asociación.

*Teorema* 3.1 (Teorema de equivalencia para csQCA). Sea $(\mathcal{V}, R, M)$ un modelo csQCA y sea $(I, T)$ el problema de minería de reglas de asociación definido anteriormente en esta sección. Entonces

$$\mathcal{T}(\mathcal{C}) \subseteq \mathcal{AR},$$

donde $\mathcal{AR}$ denota el conjunto de todas las reglas interesantes con umbrales mínimos de soporte y confianza dados por

$$\text{minsupp} = \frac{1}{m} \qquad \text{minconf} = 1.$$

*Demostración.* Sea $\mathcal{S}_k \in \mathcal{C}$ y $\mathcal{T}(\mathcal{S}_k) = X \Rightarrow R$ su regla de asociación equivalente. Como la configuración $\mathcal{S}_k = (\mathcal{V}, \mu)$ es conducente al resultado $R$, existe al menos un caso $\mathbf{c}_i = (\mathbf{x}_i, y_i) \in M$ tal que $y_i = 1$, y $\mu(V_j) = x_{ij}$ para todo $V_j \in \mathcal{V}$. Por lo tanto, $n_{xy} \geq 1$ y entonces, de la ecuación (3.7)

$$\text{supp}(X \Rightarrow R) = \frac{n_{xy}}{m} \geq \frac{1}{m}.$$

Por otro lado, si el modelo csQCA $(\mathcal{V}, R, M)$ es consistente, entonces $\text{cons}(\mathcal{S}_k) = 1$, así que de (3.8) $\text{conf}(X \Rightarrow R) = 1$.

Por lo tanto, la regla de asociación $X \Rightarrow R$ es una regla de asociación interesante con los umbrales

$$\text{mín supp} = \frac{1}{m} \quad \text{y} \quad \text{mín conf} = 1,$$

así que $(X \Rightarrow R) \in \mathcal{AR}$ y entonces $\mathcal{T}(\mathcal{C}) \subseteq \mathcal{AR}$. □

*Observación* 3.4. Nótese que de los dos umbrales establecidos en el teorema, el primero se fija en $1/m$ para que sea suficiente que exista un caso para que se incluya en el conjunto solución del problema csQCA, mientras que el segundo es 1 porque se asume que la matriz $M$ es consistente y por lo tanto no hay contradicciones.

## 3.4 REGLAS DE ASOCIACIÓN REDUCIBLES

Uno de los principales problemas con las reglas de asociación es el número de soluciones que genera. Khedr et al. (2012) afirman que muchas soluciones generadas por los algoritmos de reglas de asociación pueden ser redundantes e irrelevantes. Por lo tanto, reducir el número de reglas es uno de los objetivos principales de la investigación en extracción de reglas de asociación. Con esto en mente, el objetivo principal de esta sección es diseñar un algoritmo de minimización de reglas de asociación que posteriormente se comparará con las soluciones producidas por los algoritmos QCA.

En el caso particular de reglas de asociación provenientes de un problema QCA, como las definidas en la sección anterior, demostraremos a continuación un resultado que nos permitirá simplificar reglas de asociación de manera simple y eficiente, dando los mismos resultados que con la reducción Quine-McCluskey vista para csQCA.

*Teorema* 3.2. Sean $X_1 \Rightarrow R$ y $X_2 \Rightarrow R$ dos reglas de asociación tales que sus $k$-configuraciones correspondientes, $S_k^{(1)}$ y $S_k^{(2)}$, son reducibles a la $(k-1)$-configuración $S_{k-1}^{(*)}$, cuya regla de asociación correspondiente se denotará como $X^* \Rightarrow R$ (es decir, $\mathcal{T}\left(S_{k-1}^{(*)}\right) = X^* \Rightarrow R$). Entonces, si tanto $X_1 \Rightarrow R$ como $X_2 \Rightarrow R$ son reglas interesantes, entonces $X^* \Rightarrow R$ también es una regla interesante.

*Demostración.* Si $X_1 \Rightarrow R$ y $X_2 \Rightarrow R$ son reglas interesantes, entonces ambas reglas cumplen con los criterios de soporte mínimo y confianza mínima:

$$\text{supp}(X_i \Rightarrow R) \geq \text{mín supp}, \qquad \text{conf}(X_i \Rightarrow R) \geq \text{mín conf}, \qquad i = 1, 2. \tag{3.10}$$

Para probar que $X^* \Rightarrow R$ también es una regla interesante, necesitamos mostrar que esta regla también verifica los criterios de soporte mínimo y confianza.

Como $S_k^{(1)} = (\mathcal{V}_k^{(1)}, \mu_1)$ y $S_k^{(2)} = (\mathcal{V}_k^{(2)}, \mu_2)$ son reducibles a la $(k-1)$-configuración $S_{k-1}^{(*)} = (\mathcal{V}^{(*)}, \mu_*)$, según la Definición 3.6, podemos considerar, sin pérdida de generalidad, que $\mathcal{V}_k^{(1)} = \mathcal{V}_k^{(2)} = \{V_1, \ldots, V_k\}$, y existe una variable, que podemos asumir es $V_k$, tal que

$$\mathcal{V}^{(*)} = \{V_1, \ldots, V_{k-1}\},$$

y

$$\mu_*(V_i) = \mu_1(V_i) = \mu_2(V_i), \quad i = 1, \ldots, k-1.$$

Además, como la variable $V_k$ es la única en la que difieren las configuraciones $S_k^{(1)}$ y $S_k^{(2)}$, podemos asumir que, por ejemplo,

$$\mu_1(V_k) = 1, \quad \mu_2(V_k) = 0.$$

Tras estas consideraciones, las reglas de asociación relacionadas con estas configuraciones: $X_1 \Rightarrow R$, $X_2 \Rightarrow R$, y $X^* \Rightarrow R$ están dadas por

$$X_1 = \{\psi_1, \ldots, \psi_{k-1}, V_k\}, \quad X_2 = \{\psi_1, \ldots, \psi_{k-1}, v_k\}, \quad X^* = \{\psi_1, \ldots, \psi_{k-1}\},$$

donde,

$$\psi_i = \begin{cases} V_i & \text{si } \mu_1(V_i) = 1 \\ v_i & \text{si } \mu_1(V_i) = 0, \end{cases}$$

para $i = 1, \ldots k-1$. Así, $X_1 = X^* \cup \{V_k\}$ y $X_2 = X^* \cup \{v_k\}$.

Ahora mostraremos que cualquier transacción que contenga $X^*$ contiene $X_1$ o $X_2$, es decir,

$$\{\tau \in T : X^* \subset \tau\} = \{\tau \in T : X_1 \subset \tau\} \cup \{\tau \in T : X_2 \subset \tau\} = T_1 \cup T_2.$$

Para este propósito, usaremos el método de doble inclusión.

$\subseteq$ Sea $\tilde{\tau} \in \{\tau \in T : X^* \subset \tau\}$, entonces de la Observación 3.2 sabemos que todas las transacciones tienen longitud $n+1$, y que contienen cada variable una vez, ya sea en su forma afirmativa ($V_i$) o negativa ($v_i$). Por lo tanto, como $X^* \subset \tilde{\tau}$, podemos asegurar que

$$\tilde{\tau} = \{\psi_1, \ldots, \psi_{k-1}, V_k, \ldots\}, \quad \text{o} \quad \tilde{\tau} = \{\psi_1, \ldots, \psi_{k-1}, v_k, \ldots\}.$$

En cualquier caso $\tilde{\tau} \in T_1 \cup T_2$.

$\supseteq$ Sea $\tilde{\tau} \in T_1 \cup T_2$. Entonces, si $\tilde{\tau} \in T_1$ entonces $\tilde{\tau} = \{\psi_1, \ldots, \psi_{k-1}, V_k, \ldots\}$, y si $\tilde{\tau} \in T_2$, entonces $\tilde{\tau} = \{\psi_1, \ldots, \psi_{k-1}, v_k, \ldots\}$. En cualquier caso, $X^* = \{\psi_1, \ldots, \psi_{k-1}\} \subset \tilde{\tau}$, y por lo tanto $\tilde{\tau} \in \{\tau \in T : X^* \subset \tau\}$.

Además, es trivial que $T_1$ y $T_2$ son conjuntos disjuntos, ya que si una transacción contiene $X_1$, entonces contiene la variable afirmativa $V_K$, y por lo tanto no puede contener $X_2$. Por lo tanto,

$$n_{x^*} = |\{\tau \in T : X^* \subset \tau\}| = |T_1 \cup T_2| = |T_1| + |T_2| = n_{x_1} + n_{x_2}.$$

Análogamente, también puede mostrarse que

$$n_{x^*y} = n_{x_1y} + n_{x_2y}.$$

Por lo tanto,

$$\text{supp}(X^* \Rightarrow R) = \frac{n_{x^*}}{m} = \frac{n_{x_1} + n_{x_2}}{m} = \text{supp}(X_1 \Rightarrow R) + \text{supp}(X_2 \Rightarrow R), \tag{3.11}$$

y

$$\text{conf}(X^* \Rightarrow R) = \frac{n_{x^*y}}{n_{x^*}} = \frac{n_{x_1y} + n_{x_2y}}{n_{x_1} + n_{x_2}}. \tag{3.12}$$

Ahora estamos en posición de probar que $X^* \Rightarrow R$ es una regla interesante. De las ecuaciones (3.10) y (3.11) es obvio que

$$\text{supp}(X^* \Rightarrow R) \geq 2\,\text{mín}\,\text{supp} > \text{mín}\,\text{supp}.$$

Por otro lado, teniendo en cuenta la propiedad algebraica que establece que si $a, b, c, d > 0$ y $a/b \leq c/d$ entonces

$$\frac{a}{b} \leq \frac{a+c}{b+d} \leq \frac{c}{d},$$

y si asumimos, sin pérdida de generalidad, que $n_{x_1y}/n_{x_1} \leq n_{x_2y}/n_{x_2}$, entonces de (3.12),

$$\frac{n_{x_1y}}{n_{x_1}} \leq \frac{n_{x_1y} + n_{x_2y}}{n_{x_1} + n_{x_2}} \leq \frac{n_{x_2y}}{n_{x_2}}.$$

Así,

$$\text{conf}(X_1 \Rightarrow R) \leq \text{conf}(X^* \Rightarrow R) \leq \text{conf}(X_2 \Rightarrow R),$$

y, de (3.10) podemos concluir que

$$\text{conf}(X^* \Rightarrow R) \geq \text{conf}(X_1 \Rightarrow R) \geq \text{mín}\,\text{conf},$$

y, por lo tanto, $X^* \Rightarrow R$ es una regla interesante. □

Este resultado nos permite obtener un método para minimizar reglas de asociación en el sentido de Quine-McCluskey, pero con la ventaja de no necesitar usar tablas de verdad cuyo tamaño crece exponencialmente con el número de variables. El método Apriori proporciona una lista de reglas de asociación interesantes que sabemos contienen todas las soluciones del problema csQCA. Ahora, como sabemos por el Teorema 3.2, si dos reglas

interesantes son reducibles, su reducción también es interesante, y por lo tanto ya estará en la lista. Por lo tanto, solo tenemos que eliminar de la lista aquellas reglas que son reducibles.

## 3.5 EJEMPLOS

### 3.5.1 *Bloqueos de Internet en períodos electorales*

En este primer ejemplo, se ilustran las metodologías csQCA y ARM utilizando un estudio de caso derivado de Freyburg y Garbe (2018), que analiza la presencia o ausencia de apagones de Internet durante períodos electorales, incluyendo los días previos a las elecciones, el día de las elecciones y los días posteriores. Los datos utilizados, presentados en la Tabla 3.1, se obtuvieron de un estudio comparativo que abarca 33 países de la región subsahariana (tratados como casos/transacciones). Con fines educativos, este trabajo explora tres condiciones causales (ISP, AT, EV)[1]3.1 junto con un resultado (IS)[2]Freyburg y Garbe (2018).

#### 3.5.1.1 *Enfoque QCA*

Siguiendo la notación introducida en la Subsección 3.1, el modelo csQCA está dado por $(\mathcal{V}, R, M)$, donde:

- $\mathcal{V} = \{\text{ISP}, \text{AT}, \text{EV}\}$ es el conjunto de variables. Más precisamente: ISP se refiere a "Proveedores de Servicios de Internet de propiedad estatal". Estos ISP juegan un papel crucial, especialmente en contextos donde el gobierno tiene control directo sobre la infraestructura de telecomunicaciones, como la red troncal de Internet nacional, permitiéndole ejercer una influencia significativa sobre el acceso y disponibilidad de Internet. AT es la variable "Autocracia". Se utiliza para categorizar el régimen político de un país, distinguiendo entre "1" si el régimen es autocrático y "0" si es una democracia. EV se refiere a la ocurrencia de violencia asociada al proceso electoral.
- $R = \text{IS}$. La variable resultado es "Apagón de Internet".
- $M$ es la matriz de datos dada en la Tabla 3.1.

La Tabla 3.2 presenta la tabla de verdad para la ocurrencia de apagones de Internet. En el presente estudio, con 3 condiciones causales (variables) consideradas, hay $2^3 = 8$ configuraciones teóricamente posibles entre los 33 casos examinados. Si bien todas estas 8 configuraciones aparecen en los datos empíricos, lograr tal completitud empírica no es típico. Cuando ciertas configuraciones carecen de representación empírica, resulta en diversidad limitada, una restricción frecuentemente encontrada en estudios comparativos de pequeña y mediana escala.

Ahora realizaremos el llamado análisis de suficiencia, cuyo objetivo central es evaluar todas las combinaciones posibles de condiciones causales e identificar cuáles de estas combinaciones constituyen subconjuntos consistentes del resultado (Rihoux y Ragin, 2009). En esta situación, la configuración se considera suficiente para el resultado (IS) (Oana et al.,

Tabla 3.1: Resumen de casos y variables

| N. | Caso (País–Año electoral) | ISP Propiedad estatal | Autocracia | Violencia electoral | Apagón de Internet |
|---|---|---|---|---|---|
| 1 | Benín 2015 | 1 | 0 | 0 | 0 |
| 2 | Benín 2016 | 1 | 0 | 0 | 0 |
| 3 | Botsuana 2014 | 1 | 0 | 0 | 0 |
| 4 | Burkina Faso 2015 | 0 | 0 | 0 | 0 |
| 5 | Burundi 2015 | 1 | 1 | 1 | 1 |
| 6 | RCA 2015 | 0 | 1 | 0 | 0 |
| 7 | RCA 2016 | 0 | 0 | 0 | 0 |
| 8 | Chad 2016 | 1 | 1 | 1 | 1 |
| 9 | Yibuti 2016 | 0 | 1 | 1 | 1 |
| 10 | Guinea Ecuatorial 2016 | 1 | 0 | 0 | 0 |
| 11 | Etiopía 2015 | 1 | 1 | 0 | 1 |
| 12 | Gabón 2016 | 1 | 1 | 0 | 1 |
| 13 | Gambia 2016 | 0 | 0 | 1 | 1 |
| 14 | Ghana 2016 | 1 | 1 | 0 | 1 |
| 15 | Guinea 2015 | 0 | 0 | 0 | 0 |
| 16 | Guinea-Bisáu 2014 | 1 | 0 | 0 | 0 |
| 17 | Costa de Marfil 2015 | 0 | 0 | 0 | 0 |
| 18 | Costa de Marfil 2016 | 0 | 0 | 0 | 0 |
| 19 | Lesoto 2015 | 1 | 0 | 0 | 0 |
| 20 | Malaui 2014 | 0 | 0 | 0 | 0 |
| 21 | Mauritania 2014 | 0 | 0 | 0 | 0 |
| 22 | Mozambique 2014 | 0 | 1 | 0 | 0 |
| 23 | Namibia 2014 | 1 | 0 | 1 | 0 |
| 24 | Níger 2016 | 1 | 0 | 0 | 0 |
| 25 | Nigeria 2015 | 1 | 0 | 0 | 0 |
| 26 | República del Congo 2016 | 1 | 0 | 0 | 0 |
| 27 | Sudáfrica | 1 | 0 | 0 | 0 |
| 28 | Sudán del Norte 2015 | 0 | 1 | 1 | 1 |
| 29 | Tanzania 2015 | 1 | 0 | 1 | 0 |
| 30 | Togo 2015 | 1 | 1 | 0 | 1 |
| 31 | Uganda 2016 | 0 | 1 | 1 | 1 |
| 32 | Zambia 2015 | 1 | 0 | 0 | 0 |
| 33 | Zambia 2016 | 1 | 0 | 1 | 0 |

2021). Analizando la Tabla 3.2, podemos observar que hay cuatro 4-configuraciones conducentes al resultado IS. A saber:

$$S_3^{(1)} = \begin{bmatrix} \text{ISP} & \text{AT} & \text{EV} \\ 0 & 0 & 1 \end{bmatrix} \quad S_3^{(2)} = \begin{bmatrix} \text{ISP} & \text{AT} & \text{EV} \\ 0 & 1 & 1 \end{bmatrix}$$

$$S_3^{(3)} = \begin{bmatrix} \text{ISP} & \text{AT} & \text{EV} \\ 1 & 1 & 0 \end{bmatrix} \quad S_3^{(4)} = \begin{bmatrix} \text{ISP} & \text{AT} & \text{EV} \\ 1 & 1 & 1 \end{bmatrix}$$

Tabla 3.2: Tabla de verdad, resultado 'Apagón de Internet'

| ISP | AT | EV | IS | n | Cons.[3] | Casos |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 1 | 1 | Gambia_16 |
| 0 | 1 | 1 | 1 | 3 | 1 | Yibuti_16, Sudán del Norte_15, Uganda_16 |
| 1 | 1 | 0 | 1 | 4 | 1 | Gabón_16, Ghana_16, Etiopía_15, Togo_15 |
| 1 | 1 | 1 | 1 | 2 | 1 | Burundi_15, Chad_16 |
| 0 | 0 | 0 | 0 | 7 | 0 | Burkina Faso_15, RCA_15, Guinea_15, Costa de Marfil_16, Malaui_14, Mauritania_14 |
| 0 | 1 | 0 | 0 | 2 | 0 | RCA_15, Mozambique_14 |
| 1 | 0 | 0 | 0 | 11 | 0 | República del Congo_16, Benín_15, Benín_16, Botsuana_14, Lesoto_15, Guinea-Bisáu_16, Costa de Marfil_16, Níger_16, Nigeria_15, Guinea Ecuatorial_16, Zambia_15, Sudáfrica_14 |
| 1 | 0 | 1 | 0 | 3 | 0 | Namibia_14, Tanzania_15, Zambia_16 |

En este ejemplo de tamaño reducido, es fácil ver que todas las 3-configuraciones son reducibles a tres 2-configuraciones. En particular:

$$\left.\begin{matrix} S_3^{(1)} \\ S_3^{(2)} \end{matrix}\right\} \to S_2^{(1)} = \begin{bmatrix} \text{ISP} & \text{EV} \\ 0 & 1 \end{bmatrix}, \qquad \left.\begin{matrix} S_3^{(3)} \\ S_3^{(4)} \end{matrix}\right\} \to S_2^{(2)} = \begin{bmatrix} \text{ISP} & \text{AT} \\ 1 & 1 \end{bmatrix}.$$

$$\left.\begin{matrix} S_3^{(2)} \\ S_3^{(4)} \end{matrix}\right\} \to S_2^{(3)} = \begin{bmatrix} \text{AT} & \text{EV} \\ 1 & 1 \end{bmatrix}$$

Dado que las 2-configuraciones obtenidas ya no son reducibles, el conjunto solución del problema QCA es:

$$\mathcal{C} = \mathcal{C}_2 = \left\{ S_2^{(1)}, S_2^{(2)}, S_2^{(3)} \right\}.$$

Para interpretar la solución obtenida, seguiremos la notación estándar en QCA, en la cual las variables se escriben en mayúsculas y minúsculas para indicar la presencia y ausencia de una condición causal, respectivamente. El símbolo '*' se usa para agregar variables, y el signo '+' denota la unión de configuraciones. Así, con esta notación, las 2-configuraciones obtenidas pueden escribirse como:

$$S_2^{(1)} = isp * EV, \qquad S_2^{(2)} = ISP * AT, \qquad S_2^{(3)} = AT * EV,$$

y por lo tanto, la minimización de la tabla de verdad para el corte de Internet conduce a la siguiente solución:

3 Consistencia

$$\mathcal{C} : isp * EV + ISP * AT + AT * EV \rightarrow IS.$$

En resumen, se han identificado tres soluciones para los cortes de Internet durante elecciones recientes en países del África subsahariana (SSA). En el primer escenario, la propiedad del ISP no es pública y ocurrieron incidentes significativos de violencia electoral. En el segundo escenario, un estado autocrático que posee la mayoría de los ISP inicia una interrupción del servicio de Internet durante las elecciones. Finalmente, la combinación de un régimen autoritario con episodios de violencia electoral también condujo a interrupciones en el acceso a Internet, independientemente de si la propiedad del ISP es pública o privada.

En cuanto a las métricas, cabe destacar que, como se muestra en la Tabla 3.3, las 3 configuraciones solución tienen consistencia 1, ya que no existen casos contradictorios. La cobertura de los casos es de 0.4, 0.6 y 0.5, lo que muestra que, cuando ocurre una interrupción del servicio de Internet, la segunda configuración causal ($ISP * AT$) es la más frecuente (ejemplos de esto pueden observarse en eventos de 2015 (Etiopía, Togo y Burundi) y 2016 (Gabón, Ghana y Chad).

Tabla 3.3: Tabla de Soluciones con Consistencia y Cobertura

| **Solución** | $X_i$ | $Y_i$ | $n_x$ | $n_y$ | $n_{xy}$ | **Consistencia** | **Cobertura** |
|---|---|---|---|---|---|---|---|
| $isp * EV \rightarrow IS$ | isp * EV | $IS$ | 4 | 10 | 4 | 1 | 0.4 |
| $ISP * AT \rightarrow IS$ | ISP * AT | $IS$ | 6 | 10 | 6 | 1 | 0.6 |
| $AT * EV \rightarrow IS$ | $AT * EV$ | $IS$ | 5 | 10 | 5 | 1 | 0.5 |

#### 3.5.1.2 *Enfoque AR*

Desde el punto de vista de la minería de reglas de asociación, para este problema, definimos el conjunto de ítems como:

$$I = \{ISP, isp, AT, at, EV, ev, IS, is\}, \tag{3.13}$$

y el conjunto de transacciones $T$ estaría definido por las filas de la Tabla 3.1. Por ejemplo, la fila 1, correspondiente al proceso electoral de 2015 en Benín, sería la transacción $\tau = \{ISP, at, ev, is\}$.

El algoritmo Apriori para la minería de reglas de asociación booleanas positivas y negativas sigue cuatro pasos principales (Cornelis et al., 2006):

a) Generar conjuntos de ítems candidatos como subconjuntos del conjunto universal de ítems $I$ dado en (3.13): Comenzar con ítems individuales (1-itemsets), luego combinar iterativamente los conjuntos de ítems frecuentes del paso anterior para formar conjuntos de ítems candidatos más grandes ($k$-itemsets).

b) Poda utilizando el Soporte Mínimo: Eliminar candidatos que no cumplan con un umbral mínimo de soporte definido por el usuario (frecuencia en el conjunto de datos). Esto se basa en el **principio de Apriori**: cualquier subconjunto de un conjunto frecuente también debe ser frecuente.

c) Repetir: Continuar generando y podando conjuntos de ítems de tamaño creciente hasta que no se encuentren nuevos conjuntos frecuentes.

d) Formar Reglas de Asociación: A partir de los conjuntos de ítems frecuentes finales, generar reglas (por ejemplo, X $\rightarrow$ Y) y filtrarlas utilizando un umbral mínimo de confianza (frecuencia con la que Y aparece cuando X está presente).

Si $C_k$ denota el conjunto candidato de longitud $k$, entonces se puede demostrar que, si hay $n$ variables y un resultado, el número de ítems en $C_k$, considerando que $C_k$ no puede contener una variable y su negación, se da por:

$$|C_k| = \frac{\prod_{i=0}^{k-1} 2(n+1-i)}{k!}, \quad k = 1, 2, \ldots, n+1.$$

Por lo tanto, en este ejemplo, $|C_1| = 8$, $|C_2| = 24$, $|C_3| = 32$ y $|C_4| = 16$.

Como se mencionó anteriormente, Apriori es un algoritmo ascendente que comienza con los conjuntos más pequeños y aumenta de tamaño, eliminando combinaciones que no superan el umbral de soporte mínimo. En nuestro caso, y siguiendo el Teorema 3.1, el umbral de soporte mínimo es:

$$\text{mín supp} = \frac{1}{m} = \frac{1}{33} \approx 0.03.$$

Es importante destacar que la minería de reglas de asociación no considera inicialmente algunos ítems como antecedentes y otros como consecuentes, como lo hace QCA al separar las variables antecedentes y el resultado a analizar. Por consiguiente, el paso preliminar consiste en extraer los conjuntos de ítems que contienen la variable de respuesta $'IS'$ y analizar cuáles de estos conjuntos cumplen las condiciones mínimas de soporte y confianza.

La Tabla 3.4 muestra los conjuntos de ítems que han superado el umbral mínimo de soporte, la regla de asociación que definen y la confianza de la regla. Las reglas de asociación que tienen confianza 1 están resaltadas en negrita.

Tabla 3.4: Reglas de asociación que conducen al resultado "IS" y que cumplen la condición de soporte mínimo

| **Itemset** | **Frecuencia** | **Regla de Asociación** | **Frecuencia Ant.** | **Soporte** | **Confianza** |
|---|---|---|---|---|---|
| {"isp","IS"} | 4 | isp → IS | 13 | 0.12 | 0.31 |
| {"at","IS"} | 1 | at → IS | 22 | 0.03 | 0.05 |
| {"ev","IS"} | 4 | ev → IS | 24 | 0.12 | 0.17 |
| {"ISP","IS"} | 6 | ISP → IS | 20 | 0.18 | 0.30 |
| {"AT","IS"} | 9 | AT → IS | 11 | 0.27 | 0.82 |
| {"EV","IS"} | 6 | EV → IS | 9 | 0.18 | 0.67 |
| **{"isp","EV","IS"}** | **4** | **isp * EV → IS** | **4** | **0.12** | **1.00** |
| {"ISP","EV","IS"} | 2 | ISP * EV → IS | 5 | 0.06 | 0.40 |
| {"ISP","ev","IS"} | 4 | ISP * EV → IS | 15 | 0.12 | 0.27 |
| **{"ISP","AT", "IS"}** | **6** | **ISP * AT → IS** | **6** | **0.18** | **1.00** |
| {"isp","AT","IS"} | 3 | isp * AT → IS | 5 | 0.09 | 0.60 |
| {"AT","ev","IS"} | 4 | AT * ev → IS | 6 | 0.12 | 0.67 |
| **{"AT","EV", "IS"}** | **5** | **AT * EV → IS** | **5** | **0.15** | **1.00** |
| {"at", "EV","IS"} | 1 | AT * ev → IS | 4 | 0.03 | 0.25 |
| {"isp", "at","IS"} | 1 | AT * ev → IS | 8 | 0.03 | 0.13 |
| **{"isp","at","EV","IS"}** | **1** | **ISP * AT * ev → IS** | **1** | **0.03** | **1.00** |
| **{"isp","AT","EV","IS"}** | **3** | **ISP * AT * ev → IS** | **3** | **0.09** | **1.00** |
| **{"ISP","AT","ev","IS"}** | **4** | **ISP * AT * ev → IS** | **4** | **0.12** | **1.00** |
| **{"ISP","AT","EV","IS"}** | **2** | **ISP * AT * EV → IS** | **2** | **0.06** | **1.00** |

Dado los resultados mostrados en la Tabla 3.4, el conjunto $\mathcal{AR}$ de reglas de asociación interesantes que conducen al resultado "Apagón de Internet" estaría formado por las reglas:

$$
\begin{aligned}
r_1 &: isp * EV \rightarrow IS \\
r_2 &: ISP * AT \rightarrow IS \\
r_3 &: AT * EV \rightarrow IS \\
r_4 &: isp * at * EV \rightarrow IS \\
r_5 &: isp * AT * EV \rightarrow IS \\
r_6 &: ISP * AT * ev \rightarrow IS \\
r_7 &: ISP * AT * EV \rightarrow IS.
\end{aligned}
$$

Es evidente que puede establecerse una relación entre las configuraciones causales del problema QCA y las reglas de asociación obtenidas. De hecho, recordando la definición del operador $\mathcal{T}$ en la Observación 3.3,

$$\mathcal{T}(\mathcal{C}_2) = \mathcal{T}\left(\left\{\mathcal{S}_2^{(1)}, \mathcal{S}_2^{(2)}, \mathcal{S}_2^{(3)}\right\}\right) = \{r_1, r_2, r_3\}$$
$$\mathcal{T}(\mathcal{C}_3) = \mathcal{T}\left(\left\{\mathcal{S}_3^{(1)}, \mathcal{S}_3^{(2)}, \mathcal{S}_3^{(3)}, \mathcal{S}_4^{(3)}\right\}\right) = \{r_4, r_5, r_6, r_7\}$$

En consecuencia, el conjunto solución del problema QCA está contenido en AR. Además, debe destacarse que el conjunto AR abarca tanto las configuraciones causales iniciales como las minimizadas. Para obtener el mismo resultado, basta con determinar si las reglas $r_1$, $r_2$ y $r_3$ pueden obtenerse minimizando las reglas $r_4, \ldots, r_7$. Está claro que este es el caso y que, si lleváramos a cabo esta minimización, obtendríamos el mismo conjunto de soluciones.

### 3.5.2 *Oposición a la inmigración en Europa*

En el siguiente estudio de caso, mostraremos cómo la metodología de reglas de asociación también puede utilizarse eficientemente cuando el número de casos es mayor. Para ello, adaptaremos los datos de Emmenegger, Schraff y Walter (2014), para mostrar otra comparación entre QCA y Reglas de Asociación.

Estos datos consisten en una muestra robusta de 2.223 observaciones calibradas de un estudio que investiga la oposición a la inmigración en Europa. Los datos originales se recopilaron utilizando una escala ordinal de 10 puntos, que permitió a los participantes expresar sus niveles de acuerdo y desacuerdo sobre el tema de la inmigración.

Este conjunto de datos cubre un total de seis variables, cinco de las cuales representan posibles factores condicionantes: 1- Baja educación [LOWEDU], 2- Preferencia por la unidad cultural [UNITY], 3- Falta de amigos inmigrantes [NOFRIEND], 4- Amenaza económica percibida y 5- Género [MAN].

Además de estas, hay una variable de resultado llamada [OPPOSITION].

La Tabla 3.5 resume la muestra utilizada en este estudio, destacando las frecuencias correspondientes a cada posible combinación de respuestas de los participantes. Dado que tenemos cinco variables independientes y dos opciones de respuesta (sí o no) para cada una de ellas, existen 32 combinaciones posibles de respuestas, como se muestra en la Tabla 3.5. Por ejemplo, en la primera fila de la tabla, vemos que 22 individuos respondieron negativamente a la primera variable y afirmativamente a las otras cuatro.

#### 3.5.2.1 *Solución csQCA*

En este contexto, presentamos las soluciones obtenidas con el algoritmo csQCA para analizar la oposición a la inmigración en Europa. Estas soluciones comprenden cuatro combinaciones causales que llevan al fenómeno analizado, a saber:

1. LOWEDU*UNITY

2. LOWEDU*NOFRIEND*THREAT
3. LOWEDU*NOFRIEND*MAN
4. UNITY*NOFRIEND*THREAT*MAN

Se observa que la solución más concisa, caracterizada por el menor número de condiciones causales, es la primera, compuesta por solo dos condiciones: baja escolaridad y preferencia por la unidad cultural. Por otro lado, la solución con mayor número de condiciones causales es la cuarta, que comprende cuatro variables causales diferentes.

#### 3.5.2.2 *Solución de Reglas de Asociación Positivas y Negativas*

El análisis de datos en este subapartado se centra en examinar las soluciones derivadas de las reglas de asociación binarias. En este contexto:

a) Los datos se prepararon adecuadamente para ser leídos en el software R. Específicamente, las respuestas individuales que representan niveles de acuerdo (presente) se transformaron en transacciones y se codificaron como 1. Por otro lado, las respuestas que indicaban desacuerdo se codificaron como 0.

b) La siguiente etapa consistió en convertir los objetos en transacciones y aplicar el algoritmo Apriori para extraer las reglas de asociación. Para garantizar la comparabilidad entre los enfoques, se estableció un soporte mínimo de regla de 0.0004. Esto fue esencial debido a la posibilidad de que, dentro del alcance de la metodología QCA, surgieran soluciones con frecuencias extremadamente bajas o incluso ausentes. Además, la confianza mínima se fijó en 1, lo que representa una probabilidad condicional del 100 %.

c) Se inspeccionaron las reglas para permitir su visualización. La Tabla 3.6 enumera el conjunto completo de soluciones obtenidas utilizando el algoritmo de reglas de asociación.

La Tabla 3.6 muestra el conjunto de soluciones de reglas de asociación derivadas del problema propuesto, junto con las métricas correspondientes. Como se puede observar, se generaron un total de 39 reglas, organizadas en orden creciente de complejidad. La regla menos compleja implica combinar solo dos variables, mientras que las reglas más complejas implican combinar cinco variables diferentes.

Sin embargo, gracias al Teorema 3.2, sabemos que si dos reglas de asociación interesantes son reducibles, entonces la regla reducida también es interesante. Esta idea permite un algoritmo de minimización muy simple: podemos eliminar de la lista todas las reglas para las cuales existe otra que difiera de ella en solo un bit. Por ejemplo, las reglas 2 y 7, 3 y 6, y 4 y 5 son todas reducibles a la regla 1. Del mismo modo, todas las reglas de longitud 5 (es decir, reglas 28 a 39) son reducibles a reglas de longitud 4, que ya están presentes en la lista, y así sucesivamente.

Después de filtrar todas las reglas reducibles, las restantes son:

1. Rule 1: LOWEDU*UNITY
2. Rule 8: LOWEDU*MAN*NOFRIEND
3. Rule 9: LOWEDU*NOFRIEND*THREAT

4. Rule 15: MAN*NOFRIEND*THREAT*UNITY

Estas coinciden con la solución obtenida para el problema csQCA en el subapartado anterior.

### 3.5.3 *Experimentos numéricos*

El costo computacional del análisis de tablas de verdad y la minimización de Quine–McCluskey ha sido ampliamente discutido en la literatura de QCA. Está bien establecido que los estudios de QCA típicamente involucran un número limitado de variables (raramente más de diez) y un número relativamente pequeño de casos (a menudo solo unas decenas). Ejemplos que involucran cientos o miles de casos—como el mostrado en el ejemplo anterior—son excepcionalmente raros.

Inicialmente, los métodos de análisis de QCA eran especialmente aplicables a estudios con un número intermedio de casos (entre 10 y 50 (Ragin y Rihoux, 2004)). A lo largo de los años, varios investigadores han desarrollado estudios que manejan miles de casos con diferentes propósitos (Greckhamer et al., 2013; Mendel y Korjani, 2013; Vis, 2012). De hecho, en teoría, el tamaño de la muestra está limitado únicamente por las restricciones del hardware y software, y no necesariamente por limitaciones impuestas por la metodología en sí.

Realizamos una serie de experimentos numéricos para ilustrar las diferencias prácticas entre csQCA y la minería de reglas de asociación, enfocándonos particularmente en la viabilidad computacional. Estos experimentos simulan problemas con diferentes cantidades de variables y casos, lo que permite una comparación sistemática de la escalabilidad y eficiencia de ambos enfoques. Para cada escenario, generamos aleatoriamente tablas de datos consistentes (es decir, sin contradicciones) con un número fijo de variables y casos. Cada experimento se repitió cinco veces, y se registró el tiempo medio de cómputo y el uso de memoria para el algoritmo de csQCA (implementado mediante el paquete QCA de R (Duşa, 2019)), el algoritmo Apriori de minería de reglas (utilizando el paquete `arules` de R (Hahsler, Chelluboina et al., 2011; Hahsler, Gruen y Hornik, 2005)), y una implementación personalizada de un algoritmo de minimización de reglas de asociación en R. Todos los cálculos se realizaron en un procesador Intel(R) Core(TM) i7-1360P de 13ª generación (2.20 GHz) con 16 GB de RAM y sin uso de aceleración por GPU. El código para este experimento está disponible en el repositorio de GitHub https://github.com/rbensua/csQCAARM/.

En el primer experimento, se fijó un número pequeño de variables (5 variables) y se varió el número de casos desde 5 hasta 10000. Los resultados, mostrados en la Figura 3.1, indican que csQCA requirió casi mil veces más uso de memoria que Apriori. La minimización de reglas de asociación, por otro lado, requirió solo alrededor de 300 Kb, independientemente del número de casos.

En el segundo experimento, se fijó el número de casos en 60 y se varió el número de variables de 5 a 21. Los resultados, mostrados en la Figura 3.2, revelan que tanto en csQCA como en la minería de reglas de asociación (AR), el uso de memoria y los tiempos de procesamiento aumentan exponencialmente con el número de variables (nótese que

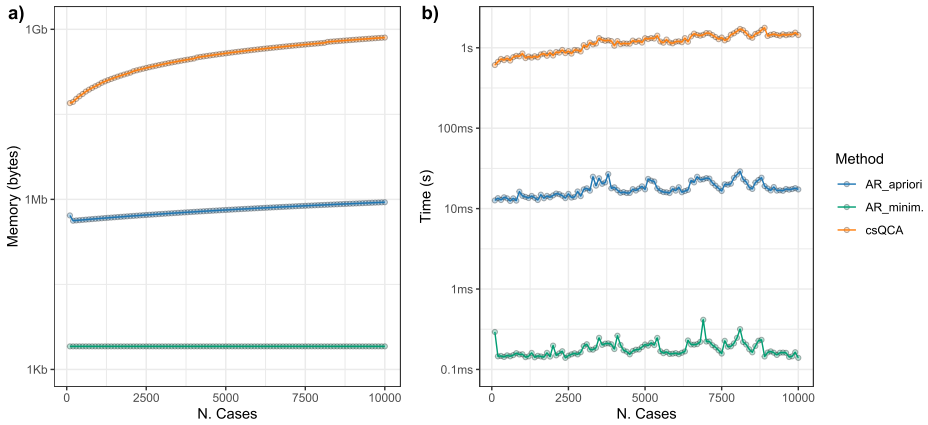


Figura 3.1: Una comparación entre el uso de memoria (izquierda) y el tiempo de cómputo (derecha) de csQCA y la minería de reglas de asociación a medida que aumenta el número de casos en el conjunto de datos. Ambas escalas son logarítmicas. El número de variables se fijó en 5.

ambas escalas son logarítmicas). Sin embargo, dentro del rango típico de un problema de csQCA (5–15 variables), el uso de memoria de ARM—tanto en la generación de reglas mediante el algoritmo Apriori como en el procedimiento de minimización posterior—es significativamente menor que el de csQCA.

Para un número mayor de variables, el uso de memoria del algoritmo Apriori aumenta rápidamente (lo cual es una limitación bien conocida en la minería de reglas de asociación). En cuanto al tiempo de cómputo, los algoritmos de AR también tuvieron un rendimiento significativamente mejor que csQCA, el cual enfrenta rápidamente severas limitaciones computacionales a medida que aumenta el número de variables (llegando a tener dificultades para encontrar soluciones con más de 20 variables, como también lo señala Chiu y Xu (2023)).

En el segundo experimento, se fijó el número de casos en 60 y se varió el número de variables entre 5 y 21. Los resultados, mostrados en la Figura 3.2, indican que ambos métodos, csQCA y AR, presentan aumentos exponenciales en el uso de memoria y en el tiempo de procesamiento a medida que crece el número de variables (nótese la escala logarítmica en ambos ejes).

Dentro del rango típico de aplicaciones de csQCA (5–15 variables), ambos pasos de AR (la generación de reglas mediante el algoritmo Apriori y el procedimiento de minimización posterior) utilizan significativamente menos memoria que csQCA. Sin embargo, para un número mayor de variables, el uso de memoria de Apriori aumenta rápidamente, lo cual constituye una limitación bien conocida en la minería de reglas de asociación, alcanzando valores del mismo orden de magnitud que csQCA.

En cuanto al tiempo de cómputo, los algoritmos de AR también superan claramente a csQCA. Los tiempos de cómputo de csQCA están en el rango de $[0{,}9\text{ s} - 236\text{ s}]$, mientras que los tiempos combinados de AR se encuentran en el rango de $[0{,}02\text{ s} - 105{,}7\text{ s}]$. En este sentido, csQCA muestra un incremento pronunciado en el tiempo de procesamiento y se vuelve computacionalmente inviable cuando se trabaja con más de 20 variables, lo cual es consistente con los hallazgos de Chiu y Xu (2023).

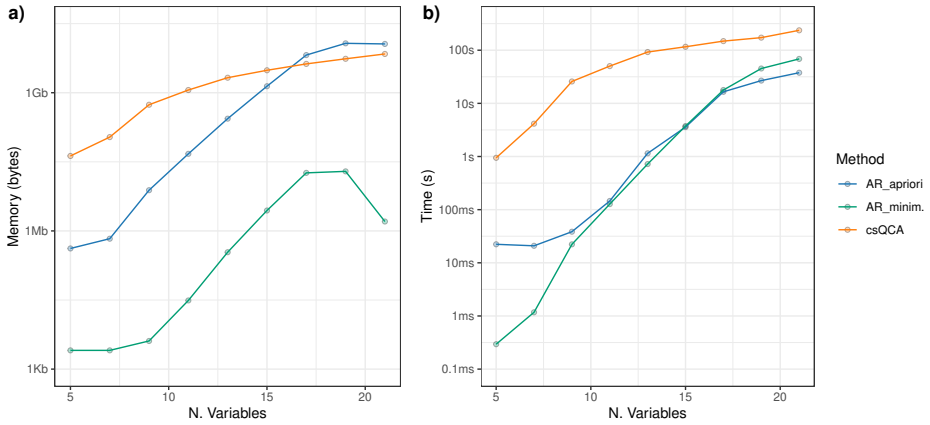


Figura 3.2: Una comparación entre el uso de memoria (izquierda) y el tiempo de cómputo (derecha) de csQCA y la minería de reglas de asociación a medida que aumenta el número de variables (condiciones). El número de casos se fijó en 60.

*Observación* 3.5. Es importante mencionar que, aunque las soluciones obtenidas con csQCA y ARM fueron exactamente las mismas en los ejemplos presentados en las secciones anteriores, en los experimentos numéricos no fue así. Obviamente, según el Teorema 1, las soluciones de csQCA estaban incluidas en la lista de reglas interesantes e irreducibles de ARM, pero ARM encontró más reglas. Esta discrepancia se debe a las diferencias en los algoritmos de minimización utilizados. En csQCA, el algoritmo de Quine–McCluskey sigue un enfoque descendente (top-down), lo que significa que las configuraciones más pequeñas siempre se derivan como simplificaciones de otras más complejas. En contraste, ARM comienza generando todas las reglas potencialmente interesantes y luego elimina aquellas que son reducibles. Como resultado, csQCA no puede producir configuraciones que no se originen a partir de la minimización de reglas más complejas. No incluimos la lista completa de soluciones en los experimentos numéricos porque, en casos de alta dimensionalidad, el número de reglas resultantes alcanzó varios millones.

## 3.6 FUNDAMENTOS TEÓRICOS Y EQUIVALENCIAS ENTRE CSQCA Y ARM

A primera vista, tanto el Análisis Cualitativo Comparativo (QCA) como la minería de reglas de asociación se basan en lógica booleana condicional del tipo "si–entonces" centrada en los casos, y buscan identificar patrones de condiciones o atributos que coocurren o se vinculan con un resultado particular (R. Agrawal et al., 1993; Ragin, 1987). Sin embargo, sus objetivos divergen: QCA se involucra explícitamente con la complejidad causal, permitiendo múltiples caminos interrelacionados hacia un resultado, mientras que las reglas de asociación capturan coocurrencias estadísticas sin implicar causalidad (A. Agrawal y Knoeber, 1996; Ragin, 1987).

El algoritmo de minimización de Quine–McCluskey, originalmente ideado para optimizar circuitos lógicos en ingeniería eléctrica (Kumar et al., 2022), es aplicado por QCA en todas sus variantes. Este algoritmo se utiliza para reducir configuraciones complejas de condiciones en expresiones más simples. En contraste, el proceso de minería de reglas de

asociación comienza con la enumeración de conjuntos frecuentes de ítems, derivando posteriormente reglas basadas en umbrales de soporte y confianza (véase R. Agrawal et al. (1993)).

A pesar de estas diferencias conceptuales y procedimentales, ambas técnicas enfrentan la misma explosión combinatoria: el espacio de búsqueda de todas las combinaciones posibles de condiciones crece exponencialmente con el número de atributos o condiciones distintos (R. Agrawal et al., 1993; Ragin y Sonnett, 2005). Además, existe un paralelismo formal en la manera en que se interpretan las reglas. En minería de reglas de asociación, una implicación $X \rightarrow Y$ se considera fuerte si $X$ satisface un umbral mínimo de frecuencia, y cada transacción que contiene $X$ también incluye $Y$ (R. Agrawal et al., 1993). De manera similar, en QCA, el conjunto de condiciones $X$ se considera necesario para el resultado $Y$ precisamente cuando $Y \subseteq X$ en los casos observados (Ragin y Sonnett, 2005), lo cual puede interpretarse como una formulación alternativa del mismo concepto.

Este estudio proporciona un análisis comparativo detallado de las soluciones obtenidas mediante las metodologías de reglas de asociación y de Análisis Comparativo Configuracional (csQCA).

Considerando la evidencia presentada, emergen los siguientes puntos clave:

1. Puede establecerse una relación uno-a-uno entre un problema de csQCA y un tipo particular de problema de minería de reglas de asociación binarias positivas y negativas (Teorema 3.1). Hasta donde sabemos, esta relación no ha sido explorada previamente.

2. Todas las soluciones generadas por el algoritmo csQCA ya habían sido capturadas previamente por el algoritmo de minería de reglas de asociación. En otras palabras, el conjunto de soluciones emergentes de las reglas de asociación actúa como un superconjunto de las soluciones generadas por el enfoque csQCA. De este modo, las reglas de asociación muestran mayor potencial para generar un espectro más amplio de soluciones en comparación con la metodología csQCA.

3. El procedimiento utilizado para obtener soluciones es completamente diferente. Mientras que csQCA sigue un enfoque de arriba hacia abajo, en el cual se obtienen primero las soluciones más complejas y, a partir de la minimización de Quine–McCluskey, se encuentran soluciones irreducibles más simples, en ARM se emplea un enfoque de abajo hacia arriba (algoritmo Apriori), comenzando con las soluciones más simples y obteniendo soluciones cada vez más complejas.

   No obstante, el algoritmo Apriori no simplifica soluciones reducibles, ya que su objetivo es obtener todas las configuraciones frecuentes. A pesar de ello, mostramos que no es necesario realizar una minimización de Quine–McCluskey ya que, como se muestra en el Teorema 3.2, si dos reglas interesantes son reducibles, su reducción también constituye una regla interesante; por tanto, basta con filtrar y eliminar las reglas reducibles.

4. QCA se originó en el contexto de la investigación en ciencias sociales, donde tradicionalmente los conjuntos de datos son relativamente pequeños para facilitar la

extracción de soluciones interpretables. En cambio, la minería de reglas de asociación surgió de la necesidad de identificar patrones en bases de datos a gran escala. Como resultado, los algoritmos desarrollados para ARM están diseñados específicamente para operar de manera eficiente sobre grandes conjuntos de datos. Dada la relación establecida entre QCA y la minería de reglas de asociación, esta conexión permite utilizar técnicas de ARM para obtener soluciones tipo QCA incluso cuando se trabaja con bases de datos extensas, superando las limitaciones computacionales tradicionales de QCA.

5. Además, es importante destacar que las soluciones obtenidas mediante ambas metodologías presentan métricas de robustez idénticas, con la confianza en minería de reglas de asociación correspondiendo directamente a la consistencia en csQCA. Adicionalmente, todas las soluciones identificadas mantienen una estructura "si–entonces", reforzando su interpretabilidad lógica. Finalmente, ambos enfoques se fundamentan teóricamente en la teoría de conjuntos.

En este estudio mostramos que las reglas de asociación (AR) producen soluciones equivalentes a las de csQCA. En contextos con diversidad de datos limitada, csQCA típicamente produce menos soluciones que AR. Sin embargo, los métodos difieren significativamente. csQCA se enfoca en un único resultado predefinido (Ragin, 2008), mientras que AR puede identificar múltiples resultados y sus complementos (Cornelis et al., 2006), con resultados ya sean predefinidos o derivados de los datos.

Ambos métodos estructuran sus soluciones como afirmaciones del tipo "si–entonces" unidas por operadores disyuntivos (OR) (Duşa, 2019); sin embargo, solo csQCA sugiere causalidad, mientras que AR identifica asociaciones sin pretensión causal (Tan et al., 2014). Los elementos condicionales deben ser seleccionados previamente: condiciones causales en csQCA (Rihoux y Ragin, 2009) e ítems en AR (R. Agrawal et al., 1993). Ambos enfoques se basan en estructuras booleanas y permiten términos negativos. Las combinaciones conjuntivas (AND) expresan causalidad coyuntural en csQCA (Ragin, 2008), mientras que en AR los conjuntos de ítems se conectan de manera similar, pero sin interpretación causal, la cual debe ser inferida por el investigador.

La generación combinatoria también difiere: csQCA genera $2^k$ combinaciones (Ragin y Rihoux, 2004), mientras que AR inicialmente genera $3^k - 2^k + 1$ posibilidades, que normalmente se filtran hasta $2^k - 1$ (Tan et al., 2014).

En este capitulo, exploramos las conexiones teóricas y prácticas entre el csQCA y la minería de reglas de asociación. Utilizando un marco matemático formal, demostramos que todo problema de csQCA puede representarse como un problema específico de ARM que involucra reglas positivas y negativas. Al revelar la mecánica interna de csQCA desde esta perspectiva matemática, abrimos el camino hacia una comprensión más profunda de sus propiedades computacionales y su compatibilidad con los enfoques de ARM.

Mostramos que conceptos clave de QCA, como la consistencia y la cobertura, corresponden de manera natural a métricas de ARM como la confianza y el lift. Esta equivalencia fue ilustrada mediante dos aplicaciones: un caso de tamaño reducido sobre cortes de Internet

en África y un conjunto de datos de gran tamaño sobre actitudes hacia la inmigración en Europa.

Una de las principales ventajas de ARM—particularmente en conjuntos de datos grandes—radica en su eficiencia computacional y en la capacidad para descubrir un conjunto más amplio de reglas relevantes, incluidas aquellas que son pasadas por alto por los algoritmos tradicionales de QCA. Propusimos un algoritmo de minimización para reglas de asociación que imita el enfoque de Quine-McCluskey en csQCA, evitando así la necesidad de construir grandes tablas de verdad.

Estos hallazgos proporcionan una base para que los investigadores seleccionen entre csQCA y ARM en función del tamaño de la muestra, las limitaciones computacionales y los objetivos de investigación. Investigaciones futuras podrían extender esta comparación al QCA de conjuntos difusos (fsQCA) y explorar la integración de métodos basados en AR dentro de diseños de investigación de métodos mixtos.

Tabla 3.5: Summary data for case study 2, opposition to immigration*

| | LOWEDU | UNITY | NOFRIEND | THREAT | MAN | OPPOSITION | n |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 1 | 1 | 22 |
| 2 | 1 | 0 | 1 | 0 | 1 | 1 | 47 |
| 3 | 1 | 0 | 1 | 1 | 0 | 1 | 163 |
| 4 | 1 | 0 | 1 | 1 | 1 | 1 | 64 |
| 5 | 1 | 1 | 0 | 0 | 0 | 1 | 41 |
| 6 | 1 | 1 | 0 | 0 | 1 | 1 | 24 |
| 7 | 1 | 1 | 0 | 1 | 0 | 1 | 98 |
| 8 | 1 | 1 | 0 | 1 | 1 | 1 | 76 |
| 9 | 1 | 1 | 1 | 0 | 0 | 1 | 83 |
| 10 | 1 | 1 | 1 | 0 | 1 | 1 | 40 |
| 11 | 1 | 1 | 1 | 1 | 0 | 1 | 196 |
| 12 | 1 | 1 | 1 | 1 | 1 | 1 | 193 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 93 |
| 14 | 0 | 0 | 0 | 0 | 1 | 0 | 123 |
| 15 | 0 | 0 | 0 | 1 | 0 | 0 | 26 |
| 16 | 0 | 0 | 0 | 1 | 1 | 0 | 40 |
| 17 | 0 | 0 | 1 | 0 | 0 | 0 | 56 |
| 18 | 0 | 0 | 1 | 0 | 1 | 0 | 47 |
| 19 | 0 | 0 | 1 | 1 | 0 | 0 | 19 |
| 20 | 0 | 0 | 1 | 1 | 1 | 0 | 30 |
| 21 | 0 | 1 | 0 | 0 | 0 | 0 | 22 |
| 22 | 0 | 1 | 0 | 0 | 1 | 0 | 37 |
| 23 | 0 | 1 | 0 | 1 | 0 | 0 | 5 |
| 24 | 0 | 1 | 0 | 1 | 1 | 0 | 17 |
| 25 | 0 | 1 | 1 | 0 | 0 | 0 | 21 |
| 26 | 0 | 1 | 1 | 0 | 1 | 0 | 49 |
| 27 | 0 | 1 | 1 | 1 | 0 | 0 | 7 |
| 28 | 1 | 0 | 0 | 0 | 0 | 0 | 101 |
| 29 | 1 | 0 | 0 | 0 | 1 | 0 | 59 |
| 30 | 1 | 0 | 0 | 1 | 0 | 0 | 172 |
| 31 | 1 | 0 | 0 | 1 | 1 | 0 | 156 |
| 32 | 1 | 0 | 1 | 0 | 0 | 0 | 96 |

* Fuente: Adaptado de Emmenegger, Schraff y Walter (2014)

Tabla 3.6: Association rules with 100 % confidence implying {OPPOSITION}.

| | **LHS → OPPOSITION** | **Support** | **Coverage** | **Count** |
|---|---|---|---|---|
| 1 | LOWEDU*UNITY | 0.34 | 0.34 | 751 |
| 2 | LOWEDU*threat*UNITY | 0.08 | 0.08 | 188 |
| 3 | LOWEDU*MAN*UNITY | 0.15 | 0.15 | 333 |
| 4 | LOWEDU*nofriend*UNITY | 0.11 | 0.11 | 239 |
| 5 | LOWEDU*NOFRIEND*UNITY | 0.23 | 0.23 | 512 |
| 6 | LOWEDU*man*UNITY | 0.19 | 0.19 | 418 |
| 7 | LOWEDU*THREAT*UNITY | 0.25 | 0.25 | 563 |
| 8 | LOWEDU*MAN*NOFRIEND | 0.15 | 0.15 | 344 |
| 9 | LOWEDU*NOFRIEND*THREAT | 0.28 | 0.28 | 616 |
| 10 | LOWEDU*MAN*threat*UNITY | 0.03 | 0.03 | 64 |
| 11 | LOWEDU*nofriend*threat*UNITY | 0.03 | 0.03 | 65 |
| 12 | LOWEDU*NOFRIEND*threat*UNITY | 0.06 | 0.06 | 123 |
| 13 | LOWEDU*man*threat*UNITY | 0.06 | 0.06 | 124 |
| 14 | LOWEDU*MAN*nofriend*UNITY | 0.04 | 0.04 | 100 |
| 15 | MAN*NOFRIEND*THREAT*UNITY | 0.10 | 0.10 | 215 |
| 16 | LOWEDU*MAN*NOFRIEND*UNITY | 0.10 | 0.10 | 233 |
| 17 | LOWEDU*MAN*THREAT*UNITY | 0.12 | 0.12 | 269 |
| 18 | LOWEDU*man*nofriend*UNITY | 0.06 | 0.06 | 139 |
| 19 | LOWEDU*nofriend*THREAT*UNITY | 0.08 | 0.08 | 174 |
| 20 | LOWEDU*man*NOFRIEND*UNITY | 0.13 | 0.13 | 279 |
| 21 | LOWEDU*NOFRIEND*THREAT*UNITY | 0.17 | 0.17 | 389 |
| 22 | LOWEDU*man*THREAT*UNITY | 0.13 | 0.13 | 294 |
| 23 | LOWEDU*MAN*NOFRIEND*threat | 0.04 | 0.04 | 87 |
| 24 | LOWEDU*MAN*NOFRIEND*THREAT | 0.12 | 0.12 | 257 |
| 25 | LOWEDU*MAN*NOFRIEND*unity | 0.05 | 0.05 | 111 |
| 26 | LOWEDU*man*NOFRIEND*THREAT | 0.16 | 0.16 | 359 |
| 27 | LOWEDU*NOFRIEND*THREAT*unity | 0.10 | 0.10 | 227 |
| 28 | lowedu*MAN*NOFRIEND*THREAT*UNITY | 0.01 | 0.01 | 22 |
| 29 | LOWEDU*MAN*nofriend*threat*UNITY | 0.01 | 0.01 | 24 |
| 30 | LOWEDU*MAN*NOFRIEND*threat*UNITY | 0.02 | 0.02 | 40 |
| 31 | LOWEDU*man*nofriend*threat*UNITY | 0.02 | 0.02 | 41 |
| 32 | LOWEDU*man*NOFRIEND*threat*UNITY | 0.04 | 0.04 | 83 |
| 33 | LOWEDU*MAN*nofriend*THREAT*UNITY | 0.03 | 0.03 | 76 |
| 34 | LOWEDU*MAN*NOFRIEND*THREAT*UNITY | 0.09 | 0.09 | 193 |
| 35 | LOWEDU*man*nofriend*THREAT*UNITY | 0.04 | 0.04 | 98 |
| 36 | LOWEDU*man*NOFRIEND*THREAT*UNITY | 0.09 | 0.09 | 196 |
| 37 | LOWEDU*MAN*NOFRIEND*threat*unity | 0.02 | 0.02 | 47 |
| 38 | LOWEDU*MAN*NOFRIEND*THREAT*unity | 0.03 | 0.03 | 64 |
| 39 | LOWEDU*man*NOFRIEND*THREAT*unity | 0.07 | 0.07 | 163 |

# 4

# INTEGRACIÓN FORMAL DEL ANÁLISIS CUALITATIVO COMPARATIVO DE CONJUNTOS FUZZY Y LA MINERÍA DE REGLAS DE ASOCIACIÓN CUANTITATIVAS

En este capítulo mostramos que el método fsQCA se basa en la teoría de conjuntos y analiza los casos a partir de su pertenencia a tablas de verdad binarias, construidas con base en las puntuaciones calibradas de conjuntos difusos. Aunque estas puntuaciones representan inicialmente el grado de presencia de una condición, posteriormente son transformados en valores binarios: los puntuaciones superiores a 0.5 se convierten en 1 (presencia), mientras que los inferiores a 0.5 se transforman en 0 (ausencia) en la tabla de verdad. Esta transformación convierte al fsQCA en una metodología orientada a la binarización, en la cual cada condición es interpretada de manera estrictamente dicotómica. Como consecuencia, el método presenta limitaciones en la representación de variables con múltiples niveles lingüísticos, restringiendo el análisis a resultados con solo dos estados posibles por variable.

En contraste, las reglas de asociación aplicadas a variables cuantitativas demuestran una mayor flexibilidad analítica. Cuando estas variables son discretizadas en dos niveles, las soluciones obtenidas se asemejan a las generadas por el fsQCA. Sin embargo, las reglas de asociación permiten el uso de más de dos intervalos, lo que posibilita la formulación de reglas más matizadas y expresivas.

Por ejemplo, al emplear reglas de asociación cuantitativas, es posible construir enunciados como:

> "Si un país tiene un historial razonable de convulsiones populares tras las elecciones, si el nivel de autoritarismo es alto, y si el gobierno ejerce una influencia moderada sobre las operadoras de telefonía móvil, entonces se producirá un corte de internet."

Este tipo de regla permite el uso de términos lingüísticos graduales, como "bajo","medio" y "alto", lo que refleja de forma más fiel la complejidad de los fenómenos sociales. Se trata de una posibilidad que la metodología fsQCA no contempla, ya que sus soluciones operan con lógica booleana estricta, sin incorporar este tipo de categorías lingüísticas intermedias.

Para ilustrar esta comparación, se presenta un ejemplo de reglas de asociación utilizando variables lingüísticas relacionadas con la corrupción.

## 4.1 FORMALIZACIÓN MATEMÁTICA DE FUZZY-SET-QCA

En esta sección presentamos una descripción formal y matemática de la metodología fsQCA. Para ello, seguimos un enfoque análogo al utilizado en la formalización del Capítulo 3. Al igual que en csQCA, partimos de un conjunto de variables, un resultado que se desea analizar, y una colección de casos, en cada uno de los cuales se evalúa el grado en que se verifica cada variable.

Sin embargo, a diferencia de csQCA, en fsQCA las variables no son binarias: no es posible determinar de forma exacta y dicotómica si un caso verifica o no una determinada condición. En su lugar, las variables permiten distintos grados de pertenencia, reflejando la naturaleza difusa de los fenómenos sociales.

Habitualmente, se parte de una base de datos con valores numéricos, a partir de la cual, mediante un proceso de normalización o calibración, se asigna a cada caso un grado de pertenencia a cada variable. Este grado puede interpretarse como el nivel de verificación de la condición representada por dicha variable en ese caso concreto.

**Definición 4.1.** Un modelo fsQCA puede representarse como una **terna** $(\mathcal{V}, R, M)$, donde $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$ es el conjunto de variables antecedentes, $R$ es la variable consecuente o resultado, y $M$ es la matriz de datos muestrales original de dimensión $m \times (n+1)$, donde cada fila representa un caso $c_i = (\mathbf{x}_i, y_i)$.

$$M = \begin{array}{c|cccc|c} & V_1 & V_2 & \cdots & V_n & R \\ \hline \mathbf{c}_1 & x_{11} & x_{12} & \cdots & x_{1n} & y_1 \\ \mathbf{c}_2 & x_{21} & x_{22} & \cdots & x_{2n} & y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{c}_m & x_{m1} & x_{m2} & \cdots & x_{mn} & y_m \end{array}, \tag{4.1}$$

donde $x_{ij}, y_i \in \mathbb{R}$ para todo $i \in \{1, \ldots, m\}$ y $j \in \{1, \ldots, n\}$. Cada fila de la matriz $M$ representa un **caso** muestral, que se denota por

$$\mathbf{c_i} = (\mathbf{x}_i, y_i), \quad i = 1, \ldots, m,$$

donde $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ es el vector de condiciones antecedentes, y $y_i$ es el valor consecuente (o resultado).

Para analizar los datos mediante la metodología fsQCA, cada valor de los datos muestrales de la matriz $M$ se transforma (o calibra) mediante una función de pertenencia, que renormaliza los datos al intervalo $[0, 1]$. Existen multitud de posibilidades para la definición de la función de pertenencia, tal y como se mencionó en el Capítulo 2, pero una de las opciones más empleadas, sobre todo en el paquete `QCA` de R, es la siguiente (Duşa, 2019):

$$\mu(x) = \begin{cases} \dfrac{\exp\left(-\log\text{-}odds(x_1) \cdot \dfrac{(x - x_2)}{x_2 - x_1}\right)}{1 + \exp\left(-\log\text{-}odds(x_1) \cdot \dfrac{(x - x_2)}{x_2 - x_1}\right)}, & \text{si } x < x_2 \\ \dfrac{\exp\left(\log\text{-}odds(x_3) \cdot \dfrac{(x - x_2)}{x_3 - x_1}\right)}{1 + \exp\left(\log\text{-}odds(x_3) \cdot \dfrac{(x - x_2)}{x_3 - x_1}\right)}, & \text{si } x \geq x_2 \end{cases} \tag{4.2}$$

Esta función depende de varios parámetros. Por un lado están los llamados **puntos de anclaje**, $x_1$, $x_2$ y $x_3$, donde $x_1$ es el punto de plena exclusión (valor bajo, pertenencia = 0), $x_2$ el punto de cruce o umbral (pertenencia = 0.5) y $x_3$ corresponde al punto de plena inclusión (valor alto, pertenencia = 1).

**Definición 4.2.** La función $\mu$ transforma cada uno de los datos de la matriz $M$ en una segunda matriz $M_\mu$, denominada **matriz calibrada**.

$$M_\mu = \begin{array}{c|cccc|c} & V_1 & V_2 & \cdots & V_n & R \\ \hline \mathbf{c}_1 & \mu(x_{11}) & \mu(x_{12}) & \cdots & \mu(x_{1n}) & \mu(y_1) \\ \mathbf{c}_2 & x_\mu(x_{21}) & \mu(x_{22}) & \cdots & \mu(x_{2n}) & \mu(y_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{c}_m & \mu(x_{m1}) & \mu(x_{m2}) & \cdots & \mu(x_{mn}) & \mu(y_m) \end{array}, \tag{4.3}$$

donde $\mu(x)_{ij}$, $\mu(y)_i \in [0,1]$, para todo $i \in \{1,\ldots,m\}$ y $j \in \{1,\ldots,n\}$.

Para construir la tabla de verdad, fsQCA transforma los valores de la matriz calibrada $M_\mu$ en valores binarios, siguiendo la siguiente regla: $\mu > 0.5 \Rightarrow 1$, $\mu < 0.5 \Rightarrow 0$, y $\mu = 0.5$ indica un punto de cruce o de máxima ambigüedad. Este punto de máxima ambigüedad corresponde al punto de corte central $x_2$ dado en la ecuación (4.2).

Cada combinación de condiciones (presencia o ausencia) se representa como una esquina del **cubo causal**, según lo propuesto por Rihoux y Ragin (2009). Estas esquinas corresponden a los vértices de un hipercubo de dimensión $2^{n+1}$, donde $n$ es el número de condiciones consideradas. Cada vértice define una **configuración causal** única de presencia/ausencia de las condiciones.

Cada vértice corresponde a un caso real observado o a una configuración teóricamente posible. En el proceso de análisis, todas las combinaciones posibles son listadas en la tabla de verdad, indicando el número de casos que se ajustan a cada configuración y el resultado observado asociado. Configuraciones sin casos empíricos también son consideradas y, dependiendo del criterio del investigador, pueden ser utilizadas o descartadas en las etapas de simplificación lógica.

Formalmente, la **tabla de verdad** binaria $\mathcal{B}$ se define como una función:

$$\mathcal{B} : [0,1]^n \times [0,1] \to \{0,1\}^n \times \{0,1\},$$

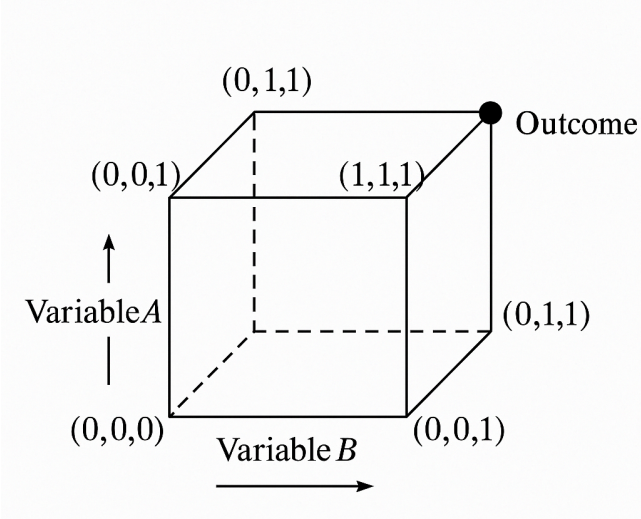


Figura 4.1: Binarización de variables y representación en el cubo causal

mediante la aplicación del operador de binarización para cada caso $\mathbf{c}_i$, con $i = 1, \ldots, m$:

$$\mathcal{B}(\boldsymbol{\mu}(\mathbf{x}_i), \mu(y_i)) = (\mathbf{z}_i, r_i),$$

donde:

- $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{in})$, con $z_{ij} = \mathbb{I}_{(0.5,\,1]}(\mu(x_{ij}))$, para $j = 1, \ldots, n$;
- $r_i = \mathbb{I}_{(0.5,\,1]}(\mu(y_i))$;
- $\mathbb{I}_{(0.5,\,1]}(t)$ es la función indicadora, tal que:

$$\mathbb{I}_{(0.5,\,1]}(t) = \begin{cases} 1, & \text{si } t > 0.5, \\ 0, & \text{si } t < 0.5. \end{cases}$$

La tabla de verdad resultante está compuesta por las configuraciones binarias $\mathbf{z}_i \in \{0,1\}^n$ y los resultados binarios $r_i \in \{0,1\}$, y se representa como el conjunto (posiblemente con repeticiones):

$$\mathcal{B} = \{(\mathbf{z}_i, r_i) \mid i = 1, \ldots, m\}.$$

A partir de $\mathcal{B}$, se identifican las configuraciones únicas, su frecuencia de ocurrencia para cada combinación causal.

**Definición 4.3.** De manera análoga al csQCA, un conjunto de datos en fsQCA se dice **consistente** si, para cualesquiera dos casos transformados en la tabla de verdad,

$$\mathbf{c}_{i_1} = (\mathbf{z}_{i_1}, r_{i_1}) \quad \text{y} \quad \mathbf{c}_{i_2} = (\mathbf{z}_{i_2}, r_{i_2}),$$

se cumple la siguiente condición:

$$\text{si } \mathbf{z}_{i_1} = \mathbf{z}_{i_2}, \text{ entonces necesariamente } r_{i_1} = r_{i_2}.$$

En otras palabras, la **consistencia** requiere que, siempre que dos casos presenten la misma configuración causal binarizada $\mathbf{z}_i \in \{0,1\}^n$, también presenten el mismo resultado binarizado $r_i \in \{0,1\}$. Es decir, no deben existir contradicciones empíricas entre los casos observados en la tabla de verdad.

**Definición 4.4.** Una configuración en la tabla de verdad $\mathbf{z}_i \in \{0,1\}^n$ se dice **conducente** al resultado si existe al menos un caso $\mathbf{c}_i = (\mathbf{z}_i, r_i)$ tal que $r_i = 1$.

**Definición 4.5.** Dos configuraciones binarias de orden $k$,

$$\mathbf{z}^{(1)} = (z_1^{(1)}, z_2^{(1)}, \ldots, z_k^{(1)}) \quad \text{y} \quad \mathbf{z}^{(2)} = (z_1^{(2)}, z_2^{(2)}, \ldots, z_k^{(2)})$$

pueden **minimizarse** (o reducirse) a una configuración de orden $k-1$ si se cumplen las siguientes condiciones:

- Ambas configuraciones son conducentes al resultado, es decir, existen casos $(\mathbf{z}^{(1)}, r^{(1)}), (\mathbf{z}^{(2)}, r^{(2)})$ en la tabla de verdad tal que $r^{(1)} = r^{(2)} = 1$;
- Existe un índice $p \in \{1, \ldots, k\}$ tal que

$$\left| \mathbf{z}^{(1)} - \mathbf{z}^{(2)} \right| = \mathbf{e}_p, \tag{4.4}$$

  donde $\mathbf{e}_p \in \{0,1\}^k$ es el $p$-ésimo vector de la base canónica (es decir, $\mathbf{e}_p = (0, \ldots, 0, 1, 0, \ldots, 0)$, con 1 en la posición $p$).

En ese caso, $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$ pueden reducirse a la configuración de orden $k-1$ eliminando la condición correspondiente a la posición $p$:

$$\mathbf{z} = (z_1^{(1)}, \ldots, z_{p-1}^{(1)}, z_{p+1}^{(1)}, \ldots, z_k^{(1)}),$$

que representa una configuración reducida pero aún conducente al resultado.

**Definición 4.6.** Si dos configuraciones no cumplen la condición de minimización anterior, se dice que son **irreducibles**.

**Definición 4.7.** Las **soluciones** del modelo fsQCA se definen como el conjunto:

$$\mathcal{C} = \bigcup_{k=1}^{n} \mathbf{C}_k,$$

donde cada conjunto $\mathbf{C}_k$ contiene todas las configuraciones binarias de orden $k$, irreducibles y conducentes al resultado. Específicamente,

$$\mathbf{C}_k = \left\{ \mathbf{z}_k^{(p)} \in \{0,1\}^k \mid \mathbf{z}_k^{(p)} \text{ es conducente y no reducible con ninguna otra en } \mathbf{C}_k \right\}.$$

En resumen, los principios lógicos y estructurales que rigen la minimización de configuraciones y la identificación de soluciones en fsQCA son conceptualmente equivalentes a los de csQCA. La diferencia radica esencialmente en el grado de pertenencia de los casos

que solo se utiliza para el cálculo de métricas, pero la lógica subyacente de reducción y de búsqueda de configuraciones irreducibles permanece inalterada.

**Definición 4.8.** La **Consistencia** en la metodología fsQCA se calcula mediante la razón entre la suma de los mínimos entre la pertenencia de cada combinación causal y la pertenencia al resultado, y la suma total de las pertenencias de las combinaciones causales. Esta medida indica en qué medida una combinación causal es suficiente para explicar el resultado observado. Su fórmula se presenta en la Ecuación 4.5.

$$\text{Cons} = \frac{\sum_{i=1}^{m} \text{mín}\left\{\boldsymbol{\mu}(\mathbf{x}_i), \boldsymbol{\mu}(y_i)\right\}}{\sum_{i=1}^{m} \boldsymbol{\mu}(\mathbf{x}_i)} \tag{4.5}$$

Por su parte, la **Cobertura** se calcula mediante la razón entre la suma de los mínimos (como en el caso de la consistencia), pero esta vez en relación con la suma total de las pertenencias al resultado. Esta métrica representa la proporción del resultado explicado por una determinada combinación causal. La Ecuación 4.6 presenta su expresión formal:

$$\text{Cov} = \frac{\sum_{i=1}^{m} \text{mín}\left\{\boldsymbol{\mu}(\mathbf{x}_i), \boldsymbol{\mu}(y_i)\right\}}{\sum_{i=1}^{m} \boldsymbol{\mu}(y_i)} \tag{4.6}$$

## 4.2 MINERÍA DE REGLAS DE ASOCIACIÓN PARA VARIABLES CUANTITATIVAS

A diferencia de las tradicionales bases de datos de "cesta de mercado", compuestas mayoritariamente por atributos categóricos, las bases de datos actuales, presentes en áreas como salud, finanzas, bioinformática y marketing, incluyen una gran cantidad de atributos cuantitativos (numéricos), lo que exige nuevas técnicas para un análisis eficaz.

En este contexto, surge el concepto de Reglas de Asociación Cuantitativas, del inglés, (Quantitative Association Rules – QARs), una extensión de las Reglas de asociación booleanas, propuesta originalmente por R. Agrawal et al. (1993), y posteriormente extendida en Aggarwal y Yu, 1998

**Definición 4.9.** Las **Reglas de Asociación Cuantitativas** (QAR, por sus siglas en inglés) son expresiones del tipo:

$$\text{SI } \langle X_1[l_{X_1}, u_{X_1}]\rangle \text{ Y } \langle X_2[l_{X_2}, u_{X_2}]\rangle \text{ O } \langle X_3[l_{X_3}, u_{X_3}]\rangle, \text{ ENTONCES } \langle R[l_R, u_R]\rangle,$$

donde cada par $\langle X[l_X, u_X]\rangle$ representa un atributo $X$ asociado a un intervalo numérico con límite inferior $l_X$ y límite superior $u_X$.

Los atributos $X_1, X_2, X_3$ y $R$ corresponden a condiciones con soporte y confianza dentro de un conjunto de datos cuantitativos. Estas reglas se extraen a partir de datos numéricos mediante la discretización de variables o intervalos numéricos.

**Definición 4.10.** Sea $I = \{x_1, x_2, \ldots, x_m\}$ un conjunto de atributos, donde cada atributo $x_j$ puede ser:

- **Categórico**, como "género", "estado civil", "nivel educativo";

- **Cuantitativo**, como “salario”, “edad”, “número de hijos”.

Cada atributo tiene un dominio de valores posibles denotado por $\text{dom}(x_j)$.

**Definición 4.11.** Un **ítem** se representa como un par:

$$x[\ell_x, u_x]$$

donde el valor del atributo $x$ está dentro del intervalo $[\ell_x, u_x]$.

- Si $x$ es categórico, entonces $\ell_x = u_x$, indicando un valor exacto.
- Si $x$ es cuantitativo, entonces $\ell_x \leq u_x$, indicando un intervalo numérico.

**Definición 4.12.** Un **ítemset** es un conjunto de ítems sobre atributos distintos:

$$X = \{x_1[\ell_{x_1}, u_{x_1}], \ldots, x_k[\ell_{x_k}, u_{x_k}]\}$$

**Definición 4.13.** Una **transacción** se define como un vector:

$$T = \langle v_1, v_2, \ldots, v_m \rangle$$

que contiene valores concretos para cada atributo del conjunto $I$.

**Definición 4.14.** Una transacción $T$ **soporta** un ítemset $X$ si, para cada ítem $x_i[\ell_i, u_i] \in X$, el valor correspondiente $v_i$ en $T$ satisface:

$$\ell_i \leq v_i \leq u_i$$

El **soporte** de un ítemset $X$, denotado $supp(X)$, es la fracción de transacciones en la base de datos $D$ que soportan $X$:

$$supp(X) = \frac{freq(X)}{|D|}$$

donde $freq(X)$ es el número de transacciones que soportan $X$.

**Definición 4.15.** Una **QAR** es una implicación del tipo $X \Rightarrow Y$, igual a las reglas binarias, donde $X$ y $Y$ son *ítemsets* tales que $\text{attr}(X) \cap \text{attr}(Y) = \emptyset$, es decir, no comparten atributos.

La regla expresa que, cuando se satisfacen las condiciones en $X$, es probable que también se cumplan las condiciones en $Y$.

Una regla cuantitativa puede expresarse de la forma:

*Si Salario* $\in [40k, 50k]$ *y Casado = Sí, entonces NumPréstamos = 2.*

En esta regla: indica que los individuos **casados** con salario entre 40k y 50k tienden a tener **dos préstamos**.

Las métricas principales utilizadas para evaluar una QAR son análogas a las reglas binarias tradicionales.

**Definición 4.16.** Una **QAR negativa** es una regla de asociación que expresa una relación inversa entre los ítemsets. Por ejemplo:

$$\text{Edad} \notin [20, 30] \Rightarrow \text{Gastos} \in [100, 300]$$

Esta regla indica que si un individuo no tiene una edad entre 20 y 30 años, entonces sus gastos tienden a estar entre 100 y 300 unidades monetarias.

Las QARs negativas amplían el poder descriptivo del modelo al capturar asociaciones inversas entre atributos.

**Definición 4.17.** La **discretización** es una técnica utilizada para transformar atributos cuantitativos continuos en intervalos discretos (o categorías), facilitando la aplicación de algoritmos de minería de reglas.

Esta transformación permite emplear algoritmos clásicos adaptados, como **Apriori**, sobre conjuntos de datos cuantitativos, generando reglas de asociación interpretables.

## 4.3 TEOREMA DE EQUIVALENCIA

A continuación estableceremos la formalización matemática demostrando que cualquier problema planteado bajo el enfoque de fsQCA puede reformularse como un caso particular de minería de reglas de asociación cuantitativa binarizadas.

Sea un modelo fsQCA definido como $(\mathcal{V}, R, M)$, donde:

- $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$ es el conjunto de condiciones causales,
- $R$ representa el resultado,
- $M = (\mathbf{x}_{ij}, y_i)$ es la matriz de datos empíricos, con $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$.

Cada variable $V_j$ y el resultado $R$ son calibrados mediante una función logística $\mu(x)$, definida por tres puntos de anclaje $k_1 < k_2 < k_3$, que transforma los valores en el intervalo $[0, 1]$. Esto genera una nueva matriz calibrada $\mu(M)$.

A partir de $\mu(M)$, discretizamos los valores usando el punto de máxima ambigüedad 0.5:

$$\text{Si } \mu(x_{ij}) > 0.5, \text{ entonces asignamos } x_{ij} = 1,$$

$$\text{Si } \mu(x_{ij}) < 0.5, \text{ entonces asignamos } x_{ij} = 0.$$

De esta forma, transformamos la matriz calibrada en una matriz binaria que facilita la construcción de una base transaccional.

Definimos ahora el conjunto de ítems binarios como:

$$I = \{v_1, V_1, v_2, V_2, \ldots, v_n, V_n, r, R\}$$

donde:

- $V_j$ representa la presencia de la condición $V_j$, es decir, $x_j = 1$, correspondiente al intervalo calibrado $x_j \in (0.5, 1]$,

- $v_j$ representa la ausencia o negación de la condición $V_j$, es decir, $x_j = 0$, correspondiente a $x_j \in [0, 0.5)$,
- $R$ indica la ocurrencia del resultado ($y_i = 1$),
- $r$ indica la ausencia del resultado ($y_i = 0$).

Así, por ejemplo:

$$x = x[0.5, 1] \text{ denota la presencia de la condición (ítem } V),$$

$$\sim x = x[0, 0.5] \text{ denota su negación (ítem } v),$$

el complemento se refiere al par de ítems mutuamente excluyentes $V$ y $v$.

Con base en esto, cada caso en fsQCA puede representarse como una transacción binaria:

$$\tau_i = \{\phi_{i1}, \phi_{i2}, \ldots, \phi_{in}, \eta_i\}$$

donde:

$$\phi_{ij} = \begin{cases} V_j & \text{si } x_{ij} = 1 \\ v_j & \text{si } x_{ij} = 0 \end{cases}, \quad \eta_i = \begin{cases} R & \text{si } y_i = 1 \\ r & \text{si } y_i = 0 \end{cases}$$

Esto define la base de datos transaccional $T = \{\tau_1, \ldots, \tau_m\}$, compuesta por $m$ transacciones.

Una configuración causal en fsQCA, por ejemplo $S_k = (\mathcal{V}_k, \mu)$, donde $\mathcal{V}_k = \{V_{j_1}, \ldots, V_{j_k}\}$, conduce al resultado $R$. Esta configuración puede representarse como una regla de asociación cuantitativa (QAR):

$$X \Rightarrow R$$

donde:

$$X = \{\psi_1, \ldots, \psi_k\}, \quad \psi_l = \begin{cases} V_{j_l} & \text{si } \mu(V_{j_l}) > 0.5 \\ v_{j_l} & \text{si } \mu(V_{j_l}) < 0.5 \end{cases}$$

Llamamos $\mathcal{T}(S_k) = X \Rightarrow R$ a la transformación de la configuración causal en una QAR. Esta transformación establece una correspondencia biunívoca entre las configuraciones fsQCA y las reglas de asociación cuantitativas, sobre la base de la discretización del conjunto de datos calibrados en dos trozos.

- Una de las particularidades es que la métricas en fsQCA se calcula sobre la base de los datos calibrados (valores continuos en el intervalo [0, 1]), mientras que en las reglas de asociación cuantitativas se calcula a partir de los datos ya discretizados (valores binarios).

Con estas correspondencias podemos afirmar:

**Teorema de Equivalencia QAR-fsQCA.** Todo modelo fsQCA puede ser interpretado como un caso especial de minería de reglas de asociación cuantitativa (QAR), en su versión binarizada en dos intervalos. Esta reformulación permite utilizar algoritmos y métricas de

minería de datos para identificar soluciones causales consistentes con los principios de fsQCA.

*Teorema* 4.1 (Teorema de equivalencia para fsQCA). Sea $(\mathcal{V}, R, M)$ un modelo fsQCA y sea $(I, T)$ el problema de minería de reglas de asociación cuantitativas definido anteriormente en esta sección. Entonces

$$\mathcal{T}(\mathcal{C}) \subseteq \mathcal{AR},$$

donde $\mathcal{AR}$ denota el conjunto de todas las reglas interesantes con umbrales mínimos de soporte y confianza dados por

$$\text{minsupp} = \frac{1}{m} \qquad \text{minconf} = 1.$$

*Demostración.* La demostración es una consecuencia directa del hecho de que, tal y como se ha descrito en anteriormente, las soluciones del problema de fsQCA se obtienen a partir de una binarización de la matriz $M$, es decir, de una transformación del problema a uno de csQCA. Por lo tanto, por el Teorema 3.1, tenemos la equivalencia con el problema de reglas de asociación binarizadas (que es equivalente al problema de NPAR descrito en dicho Teorema. □

## 4.4 EJEMPLO: CAÍDA DE LA DEMOCRACIA ENTRE LA PRIMERA Y LA SEGUNDA GUERRA MUNDIAL

### 4.4.1 *Enfoque fsQCA*

Con el objetivo de ilustrar el enfoque fsQCA y las reglas de asociación cuantitativas, se adopta un ejemplo empírico basado en los datos presentados en la Tabla 5.2 de Rihoux y Ragin (2009). Dicho estudio examina los factores asociados a la caída de la democracia en países europeos entre la Primera y la Segunda Guerra Mundial. Se consideran tres condiciones causales: desarrollo económico (DV), urbanización (UR) y alfabetización (LT), en relación con la supervivencia del estado democrático (SV) y, sobre todo, con su complementario, que sería la caída de la democracia (BD).

La calibración de las variables se realizó mediante una función de pertenencia logística directa, tal como lo propone Rihoux y Ragin (2009), lo que permitió asignar puntuaciones de pertenencia a las condiciones DV, UR y LT en una muestra de 18 países europeos, identificados como "Casos 1–18" (véase la Tabla 4.1). Para garantizar la comparabilidad entre los resultados, se utilizó un criterio uniforme para definir el punto de corte central de cada variable.

El procedimiento de calibración de los datos observados fue implementado en el software R, siguiendo la metodología descrita por Duşa (2019), basada en tres puntos de anclaje para cada variable. Los puntos de corte utilizados en la calibración para cada variable fueron los siguientes: **DV** : $(398.5;\ 552;\ 871.5)$, **UR** : $(25.87; 33.7; 53.125)$, **LT** : $(73.3; 89.7; 98)$, **BD** : $(-8.0; 1.5; 9.5)$.

Las puntuaciones resultantes de la calibración pueden consultarse en la Tabla 4.1, que resume los datos empleados en el análisis comparativo posterior.

Tabla 4.1: Matriz de datos y puntuaciones de pertenencia difusa

| N.º | País | DV | $\mu$(DV) | UR | $\mu$(UR) | LT | $\mu$(LT) | SV | $\mu$(SV) | $\mu$(BD) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Austria | 720 | 0.82 | 33.40 | 0.47 | 98.00 | 0.95 | -9.00 | 0.04 | 0.96 |
| 2 | Bélgica | 1098 | 0.99 | 60.50 | 0.98 | 94.40 | 0.84 | 10.00 | 0.96 | 0.04 |
| 3 | Checoslovaquia | 586 | 0.58 | 69.00 | 1.00 | 95.90 | 0.90 | 7.00 | 0.88 | 0.12 |
| 4 | Estonia | 468 | 0.17 | 28.50 | 0.12 | 95.00 | 0.87 | -6.00 | 0.09 | 0.91 |
| 5 | Finlandia | 590 | 0.59 | 22.00 | 0.01 | 99.10 | 0.97 | 4.00 | 0.72 | 0.28 |
| 6 | Francia | 983 | 0.98 | 21.20 | 0.01 | 96.20 | 0.91 | 10.00 | 0.96 | 0.04 |
| 7 | Alemania | 795 | 0.90 | 56.50 | 0.97 | 98.00 | 0.95 | -9.00 | 0.04 | 0.96 |
| 8 | Grecia | 390 | 0.04 | 31.10 | 0.27 | 59.20 | 0.00 | -8.00 | 0.05 | 0.95 |
| 9 | Hungría | 424 | 0.08 | 36.30 | 0.60 | 85.00 | 0.30 | -1.00 | 0.32 | 0.68 |
| 10 | Irlanda | 662 | 0.73 | 25.00 | 0.04 | 95.00 | 0.87 | 8.00 | 0.92 | 0.08 |
| 11 | Italia | 517 | 0.34 | 31.40 | 0.30 | 72.10 | 0.04 | -9.00 | 0.04 | 0.96 |
| 12 | Países Bajos | 1008 | 0.99 | 78.80 | 1.00 | 99.90 | 0.97 | 10.00 | 0.96 | 0.04 |
| 13 | Polonia | 350 | 0.02 | 37.00 | 0.62 | 76.90 | 0.09 | -6.00 | 0.09 | 0.91 |
| 14 | Portugal | 320 | 0.01 | 15.30 | 0.00 | 38.00 | 0.00 | -9.00 | 0.04 | 0.96 |
| 15 | Rumanía | 331 | 0.01 | 21.90 | 0.01 | 61.80 | 0.01 | -4.00 | 0.15 | 0.85 |
| 16 | España | 367 | 0.03 | 43.00 | 0.80 | 55.60 | 0.00 | -8.00 | 0.05 | 0.95 |
| 17 | Suecia | 897 | 0.96 | 34.00 | 0.51 | 99.90 | 0.97 | 10.00 | 0.96 | 0.04 |
| 18 | Reino Unido | 1038 | 0.99 | 74.00 | 1.00 | 99.90 | 0.97 | 10.00 | 0.96 | 0.04 |

Fuente: Adaptado de Rihoux y Ragin (2009).

La Tabla 4.2 presenta la tabla de verdad correspondiente a los datos de la Tabla 4.1, que contiene todas las combinaciones causales posibles derivadas de las tres condiciones analizadas. Las columnas 2, 3 y 4 representan las condiciones causales codificadas en valores binarios: el valor 1 indica la presencia de la condición, mientras que el valor 0 representa la presencia de su complemento. Por ejemplo, la combinación dv*UR*lt se representa como $(0,1,0)$. La columna 5 ("OUT") corresponde con el resultado (output) que queremos estudiar. En este caso el resultado es caída de la democracia (BD) que corresponde con la negación de la variable SV de la Tabla 4.1, es decir los datos calibrados de la variable BD, serían los complementarios de los datos calibrados de la variable BD.

En la última columna, se indican los números de identificación de los países correspondientes a cada configuración, de acuerdo con su posición en la tabla de datos original 4.1. Por ejemplo, el número 16 representa a España.

Del análisis se observa que tres combinaciones específicas de condiciones están asociadas con la caída de la democracia.

La columna "Vector" expresa cada configuración causal utilizando la notación conjuntiva, destacando la interacción entre las condiciones presentes y ausentes. La columna "n" indica el número de casos cuya puntuación de pertenencia ($\mu$) supera el umbral de 0.5.

Las medidas de consistencia y cobertura son calculadas utilizando las Ecuaciones 4.5 y 4.6, respectivamente. Estos cálculos se basan en los valores de pertenencia fuzzy, obteni-

Tabla 4.2: Tabla de verdad

| | DV | UR | LT | Vector | OUT (BD) | n | Cons | cov | cases |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | **dv*ur*lt** | 1 | 4 | 0.955 | 0.942 | 8,11,14,15 |
| 3 | 0 | 1 | 0 | **dv*UR*lt** | 1 | 3 | 1.000 | 1.000 | 9,13,16 |
| 2 | 0 | 0 | 1 | **dv*ur*LT** | 1 | 1 | 0.861 | 0.741 | 4 |
| 8 | 1 | 1 | 1 | **DV*UR*LT** | 0 | 6 | 0.357 | 0.272 | 2,3,7,12,17,18 |
| 6 | 1 | 0 | 1 | **DV*ur*LT** | 0 | 4 | 0.374 | 0.202 | 1,5,6,10 |
| 4 | 0 | 1 | 1 | **dv*UR*LT** | ? | 0 | - | - | - |
| 5 | 1 | 0 | 0 | **DV*ur*lt** | ? | 0 | - | - | - |
| 7 | 1 | 1 | 0 | **DV*UR*lt** | ? | 0 | - | - | - |

dos a partir de la función de pertenencia $\mu$, para los 18 casos considerados en el análisis. Los resultados obtenidos se resumen en la Tabla 4.2.

Aplicando un umbral de consistencia estándar de 0.75, se identifican tres combinaciones causales efectivas que conducen a la caída de la democracia. Estas combinaciones corresponden a las tres primeras filas de la Tabla 4.2 y están representadas por las siguientes configuraciones:

$$S_3^{(1)} = \begin{bmatrix} \text{DV} & \text{UR} & \text{LT} \\ 0 & 0 & 0 \end{bmatrix} \quad S_3^{(2)} = \begin{bmatrix} \text{DV} & \text{UR} & \text{LT} \\ 0 & 1 & 0 \end{bmatrix}$$
$$S_3^{(3)} = \begin{bmatrix} \text{DV} & \text{UR} & \text{LT} \\ 0 & 0 & 1 \end{bmatrix}$$

El *implicante primo* para las tres condiciones causales se obtiene de forma sencilla, a partir del agrupamiento de combinaciones causales que difieren en solo una condición. Este proceso permite la reducción de términos mediante el algoritmo de Quine–McCluskey. A continuación, se ilustra esta simplificación:

$$\left.\begin{matrix} S_3^{(1)} \\ S_3^{(2)} \end{matrix}\right\} \rightarrow S_2^{(1)} = \begin{bmatrix} \text{DV} & \text{LT} \\ 0 & 0 \end{bmatrix}, \qquad \left.\begin{matrix} S_3^{(1)} \\ S_3^{(3)} \end{matrix}\right\} \rightarrow S_2^{(2)} = \begin{bmatrix} \text{DV} & \text{UR} \\ 0 & 0 \end{bmatrix}.$$

Por tanto, la minimización de la tabla de verdad para este caso conduce a la siguiente fórmula explicativa:

$$\mathcal{C} : dv * ur + dv * lt \rightarrow \text{BD}$$

En resumen, se han identificado dos configuraciones causales que explican la caída de la democracia en Europa durante las primera e segunda guerra.

1. El bajo desarrollo económico combinado con un bajo nivel educativo de la población.

2. El bajo desarrollo económico junto con una urbanización limitada.

En lo que respecta a las métricas, cabe destacar que, como se muestra en la Tabla 4.2, las solución presentan un nivel de consistencia superior a 0.86, lo que indica una fuerte relación entre las configuraciones causales y el resultado observado.

### 4.4.2 *Enfoque con Reglas de Asociación Quantitativas*

Desde el punto de vista de la extracción de reglas de asociación cuantitativas, aplicamos el método al problema de la caída de la democracia en Europa.

Inicialmente, realizamos un mapeo de los atributos o variables de la base de datos calibrada (ver Tabla 4.1), que se presentan en dominios cuantitativos, hacia un dominio binario. Para ello, utilizamos el punto de anclaje 0.5 como umbral para cada una de las variables. Cada atributo, junto con sus respectivos intervalos de valores, da origen a nuevas columnas binarias.

En este proceso, cada variable es categorizada en dos rangos de valores: aquellos mayores que 0.5 se convierten en 1 presencia) y aquellos menores a 0.5 se convierten en 0 (no presencia). De este modo, se definen los ítems como:

$$I = \{DV, dv, UR, ur, LT, lt, BD, bd\}, \tag{4.7}$$

donde cada par representa la presencia (mayúsculas) o ausencia (minúsculas) de una condición calibrada.

La Tabla 4.3 presenta los resultados del mapeo de la Tabla 4.1 al dominio binario. En este proceso, los valores de las variables cuantitativas han sido transformados en variables binarias, según los rangos definidos anteriormente.

Para la minería de reglas de asociación, utilizamos el algoritmo Apriori, que permite identificar patrones frecuentes y extraer reglas significativas entre los atributos binarizados de la base de datos.

Como se mencionó anteriormente, Apriori es un algoritmo ascendente que comienza con los conjuntos de ítems más pequeños y va incrementando su tamaño gradualmente, eliminando combinaciones que no superan el umbral mínimo de soporte. En nuestro caso, y de acuerdo con el Teorema 3.1, el umbral mínimo de soporte se define como:

$$\text{mín supp} = \frac{1}{m} = \frac{1}{18} \approx 0.05,$$

donde $m$ representa el número total de observaciones en la base de datos.

El primer paso consiste en extraer los conjuntos de ítems que contienen la variable de respuesta BD y analizar cuáles de estos conjuntos cumplen con las condiciones mínimas de soporte y confianza.

La Tabla 4.4 presenta los conjuntos de ítems que superaron el umbral mínimo de soporte, junto con la regla de asociación que definen y la confianza correspondiente. Cabe destacar que se resaltan las reglas cuya confianza alcanza el valor máximo de 1, lo cual indica una relación determinística entre el antecedente y la consecuencia.

Tabla 4.3: Resultados de la binarización.

| # | Países | dv | DV | ur | UR | lt | LT | bd | BD |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Austria | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 2 | Bélgica | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 3 | Checoslovaquia | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 4 | Estonia | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 5 | Finlandia | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 6 | Francia | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 7 | Alemania | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 8 | Grecia | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 9 | Hungría | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 10 | Irlanda | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 11 | Italia | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 12 | Países Bajos | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 13 | Polonia | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 14 | Portugal | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 15 | Rumanía | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 16 | España | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 17 | Suecia | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 18 | Reino Unido | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |

A partir de los resultados obtenidos en la Tabla 4.4, es evidente que se puede establecer una correspondencia clara entre las configuraciones causales identificadas mediante el enfoque fsQCA y las reglas de asociación cuantitativas resultantes de la discretización en dos niveles. Se pueden destacar los siguientes puntos:

1. Todas las soluciones obtenidas por el algoritmo fsQCA son subconjuntos de las soluciones derivadas de las reglas de asociación cuantitativas discretizadas en dos intervalos. Esto implica que cualquier problema formulado bajo el enfoque fsQCA puede reformularse como un caso particular de minería de reglas de asociación cuantitativas binarizadas, tal como se enuncia en el Teorema 4.1.

2. Las soluciones obtenidas mediante las reglas de asociación presentadas en la tabla mencionada pueden simplificarse eliminando reglas redundantes. Por ejemplo: las reglas `r9` y `r10` son redundantes entre sí y pueden eliminarse; lo mismo ocurre entre `r9` y `r11`; `r3` y `r8`; `r4` y `r5`; y `r6` y `r7`. Luego de eliminar las redundancias, las soluciones no redundantes que permanecen son `r1` y `r2`.

El Diagrama 4.2 muestra la solución más parsimoniosa obtenida a partir de las reglas de asociación.

Según los datos de la Tabla 4.3, se verifica que un bajo nivel de alfabetización implica un bajo desarrollo económico. Solo Alemania y Austria, países con alto desarrollo económico, se configuraron como dictaduras entre las guerras mundiales. Es decir, en términos

Tabla 4.4: Reglas de asociación con consecuente *BD*

| Regla | LHS | RHS | Soporte | Confianza | Cobertura | Lift | Conteo |
|---|---|---|---|---|---|---|---|
| $r_1$ | {lt} | {BD} | 0.3889 | 1 | 0.3889 | 1.8 | 7 |
| $r_2$ | {dv} | {BD} | 0.4444 | 1 | 0.4444 | 1.8 | 8 |
| $r_3$ | {dv, lt} | {BD} | 0.3889 | 1 | 0.3889 | 1.8 | 7 |
| $r_4$ | {lt, ur} | {BD} | 0.2222 | 1 | 0.2222 | 1.8 | 4 |
| $r_5$ | {lt, UR} | {BD} | 0.1667 | 1 | 0.1667 | 1.8 | 3 |
| $r_6$ | {dv, ur} | {BD} | 0.2778 | 1 | 0.2778 | 1.8 | 5 |
| $r_7$ | {dv, UR} | {BD} | 0.1667 | 1 | 0.1667 | 1.8 | 3 |
| $r_8$ | {dv, LT} | {BD} | 0.0556 | 1 | 0.0556 | 1.8 | 1 |
| $r_9$ | {dv, lt, ur} | {BD} | 0.2222 | 1 | 0.2222 | 1.8 | 4 |
| $r_{10}$ | {dv, lt, UR} | {BD} | 0.1667 | 1 | 0.1667 | 1.8 | 3 |
| $r_{11}$ | {dv, LT, ur} | {BD} | 0.0556 | 1 | 0.0556 | 1.8 | 1 |

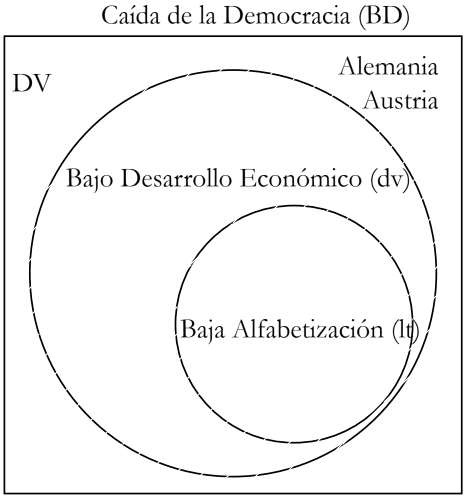


Figura 4.2: Diagrama de solución más simple

de soluciones lógicas, la existencia de un bajo nivel de desarrollo explica la caída de la democracia, como se expresa en la solución `r2`.

## 4.5 DISCRETIZACIÓN MULTI-INTERVALO DE ATRIBUTOS CONTINUOS

La discretización de atributos continuos constituye una etapa fundamental tanto en la metodología QCA como en la minería de reglas de asociación, ya que permite transformar valores numéricos en categorías discretas, facilitando así la formulación e interpretación de reglas lógicas.

Estas técnicas de discretización pueden clasificarse en **supervisadas** y **no supervisadas**, dependiendo de si utilizan o no la variable de salida durante el proceso.

### 4.5.1 *Discretización No Supervisada*

En la discretización no supervisada, los atributos explicativos se transforman en categorías sin considerar la variable objetivo. El propósito principal es identificar patrones latentes en los datos sin conocimiento previo de las clases de salida. La definición de los puntos de corte, en muchos casos, se basa en el *conocimiento experto*, ya sea proporcionado por especialistas o fundamentado en investigaciones previas. Ejemplos comunes de este enfoque son los métodos de **Anchura Igual** y **Frecuencia Igual**, ampliamente utilizados para convertir datos numéricos en nominales.

Por ejemplo, en un análisis sobre el retroceso democrático en Europa, se discretizó una variable continua en tres categorías lingüísticas (bajo, medio y alto) utilizando los percentiles 33 % y 66 % como umbrales. La Figura 4.3 ilustra gráficamente las reglas generadas mediante esta técnica.

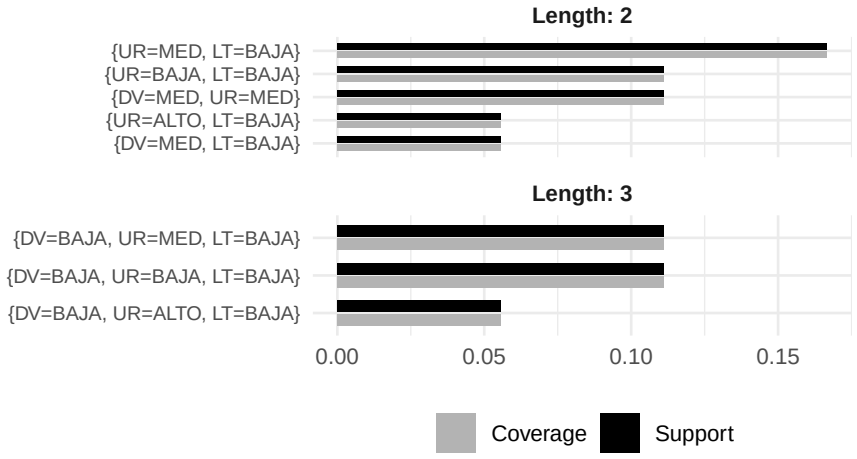


Figura 4.3: Reglas cuantitativas con tres categorías lingüísticas

#### 4.5.1.1 *Enfoque de Anchura Igual*

El enfoque de anchura igual es una técnica no supervisada que determina el valor mínimo $a_0$ y el valor máximo $a_k$ del atributo numérico, y divide el intervalo entre ellos en $k$ subintervalos de igual tamaño. El parámetro $k$ es definido por el investigador. La amplitud de discretización se calcula de la siguiente forma:

$$\lambda = \frac{a_k - a_0}{k} \tag{4.8}$$

Los puntos de corte generados forman el conjunto de límites $\{a_0, a_1, ..., a_k\}$, donde $a_i = a_0 + i\lambda$ para $i = 1, 2, ..., k$. Los intervalos resultantes son: $[a_0, a_1), [a_1, a_2), ..., [a_{k-1}, a_k]$. Esta técnica se aplica de manera independiente a cada atributo continuo.

#### 4.5.1.2 *Enfoque de Frecuencia Igual*

El enfoque de frecuencia igual también es una técnica no supervisada. Primero se ordenan los valores numéricos en forma ascendente, y luego se dividen en $k$ subintervalos, cada uno conteniendo aproximadamente la misma cantidad de observaciones. Los límites se definen como $\{b_0, b_1, ..., b_k\}$, donde cada intervalo $[b_{i-1}, b_i)$ contiene un número similar de valores.

Al utilizar diferentes valores de $k$ y aplicar distintas técnicas de discretización, es posible comparar el efecto de cada enfoque sobre la generación de reglas. Un mayor número de intervalos puede incrementar la precisión, pero tiende a reducir la cobertura de las reglas.

### 4.5.2 *Discretización Supervisada*

Las técnicas supervisadas utilizan la variable objetivo (también llamada clase o etiqueta) para guiar el proceso de discretización. El objetivo principal es optimizar la separación entre clases con base en los atributos explicativos, mejorando la capacidad predictiva del modelo. Además, estas técnicas contribuyen a eliminar categorías irrelevantes y reducir la complejidad del conjunto de datos.

Un método ampliamente utilizado es el MDLP, que realiza la discretización evaluando la ganancia de información al segmentar los datos según la variable de clase.

En términos generales, la discretización por múltiples intervalos busca dividir un atributo continuo en varios subintervalos significativos, evitando segmentaciones innecesarias y mejorando la calidad de las reglas generadas. Este proceso se fundamenta en métricas de entropía, que miden la impureza del conjunto de datos y se calculan según:

$$\text{Ent}(S) = -\sum_{i=1}^{k} P(C_i, S) \log_2(P(C_i, S))$$

donde:

- $P(C_i, S)$ es la proporción de ejemplos en $S$ que pertenecen a la clase $C_i$;
- $k$ es el número total de clases.

La entropía posterior a la partición de un atributo $A$ con punto de corte $T$ se calcula como:

$$E(A, T; S) = \frac{|S_1|}{|S|}\text{Ent}(S_1) + \frac{|S_2|}{|S|}\text{Ent}(S_2)$$

donde:

- $S_1$ contiene los ejemplos donde $A \leq T$;
- $S_2$ contiene los ejemplos donde $A > T$.

La ganancia de información se define como:

$$\text{Gain}(A, T; S) = \text{Ent}(S) - E(A, T; S)$$

Esta métrica cuantifica la reducción de incertidumbre al aplicar el punto de corte $T$, siendo clave para seleccionar divisiones informativas.

### 4.5.3 *Criterio MDLP*

El criterio MDLP establece si una partición de un atributo debe aceptarse en función del equilibrio entre la reducción de entropía y el costo adicional de codificación de clases tras la partición. La condición se expresa como:

$$\text{Gain}(A, T; S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A, T; S)}{N}$$

donde:

- $\text{Gain}(A, T; S)$ es la ganancia de información al dividir $S$ en dos subconjuntos mediante $T$;
- $N$ es el número total de ejemplos en $S$;
- $\Delta(A, T; S)$ representa el costo adicional de describir las clases tras la partición.

Esta desigualdad asegura que una partición solo se acepta si la reducción de entropía justifica el aumento en la complejidad, garantizando un balance entre simplicidad y precisión informativa.

#### 4.5.3.1 *Desempeño de los enfoques de discretización*

Utilizando el mismo conjunto de datos presentado en el estudio de Rihoux y Ragin (2009), realizamos un análisis con la base de datos completa, compuesta por seis variables. El objetivo fue evaluar el desempeño de diferentes estrategias de discretización en la minería de reglas de asociación.

La evaluación se llevó a cabo en función del número de candidatos y de reglas generadas por cada enfoque. Se compararon las técnicas de discretización por Anchura Igual, Frecuencia Igual y el criterio basado en MDLP. Para asegurar la consistencia, se fijaron los parámetros de soporte mínimo ($ms$) en los valores $ms \in \{0.04, 0.06, 0.08, 0.10, 0.12\}$. El número de intervalos ($k$) se fijó en $k = 3$ para todos los enfoques.

Los resultados correspondientes al número de candidatos y reglas generadas se presentan en la Figura 4.4.

El análisis gráfico muestra que todas las estrategias de discretización contribuyen al aumento en la generación de reglas de asociación. Sin embargo, el enfoque MDLP presentó el mejor desempeño general. Incluso bajo condiciones de soporte elevado ($soporte = 0.12$), MDLP fue capaz de generar reglas, a diferencia de las otras técnicas. La discretización por Frecuencia Igual mostró un desempeño intermedio, superando al enfoque de Anchura Igual, pero quedando por debajo del MDLP en términos de eficiencia.

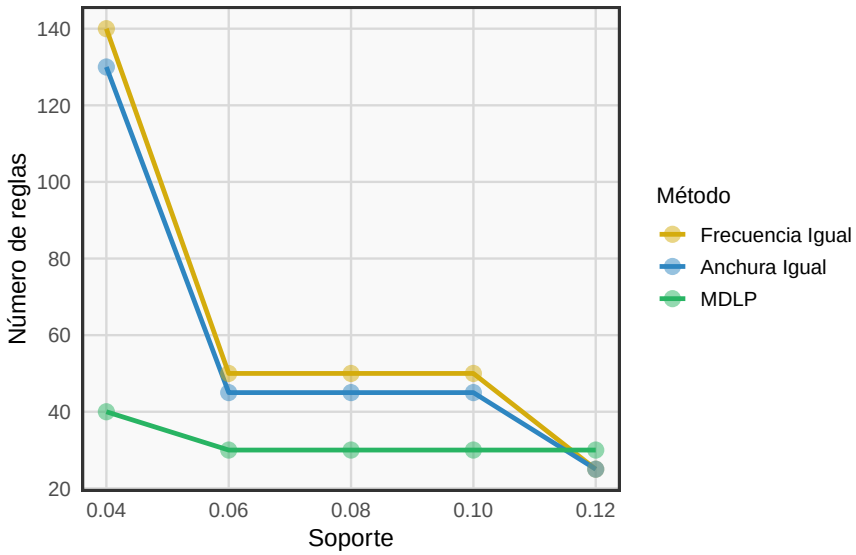


Figura 4.4: Número de reglas generadas para $k = 3$ intervalos

Los experimentos empíricos confirmaron la eficacia del criterio MDLP para discretizar variables continuas, favoreciendo la generación de reglas más cortas y con menor complejidad estructural. Esta característica resalta el potencial del MDLP como una alternativa superior en contextos de minería de reglas.

En contraste, la metodología fsQCA presenta limitaciones estructurales, ya que no permite soluciones con más de dos términos lingüísticos, lo que resulta en expresiones puramente booleanas. En cambio, la minería de reglas de asociación ofrece mayor flexibilidad lingüística y computacional, permitiendo capturar relaciones más complejas entre los atributos.

### 4.5.4 *Multi-value Qualitative Comparative Analysis (mvQCA)*

El método mvQCA es una extensión de la metodología QCA clásica propuesta originalmente por Cronqvist (2004), que permite trabajar con variables que asumen más de dos categorías, a diferencia del csQCA, que opera únicamente con condiciones dicotómicas. En lugar de transformar todas las variables en conjuntos binarios (0/1), la mvQCA acepta categorías múltiples (por ejemplo, 0, 1, 2, etc.), lo que permite mantener una mayor fidelidad a la complejidad de los datos sociales empíricos (Cronqvist, 2004; Rihoux, 2006).

Una ventaja destacada de la mvQCA es su capacidad para tratar de forma más eficaz con configuraciones contradictorias, es decir, combinaciones causales idénticas que conducen a resultados distintos. Según Cronqvist y Berg-Schlosser (2008), mientras que el csQCA exige excluir o resolver estas líneas contradictorias en la tabla de verdad mediante la introducción de nuevas condiciones, la mvQCA puede atenuarlas gracias a la recalibración multivalorada, permitiendo una distinción más refinada entre los casos.

No obstante, esta estrategia implica también ciertos costos metodológicos relevantes. Uno de los problemas más significativos es la explosión combinatoria: al aumentar el número de variables con múltiples categorías, crece exponencialmente el número de com-

binaciones lógicamente posibles, lo que resulta en un aumento sustancial del número de líneas en la tabla de verdad (Ragin, 1987, 2000). La fórmula general para calcular el número de líneas en la tabla de verdad en mvQCA es:

$$\text{Número de líneas} = \prod_{i=1}^{k} n_i,$$

donde $n_i$ representa el número de categorías de la variable $i$ y $k$ es el número total de condiciones.

Por ejemplo, con dos variables dicotómicas y una tricotómica, se generan $2 \times 2 \times 3 = 12$ líneas en la tabla de verdad. Comparativamente, añadir una categoría a una condición existente genera menos líneas nuevas que añadir una condición binaria adicional, lo que representa una ventaja relativa de la mvQCA frente al csQCA. Sin embargo, esta ventaja es cualificada, ya que el número de combinaciones posibles sigue aumentando rápidamente, agravando el problema de la diversidad limitada, es decir, situaciones donde muchas combinaciones posibles no tienen casos empíricos observados (Rihoux y Ragin, 2009).

Otro aspecto crítico se refiere al proceso de minimización de soluciones. Cuando se trabaja con categorías intermedias (por ejemplo, en una escala de 0 a 2), el estatus teórico de conjunto de estas categorías se vuelve ambiguo, dificultando tanto la interpretación causal como el proceso de minimización lógica (Rihoux y Ragin, 2009). Esto es especialmente problemático en contextos con pocos casos empíricos, ya que se dificulta obtener soluciones empíricamente sostenibles y lógicamente consistentes.

En resumen, aunque la mvQCA ofrece un enfoque más flexible y realista para analizar configuraciones causales complejas —especialmente cuando las variables no pueden ser dicotomizadas sin pérdida significativa de información—, también hereda y amplifica ciertas limitaciones estructurales del QCA, particularmente en lo que respecta a la explosión combinatoria, la diversidad limitada y la ambigüedad interpretativa de las categorías intermedias.

# 5

# ANÁLISIS COMPARATIVO CUALITATIVO CON REGLAS DE ASOCIACIÓN FUZZY

Uno de los principales dificultades dentro del análisis de datos en problemas de ciencias sociales es la gran complejidad de los fenómenos que se estudian. Aunque fsQCA se diseñó precisamente para incorporar cierta flexibilidad gracias al uso de conjuntos difusos, en el fondo, tal y como hemos visto en el Capítulo 4, se acaban generando soluciones binarias, eliminando en los primeros pasos del proceso, toda relación con la lógica borrosa. Esto supone una contradicción con su planteamiento original y limita su utilidad para reflejar los matices graduales que suelen caracterizar a los datos sociales.

Al comparar el enfoque del fsQCA con el de las Reglas de Asociación, hemos visto que éstas últimas ofrecen resultados más valiosos en términos de identificación de patrones y relaciones entre las variables. Además, si en las reglas de asociación cuantitativas se emplean métodos de discretización supervisados (como MDLP), no sólo se añade objetividad y reproducibilidad en el análisis, sino que también se elimina el problema de los datos de máxima ambigüedad (con valores calibrados de 0.5 que no permiten determinar si se verifica o no una determinada condición).

Sin embargo, la naturaleza del fsQCA, pese a lo que su nombre podría indicar, no es completamente "fuzzy". Si bien trata con variables difusas y emplea funciones de pertenencia para cuantificar el grado de pertenencia o de verificación de una condición, uno de los primeros pasos de su metodología es la de-fuzzificación, es decir, volver a binarizar las variables para trabajar como si un problema de csQCA se tratara.

En este capítulo vamos a abordar un tratamiento puramente fuzzy de los problemas de análisis cualitativo comparativo mediante reglas de asociación difusas. Este tipo de reglas de asociación representan una alternativa más adecuada para muchos contextos de análisis social. A lo largo del capítulo se presentarán ejemplos concretos con datos reales que ilustran cómo esta metodología puede ayudar a descubrir patrones significativos incluso cuando las variables no están perfectamente definidas. También se abordan tanto las ventajas como las dificultades que conlleva este enfoque, con la intención de aportar una visión más realista y útil sobre su aplicabilidad en distintos campos de investigación.

## 5.1 REGLAS DE ASOCIACIÓN FUZZY

Como se discutió en la sección anterior, el modelo fsQCA, propuesto por Ragin (2008), no genera soluciones fuzzy, sino soluciones binarias. Esto se debe a las características

del modelo desde el punto de vista algorítmico-matemático, el cual se basa en el uso de tablas de verdad booleanas. En este enfoque, los valores de pertenencia fuzzy se utilizan únicamente para el cálculo de métricas, sin ser considerados en la lógica de la solución final.

En este capítulo, se presenta el modelo de extracción de reglas de asociación fuzzy como una alternativa metodológica para los investigadores en ciencias sociales que emplean el enfoque QCA.

A. Agrawal y Knoeber (1996) propusieron un algoritmo de minería de reglas para variables cuantitativas. Este método consistía en particionar los dominios de las variables, combinando particiones adyacentes mediante discretización, con el objetivo de binarizar los datos. Sin embargo, este enfoque presentó limitaciones importantes, especialmente relacionadas con el problema conocido como *sharp boundary*, que compromete la representación continua de los datos.

Posteriormente, Kuok et al. (1998) introdujeron por primera vez un algoritmo de extracción de reglas fuzzy basado en el algoritmo Apriori. Este algoritmo se compone de dos etapas principales:

1. En la primera etapa, los valores de los atributos cuantitativos se transforman en términos lingüísticos, cada uno asociado a una variable lingüística y a una función de pertenencia (particiones fuzzy). Luego, se selecciona un único término lingüístico con la máxima cardinalidad escalar, con el fin de mantener el número de regiones fuzzy igual al número original de atributos;

2. La segunda etapa consiste en identificar los itemsets frecuentes a partir de un valor mínimo de soporte. Con base en un valor mínimo de confianza, se procede a la extracción de las reglas de asociación fuzzy.

Dado un conjunto de datos $D = \{t_1, t_2, \ldots, t_n\}$ con variable $I$ y conjuntos difusos asociados a dichas variables, el objetivo es identificar relaciones potencialmente útiles e interesantes. La expresión general de una regla de asociación difusa puede representarse como:

$$\text{Si } X = \{x_1, x_2, \ldots, x_i\} \text{ es } A = \{f_1, f_2, \ldots, f_i\} \text{ entonces } Y = \{y_1, y_2, \ldots, y_j\} \text{ es } B = \{g_1, g_2, \ldots, g_j\} \tag{5.1}$$

donde:

- $f_i \in \{\text{conjuntos difusos asociados al atributo } x_i\}$,
- $g_j \in \{\text{conjuntos difusos asociados al atributo } y_j\}$,

y $X$ y $Y$ son itemsets, es decir, subconjuntos de $I$ que son mutuamente disjuntos, lo cual implica que no comparten atributos en común. Los conjuntos $A$ y $B$ contienen los términos difusos correspondientes a los atributos en $X$ y $Y$, respectivamente.

Similarmente a las reglas de asociación binarias, la parte "$X$ es $A$" se denomina el antecedente de la regla, mientras que $Y$ es $B$ es el consecuente. Para que una regla sea considerada interesante, debe presentar un valor de soporte suficiente y un valor elevado de confianza.

### 5.1.1 *Algoritmo de Extración Reglas Fuzzy*

La notación utilizada en esta sección está definida de la siguiente manera:

- $n$: el número total de transacciones de datos;
- $m$: el número total de atributos;
- $A_j$: el $j$-ésimo atributo, $1 \leq j \leq m$;
- $|A_j|$: el número de regiones difusas para $A_j$;
- $R_{jk}$: la $k$-ésima región difusa de $A_j$, $1 \leq k \leq |A_j|$;
- $D^{(i)}$: el $i$-ésimo dato de transacción, $1 \leq i \leq n$;
- $v_j^{(i)}$: el valor cuantitativo de $A_j$ para $D^{(i)}$;
- $f_j^{(i)}$: el conjunto difuso generado a partir de $v_j^{(i)}$;
- $f_{jk}^{(i)}$: el valor de pertenencia de $v_j^{(i)}$ en la región $R_{jk}$;
- $count_{jk}$: la suma de $f_{jk}^{(i)}$ para $i = 1$ hasta $n$;
- $\alpha$: el nivel mínimo de soporte predefinido;
- $\lambda$: el valor mínimo de confianza predefinido;
- $C_r$: el conjunto de *itemsets* candidatos con $r$ atributos (ítems);
- $L_r$: el conjunto de *itemsets* frecuentes con $r$ atributos (ítems).

El algoritmo transforma primero cada valor cuantitativo en un conjunto fuzzy de términos lingüísticos utilizando funciones de pertenencia. Luego, el algoritmo calcula la cardinalidad escalar de cada término lingüístico en todos los datos de transacción. Posteriormente, se lleva a cabo el proceso de minería para encontrar reglas de asociación difusas.

A continuación, se describen las principales etapas del proceso, y la Figura 5.1 presenta el diagrama de flujo del Algoritmo Fuzzy Apriori, propuesto por Hong et al. (2001).
**ENTRADA:** Un conjunto de $n$ datos, cada uno con $m$ atributos, un conjunto de funciones de pertenencia, un valor de soporte mínimo predefinido $\alpha$ y un valor de confianza mínimo predefinido $\lambda$.

**SALIDA:** Un conjunto de reglas de asociación difusas.

PASO 1: Transformar el valor cuantitativo $v_j^{(i)}$ de cada dato de transacción $D^{(i)}$, $i = 1$ hasta $n$, para cada atributo $A_j$, $j = 1$ hasta $m$, en un conjunto difuso $f_j^{(i)}$ representado como:

$$f_j^{(i)} = \left( \frac{f_{j1}^{(i)}}{R_{j1}} + \frac{f_{j2}^{(i)}}{R_{j2}} + \cdots + \frac{f_{jk}^{(i)}}{R_{jk}} \right)$$

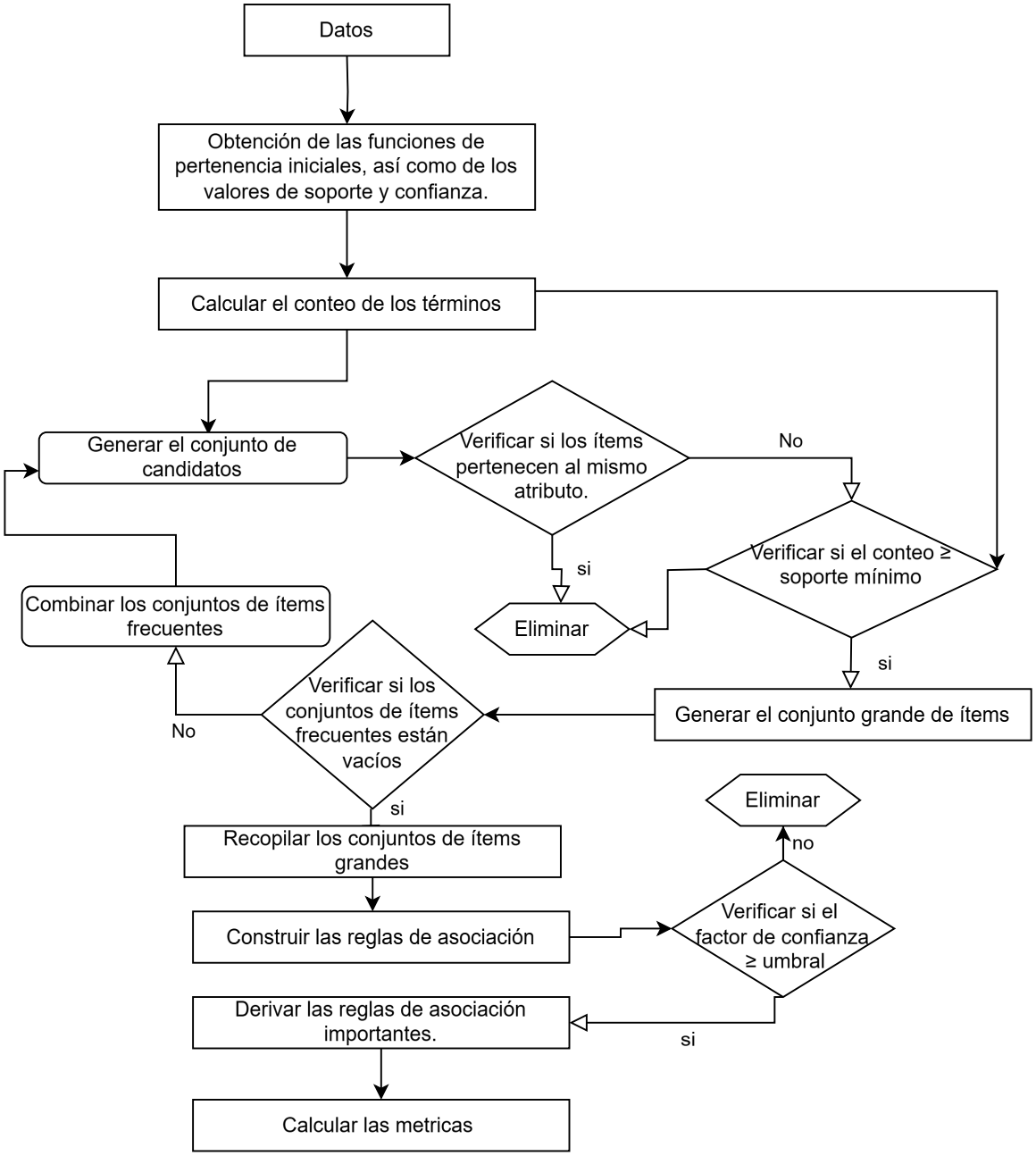


Figura 5.1: Diagrama de flujo del Algoritmo Fuzzy Apriori

utilizando las funciones de pertenencia dadas, donde $R_{jk}$ es el término lingüístico $k$-ésimo del atributo $A_j$, y $f_{jk}^{(i)}$ es el valor de pertenencia difusa de $v_j^{(i)}$ en la región $R_{jk}$, y $k = |A_j|$ es el número de regiones difusas para $A_j$.

PASO 2: Calcular la frecuencia (count) de cada región de atributo (término lingüístico) $R_{jk}$ en los datos de transacción:

$$\text{count}_{jk} = \sum_{i=1}^{n} f_{jk}^{(i)}$$

PASO 3: Recolectar cada región de atributo (término lingüístico) para formar el conjunto candidato $C_1$.

PASO 4: Verificar si $\text{count}_{jk}$ de cada $R_{jk}$ $(1 \leq j \leq m$ y $1 \leq k \leq |A_j|)$ es mayor o igual al valor de soporte mínimo predefinido $\alpha$. Si $R_{jk}$ satisface la condición anterior, incluirlo en el conjunto de itemsets grandes de 1 elemento $(L_1)$. Es decir:

$$L_1 = \left\{ R_{jk} \mid \text{count}_{jk} \geq \alpha,\ 1 \leq j \leq m,\ 1 \leq k \leq |A_j| \right\}$$

PASO 5: Si $L_1$ no está vacío, proceder al siguiente paso; de lo contrario, salir del algoritmo.

PASO 6: Establecer $r = 1$, donde $r$ representa el número de elementos en los itemsets grandes actuales.

PASO 7: Unir los itemsets grandes $L_r$ para generar el conjunto candidato $C_{r+1}$ de forma similar al algoritmo Apriori, donde se exige que dos regiones difusas (términos lingüísticos) que pertenecen al mismo atributo no puedan coexistir en un solo itemset en $C_{r+1}$. Sólo se mantienen los itemsets en $C_{r+1}$ que:

- Tienen todos sus subconjuntos como itemsets existentes en $L_r$.
- No tienen dos términos lingüísticos $R_p$ y $R_q$ $(p \neq q)$ del mismo atributo $A_j$.

PASO 8: Para cada itemset recién formado $s_i = \{s_{i1}, s_{i2}, \ldots, s_{i(r+1)}\} \in C_{r+1}$, hacer:

(a) Calcular el valor difuso de cada transacción $D^{(i)}$ como:

$$f_s^{(i)} = f_{i1}^{(i)} \wedge f_{i2}^{(i)} \wedge \cdots \wedge f_{i(r+1)}^{(i)}$$

Si se usa el operador mínimo para la intersección difusa:

$$f_s^{(i)} = \text{mín}\left( f_{i1}^{(i)}, f_{i2}^{(i)}, \ldots, f_{i(r+1)}^{(i)} \right)$$

(b) Calcular el conteo de $s$ en las transacciones como:

$$\text{count}_s = \sum_{i=1}^{n} f_s^{(i)}$$

(c) Si $\text{count}_s \geq \alpha$, entonces colocar $s$ en $L_{r+1}$.

PASO 9: Si $L_{r+1}$ no es nulo, establecer $r = r + 1$ y repetir los pasos 6 a 8; en caso contrario, proceder al paso siguiente.

PASO 10: Recolectar todos los conjuntos de ítems grandes.

PASO 11: Construir reglas de asociación para cada conjunto de ítems grande $q$ con ítems $\{s_1, s_2, \ldots, s_q\}$, $q \geq 2$, utilizando:

(a) Formar cada posible regla de asociación como:

$$s_1 \wedge \ldots \wedge s_{k-1} \wedge s_{k+1} \wedge \ldots \wedge s_q \rightarrow s_k, \quad k = 1 \text{ hasta } q$$

(b) Calcular los valores de confianza de todas las reglas de asociación usando:

$$\frac{\sum_{i=1}^{n} f_S^{(i)}}{\sum_{i=1}^{n} \left( f_{s_1}^{(i)} \wedge \ldots \wedge f_{s_{k-1}}^{(i)} \wedge f_{s_{k+1}}^{(i)} \wedge \ldots \wedge f_{s_q}^{(i)} \right)}$$

PASO 12: Salida de las reglas de asociación con valores de confianza mayores o iguales al umbral de confianza predefinido $\lambda$.

### 5.1.2 *Ejemplo: Participación electoral en países de la OCDE*

En esta sección, presentamos un ejemplo que ilustra la aplicabilidad de la metodología mediante la generación de reglas fuzzy. Para ello, utilizamos los datos del estudio de Cruz (2023), que empleó el enfoque fsQCA propuesto por Rihoux y Ragin (2009) para analizar cómo la corrupción, la educación, la desigualdad y la confianza en el parlamento afectan la participación electoral (voter turnout) en los países miembros de la Organización para la Cooperación y el Desarrollo Económico (OCDE). El objetivo principal del análisis es identificar si estos factores constituyen condiciones necesarias o suficientes para explicar niveles altos o bajos de participación electoral. El conjunto de datos incluye 29 países, como se muestra en la Tabla 5.1.

Cada caso está compuesto por cinco variables cuantitativas: Participación Electoral (VOTE), medida por el promedio de la tasa de asistencia a elecciones nacionales en el período 2012–2022; Corrupción (NOCORRUPTION), evaluada mediante el Índice de Percepción de la Corrupción (CPI) de Transparency International (2018), en una escala de 0 (muy corrupto) a 100 (muy limpio); Educación (EDUCATION), representada por la tasa bruta de matrícula en educación superior ( %); Desigualdad (INEQUALITY), medida a través del coeficiente de Gini basado en el ingreso disponible; y Confianza en el Parlamento (NOTRUST), expresada por el porcentaje de encuestados que declararon "ninguna confianza" en el parlamento.

Para representar los datos como conjuntos difusos, se utilizaron funciones de pertenencia sigmoidales. Para las condiciones causales se emplearon tres términos lingüísticos, mientras que para la variable de respuesta se utilizaron dos términos lingüísticos. Los puntos de corte de estas funciones se definieron en base a los cuantiles de las distribuciones de las variables, como se muestra en la Figura 5.2.

**Paso 1**: En primer lugar, los valores cuantitativos de cada transacción se transforman en un conjunto fuzzy. Por ejemplo, consideremos la variable NOCORRUPTION (NOC) para el Caso 1, cuya puntuación es 76. Esta puntuación se convierte en un conjunto difuso de la siguiente manera:

$$\text{NOCORRUPTION (NOC)}_{76} = \left( \frac{0.0}{\text{Bajo}} + \frac{0.7}{\text{Medio}} + \frac{0.3}{\text{Alto}} \right)$$

Este procedimiento se repite para los demás países y variables. Los resultados consolidados se presentan en la Tabla 5.2.

Tabla 5.1: Conjunto de datos

| Country | VOTE | NOCORRUPTION | EDUCATION | INEQUALITY | NOTRUST |
|---|---|---|---|---|---|
| Austria | 77.795 | 76 | 86.69 | 27.8 | 8.4 |
| Canada | 64.95 | 81 | 70.11 | 30.3 | 12.9 |
| Chile | 51.17 | 67 | 90.90 | 45.8 | 35.4 |
| Czechia | 63.115 | 59 | 63.77 | 24.4 | 38.5 |
| Denmark | 85.245 | 88 | 81.18 | 26.9 | 10.4 |
| Estonia | 63.95 | 73 | 70.37 | 30.7 | 15.4 |
| Finland | 67.79 | 85 | 90.26 | 26.0 | 10.7 |
| France | 77.455 | 72 | 67.54 | 29.8 | 23.4 |
| Germany | 76.365 | 80 | 70.34 | 29.6 | 14.4 |
| Hungary | 65.755 | 46 | 50.31 | 27.9 | 24.1 |
| Iceland | 80.645 | 76 | 73.10 | 24.6 | 10.7 |
| Ireland | 63.93 | 73 | 77.28 | 29.5 | 11.8 |
| Israel | 69.48 | 61 | 61.48 | 34.4 | 21.9 |
| Italy | 74.06 | 52 | 64.29 | 33.7 | 22.9 |
| Japan | 54.315 | 73 | 63.81 | 32.7 | 11.8 |
| Latvia | 56.69 | 58 | 93.02 | 34.8 | 8.1 |
| Lithuania | 49.22 | 59 | 73.73 | 36.0 | 17.7 |
| Netherlands | 80.32 | 82 | 87.10 | 27.3 | 12.3 |
| New Zealand | 80.995 | 87 | 82.98 | 32.9 | 7.4 |
| Norway | 77.69 | 84 | 83.02 | 25.8 | 3.5 |
| Poland | 56.33 | 60 | 68.62 | 29.2 | 31.4 |
| Portugal | 53.265 | 64 | 65.66 | 32.3 | 21.2 |
| Slovakia | 62.815 | 50 | 45.37 | 23.0 | 19.1 |
| Slovenia | 52.185 | 60 | 77.11 | 24.5 | 26.5 |
| South Korea | 77.14 | 57 | 95.86 | 33.3 | 27.0 |
| Spain | 70.8 | 58 | 91.11 | 32.8 | 23.4 |
| Sweden | 86.495 | 85 | 72.46 | 26.4 | 4.6 |
| United Kingdom | 68.43 | 80 | 61.38 | 31.7 | 16.3 |
| United States | 68.095 | 71 | 88.30 | 38.8 | 23.6 |

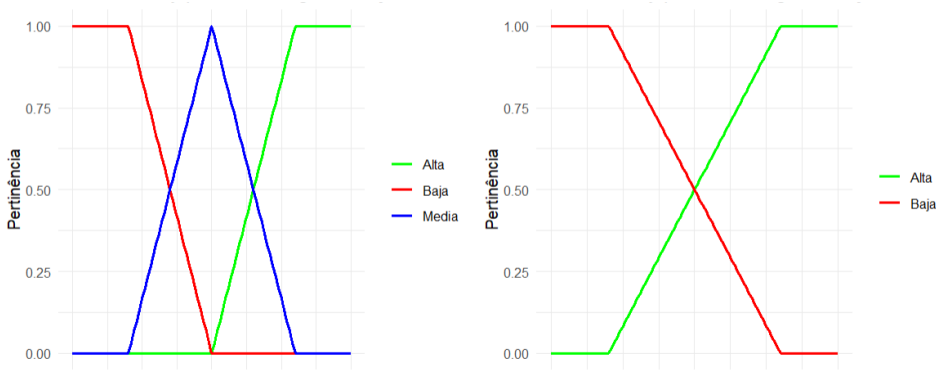


Figura 5.2: Fuzzyficación de las variables utilizando cuartiles.

Tabla 5.2: Variables calibradas lingüísticamente

| País | NOCORRUPTION | | | EDUCATION | | | INEQUALITY | | | NOTRUST | | | VOTE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bajo | Med | Alto | Bajo | Med | Alto | Bajo | Med | Alto | Bajo | Med | Alto | Bajo | Alto |
| Austria | 0,0 | 0,7 | 0,3 | 0,0 | 0,0 | 1,0 | 0,5 | 0,5 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 |
| Canada | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,9 | 0,1 |
| Chile | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 |
| Czechia | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 |
| Denmark | 0,0 | 0,0 | 1,0 | 0,0 | 0,7 | 0,3 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 |
| Estonia | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,9 | 0,1 |
| Finland | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,7 | 0,3 |
| France | 0,0 | 1,0 | 0,0 | 0,6 | 0,4 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 |
| Germany | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,1 | 0,9 |
| Hungary | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,4 | 0,6 | 0,0 | 0,0 | 0,0 | 1,0 | 0,8 | 0,2 |
| Iceland | 0,0 | 0,7 | 0,3 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 |
| Ireland | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,4 | 0,6 | 0,9 | 0,1 |
| Israel | 0,4 | 0,6 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,8 | 0,2 | 0,5 | 0,5 |
| Italy | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,3 | 0,7 | 0,2 | 0,8 |
| Japan | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 0,2 | 0,8 | 0,4 | 0,6 | 0,0 | 1,0 | 0,0 |
| Latvia | 1,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 |
| Lithuania | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 |
| Netherlands | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,8 | 0,2 | 0,0 | 0,2 | 0,8 | 0,0 | 0,0 | 1,0 |
| New Zealand | 0,0 | 0,0 | 1,0 | 0,0 | 0,5 | 0,5 | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 |
| Norway | 0,0 | 0,0 | 1,0 | 0,0 | 0,5 | 0,5 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 |
| Poland | 0,7 | 0,3 | 0,0 | 0,4 | 0,6 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 |
| Portugal | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 0,7 | 0,3 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 |
| Slovakia | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 |
| Slovenia | 0,7 | 0,3 | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 |
| South Korea | 1,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 |
| Spain | 1,0 | 0,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,1 | 0,9 | 0,0 | 0,0 | 1,0 | 0,5 | 0,5 |
| Sweden | 0,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 |
| United Kingdom | 0,0 | 0,0 | 1,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,6 | 0,4 |
| United States | 0,0 | 1,0 | 0,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,0 | 0,0 | 1,0 | 0,6 | 0,4 |
| **Count** | 9,9 | 9,6 | 9,5 | 9,0 | 10,8 | 9,3 | 9,6 | 9,4 | 10,0 | 9,0 | 10,1 | 9,9 | 15,7 | 13,3 |

**Paso 2**: En esta etapa, se calcula la cardinalidad escalar de cada región del atributo (es decir, de cada término lingüístico) a lo largo de todas las transacciones. Este valor se representa como *count*. Por ejemplo, para la región lingüística `NOC.Bajo`, el valor de *count* se calcula sumando los grados de pertenencia de todos los casos a esta categoria:

$$\text{count(NOC.Bajo)} = (0.0 + 0.0 + 0.0 + 1.0 + \ldots + 0.0) = 9.9$$

Este procedimiento se repite para todas las demás regiones linguísticas.

Los resultados de todas regiones se muestran en la fila inferior de la Tabla 5.2.

**Paso 3**: En este paso, cada región de atributo (es decir, cada término lingüístico) se considera como un candidato a 1-itemset. Esto significa que cada combinación del ti-

poVARIABLE.TérminoLingüístico (por ejemplo, NOC.Bajo, EDU.Alto, etc.) es tratada como un posible ítem individual en el proceso de análisis de asociaciones.

**Paso 4**: Para cada región lingüística, se verifica si su valor de *count* es mayor o igual a un umbral de soporte mínimo predefinido, denotado por $\alpha'$. Asumiendo $\alpha' = 2.0$ en este ejemplo, los siguientes ítems cumplen con este criterio y son incluidos en el conjunto de 1-itemsets frecuentes $L_1$ (ver Tabla 5.3):NOC_Bajo, NOC_Med, NOC_Alto, EDU_Bajo, EDU_Med, EDU_Alto, INEQ_Bajo, INEQ_Med, INEQ_Alto, NOTR_Bajo, NOTR_Med, NOTR_Alto, VOT_Bajo.

Es importante destacar que, en nuestro caso, la variable VOTE_Bajo representa la variable de respuesta. Por lo tanto, nuestro interés se centra en identificar si estos factores, ya sea de forma individual o en combinación, constituyen condiciones necesarias o suficientes para niveles bajos de participación electoral.

Tabla 5.3: Conjuntos de 1-itemsets con sus valores.

| **Itemset** | **Count** |
|---|---|
| NOC_Bajo | 9.9 |
| NOC_Med | 9.6 |
| NOC_Alto | 9.5 |
| EDU_Bajo | 9.0 |
| EDU_Med | 10.8 |
| EDU_Alto | 9.3 |
| INEQ_Bajo | 9.6 |
| INEQ_Med | 9.4 |
| INEQ_Alto | 10.0 |
| NOTR_Bajo | 9.0 |
| NOTR_Med | 10.1 |
| NOTR_Alto | 9.9 |
| VOT_Bajo | 15.7 |

**Paso 5**: Dado que el conjunto $L_1$ no está vacío, el proceso continúa hacia los siguientes pasos del algoritmo. Esto significa que existen 1-itemsets frecuentes que cumplen con el umbral mínimo de soporte, lo que permite generar candidatos para conjuntos de ítems de mayor longitud, con el objetivo de descubrir combinaciones relevantes de factores asociados a bajos niveles de participación electoral.

**Paso 6**: Se define $r = 1$, donde $r$ representa el tamaño de los conjuntos de ítems (*itemsets*) que están siendo generados en esta etapa del algoritmo. Este valor inicial indica que estamos trabajando con conjuntos de un solo ítem (1-itemsets). En las próximas iteraciones, este valor se incrementará para permitir la generación de combinaciones más complejas.

**Paso 7**: Se realiza la unión de los elementos del conjunto $L_1$ para generar el conjunto candidato $C_{r+1}$. En este caso, se genera el conjunto $C_2$ (pares de ítems) a partir de $L_1$. Por

ejemplo, algunos elementos de $C_2$ serían:(NOC_Bajo, VOT_Bajo), (NOC_Med, VOT_Bajo), ..., (NOTR_Alto, VOT_Bajo).

**Nota:** Un ítem como (NOC_Bajo, NOC_Med) no se incluye en $C_2$ porque ambos elementos pertenecen a la misma variable (NOCORRUPTION), es decir, a una misma condición causal. Además, todos los pares generados se combinan con `VOT_Bajo`, ya que esta es la variable de interés o variable de respuesta en el análisis.

**Paso 8** Para cada nuevo itemset candidato generado, se realiza el siguiente procedimiento:

(a) Se calcula el valor de pertenencia fuzzy de cada transacción utilizando el operador mínimo (*min*) para representar la intersección de los términos lingüísticos. Por ejemplo, para el itemset `(NOC_Bajo, VOT_Bajo)`, el valor correspondiente al Caso 1 (Austria) es:

$$\text{mín}\{0.0;\ 0.0\} = 0.0$$

Los resultados para los demás casos están presentados en la Tabla 5.4. El mismo procedimiento se aplica a los demás itemsets.

(b) Se suma el valor de pertenencia fuzzy de cada itemset a lo largo de todas las transacciones para obtener su cardinalidad escalar.

(c) Se verifica si estas cardinalidades son mayores o iguales al valor de soporte mínimo predefinido (en este caso, $\alpha' = 2$). Doce itemsets cumplen esta condición y son incluidos en el conjunto $L_2$, como se muestra en la Tabla 5.5. Entre ellos se encuentran: `(NOC_Bajo, VOT_Bajo)`, `(NOC_Med, VOT_Bajo)`, `(EDU_Bajo, VOT_Bajo)`, ..., `(NOTR_Alto, VOT_Bajo)`.

Tabla 5.4: Distribución de pertenencia para `NOC_Bajo`, `VOT_Bajo` y su intersección

| **País** | **NOC_Bajo** | **VOT_Bajo** | **NOC_Bajo ∩ VOT_Bajo** |
|---|---|---|---|
| Austria | 0.0 | 0.0 | 0.0 |
| Canadá | 0.0 | 0.9 | 0.0 |
| Chile | 0.0 | 1.0 | 0.0 |
| Chequia | 1.0 | 1.0 | 1.0 |
| Dinamarca | 0.0 | 0.0 | 0.0 |
| Estonia | 0.0 | 0.9 | 0.0 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| Estados Unidos | 0.0 | 0.6 | 0.0 |
| **count** | | | **7.4** |

Tabla 5.5: Itemsets y fuzzy Conteo de L_2

| Itemset | Count |
|---|---|
| NOC_Bajo, VOT_Bajo | 7.4 |
| NOC_Med, VOT_Bajo | 6.5 |
| NOC_Alto, VOT_Bajo | 2.3 |
| EDU_Bajo, VOT_Bajo | 6.5 |
| EDU_Med, VOT_Bajo | 5.4 |
| EDU_Alto, VOT_Bajo | 3.8 |
| INEQ_Bajo, VOT_Bajo | 4.1 |
| INEQ_Med, VOT_Bajo | 6.0 |
| INEQ_Alto, VOT_Bajo | 5.9 |
| NOTR_Bajo, VOT_Bajo | 2.5 |
| NOTR_Med, VOT_Bajo | 7.4 |
| NOTR_Alto, VOT_Bajo | 6.3 |

**Paso 9:** Si $L_{rsj}$ es nulo, continúe con el siguiente paso. De lo contrario, defina $r = r + 1$ y repita los PASOS 6 a 8. En el ejemplo presentado, como $L_2$ no es nulo, se tiene que $r = r + 1 = 2$, y los PASOS 6 a 8 se ejecutan nuevamente para encontrar $L_3$.

El conjunto candidato $C_3$ se genera a partir de $L_2$ (véase la Tabla 5.5), formando combinaciones de tres ítems que incluyen la variable de respuesta, como: `(EDUC_Med, INEQ_Med, VOT_Bajo)`, `(EDUC_Med, NOTRUST_Med, VOT_Bajo)`, `(EDUC_Med, INEQ_Alto, VOT_Bajo)`, entre otras.

Estos conjuntos representan todas las combinaciones posibles de tres elementos, y como todos tienen frecuencias superiores al umbral mínimo (soporte mayor que 2), se mantienen en $L_3$. Como $L_3$ es un conjunto no vacío, se continúa combinando ítems hasta que se exploren todas las posibilidades que respeten el soporte mínimo. A continuación, se pasa al PASO 10.

**Paso 10:** Se reúnen todos los grandes itemsets (conjuntos frecuentes) identificados a lo largo del proceso.

**Paso 11:** Para cada gran itemset considerado interesante, se construyen reglas de asociación. Luego, se calcula el grado de confianza para cada regla. El umbral de confianza adoptado fue $\lambda = 0.75$.

Dado que el umbral de confianza ($\lambda = 0.75$) fue utilizado como criterio de validez, las reglas de asociación consideradas significativas se presentan en las Tablas 5.6, 5.7 y 5.8.

Estas reglas representan el meta-conocimiento extraído de las transacciones analizadas y son el resultado de la aplicación sistemática del algoritmo de extracción de reglas de asociación.

Tabla 5.6: Reglas con 2 antecedentes que implican `VOT_Bajo`

| # | Antecedentes | Consequente | Suporte | Confiança |
|---|---|---|---|---|
| 1 | EDUC_Med, INEQ_Med | VOT_Bajo | 0.1034 | 0.75 |
| 2 | EDUC_Med, NOTRUST_Med | VOT_Bajo | 0.1034 | 0.75 |
| 3 | EDUC_Med, INEQ_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 4 | EDUC_Med, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 5 | NOC_Bajo, EDUC_Med | VOT_Bajo | 0.0690 | 1.00 |
| 6 | NOC_Med, INEQ_Med | VOT_Bajo | 0.1034 | 0.75 |
| 7 | INEQ_Med, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 8 | NOC_Bajo, INEQ_Med | VOT_Bajo | 0.0690 | 1.00 |
| 9 | NOC_Bajo, NOTRUST_Med | VOT_Bajo | 0.0690 | 1.00 |
| 10 | INEQ_Bajo, NOTRUST_Alto | VOT_Bajo | 0.0690 | 1.00 |
| 11 | NOC_Bajo, INEQ_Bajo | VOT_Bajo | 0.1034 | 1.00 |
| 12 | EDUC_Bajo, INEQ_Bajo | VOT_Bajo | 0.0690 | 1.00 |
| 13 | NOC_Bajo, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 14 | EDUC_Bajo, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 15 | NOC_Bajo, EDUC_Bajo | VOT_Bajo | 0.1379 | 0.80 |

Tabla 5.7: Reglas con 3 antecedentes que implican `VOT_Bajo`

| # | Antecedentes | Consecuente | Soporte | Confianza |
|---|---|---|---|---|
| 1 | NOC_Med, EDUC_Med, INEQ_Med | VOT_Bajo | 0.0690 | 1.00 |
| 2 | EDUC_Med, INEQ_Med, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 3 | NOC_Med, EDUC_Med, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 4 | EDUC_Med, INEQ_Alto, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 5 | NOC_Bajo, EDUC_Med, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 6 | EDUC_Med, INEQ_Bajo, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 7 | NOC_Bajo, EDUC_Med, INEQ_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 8 | NOC_Bajo, EDUC_Med, INEQ_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 9 | NOC_Bajo, EDUC_Med, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 10 | NOC_Med, INEQ_Med, NOTRUST_Med | VOT_Bajo | 0.0690 | 1.00 |
| 11 | NOC_Med, INEQ_Med, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 12 | NOC_Bajo, INEQ_Med, NOTRUST_Alto | VOT_Bajo | 0.0690 | 1.00 |
| 13 | NOC_Bajo, EDUC_Bajo, INEQ_Med | VOT_Bajo | 0.0690 | 1.00 |
| 14 | NOC_Bajo, INEQ_Bajo, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 15 | EDUC_Bajo, INEQ_Bajo, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 16 | NOC_Bajo, INEQ_Alto, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 17 | NOC_Bajo, EDUC_Bajo, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 18 | NOC_Med, INEQ_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 19 | NOC_Med, EDUC_Bajo, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 20 | NOC_Bajo, EDUC_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 21 | NOC_Bajo, INEQ_Bajo, NOTRUST_Alto | VOT_Bajo | 0.0690 | 1.00 |
| 22 | EDUC_Bajo, INEQ_Bajo, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 23 | NOC_Bajo, EDUC_Bajo, INEQ_Bajo | VOT_Bajo | 0.0690 | 1.00 |
| 24 | NOC_Bajo, INEQ_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 25 | EDUC_Bajo, INEQ_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 26 | NOC_Bajo, EDUC_Bajo, NOTRUST_Alto | VOT_Bajo | 0.1034 | 0.75 |

Tabla 5.8: Reglas conm 4 antecedentes que implican `VOT_Bajo`

| # | Antecedentes | Consecuente | Soporte | Confianza |
|---|---|---|---|---|
| 1 | NOC_Bajo, EDUC_Med, INEQ_Bajo, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 2 | NOC_Bajo, EDUC_Bajo, INEQ_Bajo, NOTRUST_Med | VOT_Bajo | 0.0345 | 1.00 |
| 3 | NOC_Bajo, EDUC_Bajo, INEQ_Bajo, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 4 | NOC_Bajo, EDUC_Med, INEQ_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 5 | NOC_Med, EDUC_Bajo, INEQ_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 6 | NOC_Bajo, EDUC_Bajo, INEQ_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 7 | NOC_Bajo, EDUC_Bajo, INEQ_Alto, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 8 | NOC_Bajo, EDUC_Alto, INEQ_Alto, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |
| 9 | NOC_Bajo, EDUC_Bajo, INEQ_Med, NOTRUST_Alto | VOT_Bajo | 0.0345 | 1.00 |
| 10 | NOC_Med, EDUC_Bajo, INEQ_Med, NOTRUST_Bajo | VOT_Bajo | 0.0345 | 1.00 |

### 5.1.3 *Evaluación Crítica de fuzzy-QCA frente a Reglas de Asociación Fuzzy*

En esta sección, analizamos los resultados obtenidos por Cruz, 2023, quien aplica la metodología fsQCA, y comparamos con los resultados obtenidos en la sección anterior, basados en el enfoque fuzzy de Reglas de Asociación. Nuestro esfuerzo sería en vano si los resultados de Cruz, 2023 produjeran exactamente los mismos patrones de solución o clasificaran correctamente el problema abordado en esta sección.

Uno de los principales objetivos del artículo de Cruz, 2023 era identificar las condiciones suficientes para explicar los bajos niveles de participación electoral, utilizando la metodología fsQCA. Los resultados presentados corresponden a las soluciones parsimoniosa, intermedia y compleja.

La Tabla 5.9 muestra las tres soluciones identificadas para la baja participación electoral son:

1. No educación combinada con existencia de desigualdad.
2. Existencia de corrupción combinada con existencia de desigualdad.
3. Existencia de corrupción combinada con no confianza en el parlamento.

Tabla 5.9: Condiciones suficientes para la pertenencia a baja participación electoral en Cruz, 2023

| Trayectorias | Cobertura bruta | Consistencia |
|---|---|---|
| ∼EDUCATION * INEQUALITY | 0.460666 | 0.748848 |
| ∼NOCORRUPTION * INEQUALITY | 0.570517 | 0.757291 |
| ∼NOCORRUPTION * ∼TRUST | 0.656981 | 0.769933 |

El prefijo '*' denota "y".

Al comparar con los resultados de la Tabla 5.6, que presenta soluciones con dos antecedentes utilizando reglas de asociación fuzzy, observamos que la primera solución encontrada por Cruz, 2023 (∼EDUCATION * INEQUALITY) no es propiamente una solución *fuzzy*, sino más bien **booleana**. En cambio, las soluciones 1 y 3 de la Tabla 5.6, a saber: `EDUC_Med * INEQ_Med` y `EDUC_Med * INEQ_Alto`, indican que lo que conduce a una baja participación electoral es la combinación de un nivel **medio de educación** con una **desigualdad moderada o alta**, y **no necesariamente** la combinación de **no educación con existencia de desigualdad**, como sugiere Cruz, 2023.

La segunda solución de Cruz, 2023 (∼NOCORRUPTION * INEQUALITY) correspondería, en el conjunto de soluciones fuzzy, a las combinaciones `NOC_Med * INEQ_Med` y `NOC_Bajo * INEQ_Med`, identificadas como las soluciones 6 y 8 de la Tabla 5.6. En otras palabras, lo que genera la baja participación electoral son combinaciones de **corrupción media o alta** con **desigualdad media**, y no simplemente la ausencia de corrupción combinada con desigualdad.

La tercera solución de Cruz, 2023 (∼NOCORRUPTION * ∼TRUST) encuentra su correspondencia fuzzy en las combinaciones `NOC_Bajo * NOTRUST_Med` y `NOC_Bajo * NOTRUST_Bajo`, que figuran como las soluciones 9 y 13 de la Tabla 5.6. Racionalmente, lo que lleva a una baja participación electoral son combinaciones de **alto nivel de corrupción con confianza parlamentaria moderada** o **alto nivel de corrupción con baja confianza parlamentaria**.

Por lo tanto, podemos concluir que las **reglas de asociación fuzzy** generan soluciones genuinamente **fuzzy**, a diferencia de la metodología fsQCA propuesta por Rihoux y Ragin, 2009, cuyas soluciones son esencialmente **booleanas**. Esto evidencia que el enfoque fuzzy permite un análisis más **granular y continuo** de los fenómenos sociales, captando matices que los métodos *crisp* no logran representar adecuadamente.

## 5.2 MINIMIZACIÓN DE REGLAS FUZZY

En esta sección se propone un enfoque sistemático para la minimización de reglas fuzzy utilizando el método de Quiney-McCluskey (QM). El objetivo principal consiste en reducir la complejidad del conjunto de reglas, eliminando redundancias y sintetizando expresiones lógicas equivalentes que mantienen la semántica original del sistema.

El proceso comienza con la base de reglas fuzzy generada por métodos como el algoritmo Apriori. Se considera, en este contexto, que el universo del discurso de cada variable fuzzy está particionado en tres etiquetas lingüísticas distribuidas uniformemente —por ejemplo: Pequeño, Medio y Grande. La codificación de las reglas transforma los antecedentes en vectores booleanos, permitiendo la aplicación de técnicas clásicas de simplificación lógica.

Cada vector se asocia a una clase, que representa el consecuente de la regla original. La minimización se realiza de forma iterativa para cada clase, aplicando las operaciones del método QM con el fin de generar implicantes primos que representen, de manera concisa, el conjunto original de reglas. Después de la minimización, los vectores codificados se decodifican nuevamente al lenguaje fuzzy, garantizando la interpretabilidad y aplicabilidad de las reglas resultantes.

### 5.2.1 *Codificación Binaria para la Minimización*

La aplicación del método de Quine-McCluskey requiere que las etiquetas lingüísticas sean representadas de forma binaria. Esta etapa de transformación implica la sustitución de los términos fuzzy por combinaciones de bits que preserven el orden y la proximidad semántica entre las etiquetas. La representación binaria debe construirse cuidadosamente para garantizar que las etiquetas adyacentes presenten una distancia de Hamming igual a 1, lo cual permite su combinación lógica durante el proceso de minimización.

Sea $L$ el número de etiquetas lingüísticas disponibles para una variable. El número mínimo de bits necesarios para representarlas está dado por:

$$\text{Bits} = \lfloor \log_2(L) \rfloor + 1 \tag{5.2}$$

Por ejemplo, para tres etiquetas (Pequeño, Medio, Grande), se necesitan dos bits. Una codificación posible y eficaz sería:

$$\text{Pequeño} \to 00, \quad \text{Medio} \to 01, \quad \text{Grande} \to 11$$

Esta codificación garantiza la posibilidad de combinación lógica entre etiquetas vecinas, como 00 y 01, o 01 y 11, cumpliendo con los criterios necesarios para operar con el método QM.

### 5.2.2 *Aplicación del Método de Quine-McCluskey*

Una vez codificada la base de reglas fuzzy, el proceso de minimización se lleva a cabo mediante las etapas habituales del método Quine-McCluskey. Primero, las reglas se organizan según sus clases de consecuente. Para cada clase, se construyen tablas de verdad a partir de los vectores binarios que representan los antecedentes.

Consideremos el siguiente ejemplo, donde dos reglas presentan el mismo consecuente $C_1$:

$$R_1 : \text{SI } (X_1 \text{ es Med}) \wedge (X_2 \text{ es Peq}) \wedge (X_3 \text{ es Med}) \wedge (X_4 \text{ es Peq}) \Rightarrow C_1$$
$$R_2 : \text{SI } (X_1 \text{ es Med}) \wedge (X_2 \text{ es Peq}) \wedge (X_3 \text{ es Med}) \wedge (X_4 \text{ es Med}) \Rightarrow C_1$$

Aplicando la codificación binaria, se obtienen vectores de 8 bits que representan los antecedentes. La Tabla 5.10 ilustra esta representación para los dos casos.

Tabla 5.10: Representación binaria de los antecedentes de las reglas $R_1$ y $R_2$

| Regla | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | Clase |
|---|---|---|---|---|---|---|---|---|---|
| $R_1$ | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | $C_1$ |
| $R_2$ | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | $C_1$ |

Como los vectores difieren solo en el octavo bit, la aplicación del método Quine-McCluskey permite la combinación de los dos términos en un único implicante primo, representado por:

$$(0\,1\,0\,0\,0\,1\,0\,*) \Rightarrow C_1$$

Este nuevo término representa una regla minimizada:

SI $X_1$ es Medio Y $X_2$ es Pequeño Y $X_3$ es Medio Y $X_4$ es Pequeño O Medio ENTONCES $C_1$

Este procedimiento reduce el número de reglas sin pérdida de información, promoviendo una representación más compacta de la base de conocimiento.

### 5.2.3 *Casos Típicos de Minimización*

Durante la aplicación del método de Quine-McCluskey, se observan distintos patrones que indican diferentes niveles de generalización. Por ejemplo, cuando se obtiene un patrón binario como $(0\ *)$ para una variable, esto indica que puede asumir dos valores adyacentes, como Pequeño o Medio. En casos extremos, términos como $(*\ 1)$ pueden indicar generalizaciones hacia las etiquetas superiores del dominio, como Medio o Grande.

Estos patrones surgen como resultado de la simplificación de vectores cuya única diferencia ocurre en una posición específica, y son fundamentales para garantizar la consistencia semántica de las reglas minimizadas.

La aplicación del método de Quine-McCluskey a la minimización de reglas fuzzy codificadas binariamente se muestra eficaz para reducir la complejidad del sistema de inferencia. La preservación de la estructura semántica original se garantiza mediante la codificación adecuada de las etiquetas lingüísticas y la decodificación posterior de los implicantes primos. Este enfoque permite integrar métodos clásicos de simplificación lógica con sistemas fuzzy, contribuyendo al desarrollo de modelos más compactos, eficientes e interpretables.

# 6

# CLASIFICACIÓN DE REGLAS DE ASOCIACIÓN MEDIANTE MÉTODOS MULTICRITERIO NO PONDERADOS

Como hemos visto en los capítulos anteriores, el proceso de obtención de reglas de asociación presenta una explosión combinatoria, es decir, el número de reglas interesantes se incrementa de forma exponencial con el número de variables. Por ello, uno de los retos más importantes es decidir cuáles de ellas son realmente útiles y relevantes. Aunque en capítulos anteriores ya se ha hablado del uso de métricas como el soporte y la confianza, está claro que, por sí solas, estas medidas no bastan para hacer una selección verdaderamente eficaz de las reglas que pueden aportar valor en la toma de decisiones.

Partiendo de esta necesidad, en este capítulo se propone una alternativa basada en técnicas de Análisis Multicriterio (Multiple-Criteria Decision Analysis (MCDA)) para abordar la complejidad de priorizar las reglas obtenidas. En concreto, se introduce el uso del método Unweighted TOPSIS (uwTOPSIS), una versión del TOPSIS tradicional que se diferencia por no exigir la asignación de pesos a los distintos criterios. Esta particularidad ayuda a reducir la subjetividad del proceso y aporta mayor claridad y estabilidad a los resultados obtenidos.

La propuesta proporciona una manera clara de ordenar la multitud de reglas considerando al mismo tiempo múltiples criterios relevantes. Para ilustrar su aplicabilidad, la metodología se ilustra con un caso de estudio con datos reales relacionados con el COVID-19 en el estado de Espírito Santo (Brasil), mostrando cómo el método uwTOPSIS puede contribuir a tomar decisiones más sólidas y fundamentadas.

## 6.1 INTRODUCCIÓN

El crecimiento exponencial de los datos disponibles y el desarrollo del poder computacional han contribuido significativamente al avance de técnicas de minería de datos orientadas a apoyar la toma de decisiones (Bedhouche et al., 2023). Estas técnicas se han vuelto aplicables a prácticamente cualquier contexto en el que se generen datos, convirtiéndose en un componente clave para extraer conocimiento útil a partir de grandes volúmenes de información.

Uno de los problemas más importantes en la minería de reglas de asociación es la explosión combinatoria, que genera una cantidad masiva de reglas, muchas de las cuales son redundantes, triviales o poco útiles (Ashrafi et al., 2005). Varios autores han propuesto

mecanismos para tratar de paliar este problema, como por ejemplo Aggarwal y Yu (1998), Ashrafi et al. (2004), B. Liu et al. (2000) y M. J. Zaki (2000). En particular, Aggarwal y Yu (1998) propusieron una técnica para eliminar reglas redundantes derivadas del mismo conjunto frecuente, considerando una regla como redundante si presenta el mismo soporte que otra pero menor confianza. No obstante, Ashrafi et al. (2004) evidenciaron limitaciones en este tipo de enfoque, al demostrar que puede conducir a la exclusión de reglas potencialmente relevantes en determinados casos.

A pesar de la variedad de enfoques propuestos, aún se observa escasa atención en la literatura sobre la selección y clasificación de reglas de asociación utilizando criterios múltiples de decisión. Muchos algoritmos dependen exclusivamente de umbrales arbitrarios de soporte y confianza para reducir el número de reglas, lo que puede resultar en la pérdida de reglas de alto valor semántico (M.-C. Chen, 2007; Choi et al., 2005). Es habitual que, incluso después de aplicar tales filtros, el número de reglas extraídas siga siendo elevado, dificultando su análisis e interpretación por parte de los agentes decisores (Z. Zhang y Guo, 2009).

La literatura distingue entre medidas objetivas y subjetivas para evaluar la calidad de una regla (Geng y Hamilton, 2006; Z. Zhang y Guo, 2009). Las primeras se fundamentan en teorías estadísticas o probabilísticas, como el soporte, la confianza o el *lift*; mientras que las segundas requieren la interacción directa del usuario, quien aporta conocimiento experto o preferencias contextuales. En función de esta distinción, diversos estudios han sugerido la necesidad de integrar medidas objetivas y subjetivas a través de métodos multicriterio para una evaluación más global (Z. Zhang y Guo, 2009).

De este modo, se torna relevante el posprocesamiento de las reglas, permitiendo clasificarlas o priorizarlas mediante técnicas de toma de decisiones multicriterio (MCDA). Según Roy (1996), las decisiones reales rara vez se basan en un único criterio, lo cual justifica la utilidad de herramientas MCDA en escenarios complejos con criterios en conflicto.

Aunque algunos trabajos han aplicado enfoques multicriterio a la minería de reglas de asociación, su implementación aún es incipiente. Por ejemplo, Choi et al. (2005) combinaron los métodos ELECTRE II y AHP para clasificar reglas tomando en cuenta tanto medidas objetivas como la experiencia del analista. No obstante, Z. Zhang y Guo (2009) argumentaron que esta estrategia es inviable cuando se manejan grandes volúmenes de reglas, dada la carga cognitiva y el tiempo requerido para asignar pesos precisos a cada criterio.

Otros estudios han empleado técnicas del Análisis Envolvente de Datos (DEA) para identificar reglas eficientes (M.-C. Chen, 2007). Aunque el DEA permite considerar simultáneamente múltiples medidas, su aplicación exige resolver un modelo de programación lineal por cada regla, lo que representa una carga computacional significativa. Algunas mejoras como los modelos DEA integrados permiten calcular una única frontera eficiente, pero aún presentan dificultades para establecer un orden completo de reglas y para incorporar juicios subjetivos sobre la importancia relativa de los criterios(M.-C. Chen, 2007; Toloo et al., 2009).

Recientemente, Yacoubi et al. (2022) propusieron una estrategia basada en TOPSIS multiobjetivo, validada con funciones de prueba WFG y datos reales del sector industrial. Aunque eficaz, esta propuesta presenta un costo computacional elevado, lo que limita su aplicabilidad práctica en contextos con recursos limitados.

Por otra parte, algunos trabajos han empleado variantes ponderadas del método TOPSIS para ordenar reglas de asociación (Ermatita y Febry, 2017; Z. Zhang y Guo, 2009), pero este enfoque depende de la asignación precisa de pesos a cada criterio, lo cual introduce subjetividad y puede generar inconsistencias si los agentes decisores carecen de información suficiente para realizar esa asignación (Jacquet-Lagreze y Siskos, 1982; Liern y Pérez-Gladish, 2022). En este contexto, surge el método **Unweighted TOPSIS** (uwTOPSIS) como alternativa que prescinde de la necesidad de ponderaciones explícitas, permitiendo establecer límites flexibles y aproximados sobre las medidas consideradas (Benítez y Liern, 2021; Liern y Pérez-Gladish, 2022). Este enfoque ofrece una solución intermedia entre los métodos de ponderación subjetiva y los completamente objetivos.

El método uwTOPSIS ha ganado notoriedad recientemente por su aplicabilidad en contextos donde no es viable determinar pesos con precisión. Su eficacia ha sido demostrada en áreas tan diversas como el desarrollo sostenible (Blanca Pérez-Gladish y Zopounidis, 2021; Omar Elfarouk y Wong, 2022; Saunders y Luukkanen, 2022), la educación (M. C. Bas et al., 2024), la agricultura digital (Ankita Das et al., 2024), el turismo, la ingeniería, la gestión urbana, la química y el análisis de calidad del agua (Camarillo-Cisneros et al., 2024; Fernández-Portillo et al., 2024; Müller et al., 2024; Zende y Pawade, 2024; Zientara et al., 2024). No obstante, hasta la fecha, no se han encontrado evidencias en la literatura que reporten la aplicación de uwTOPSIS en el contexto específico de la extracción y priorización de reglas de asociación. En este sentido, el presente trabajo se posiciona como el primero en proponer e implementar este enfoque innovador, abriendo una nueva línea de investigación en el cruce entre métodos MCDA y técnicas de minería de datos.

En este capítulo, proponemos una estrategia con el enfoque MCDA propuesto por Benítez y Liern (2021) y Liern y Pérez-Gladish (2022). Esta integración tiene como objetivo proporcionar una base metodológica sólida para clasificar reglas de asociación extraídas, promoviendo una toma de decisiones más robusta, eficiente y coherente en contextos de alta complejidad.

## 6.2 EL ALGORITMO TOPSIS NO PONDERADO

Consideremos $n$ alternativas $A_i$, $i = 1, \ldots, n$, cuyos desempeños se evalúan en $m$ criterios $C_j$, $j = 1, \ldots, m$. Cada criterio puede ser de tipo beneficio o de tipo costo. El objetivo es agregar los desempeños de las alternativas de manera compensatoria, de modo que se pueda obtener una evaluación global de cada $A_i$ y, por lo tanto, establecer una clasificación general de las alternativas (Benítez y Liern, 2021).

Existen diversos métodos de ayuda a la toma de decisiones multicriterio que permiten alcanzar dicho objetivo. Por ejemplo, Z. Zhang y Guo (2009) aplica el algoritmo TOPSIS

clásico propuesto por Deng et al. (2000) y Hwang y Yoon (1981) para ordenar reglas de asociación, descrito a continuación:

**Algoritmo 2.1:** TOPSIS Clásico

**Entradas:**

(1) $A_1, A_2, \ldots, A_n$: alternativas evaluadas en $m$ criterios $[x_{ij}]$.

(2) Importancia relativa (peso) de cada criterio, $w_j$, $1 \leq j \leq m$.

**Paso 1:** Matriz normalizada:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}, \quad r_{ij} \in [0,1], \quad 1 \leq i \leq n, \quad 1 \leq j \leq m. \tag{6.1}$$

**Paso 2:** Matriz ponderada: $[w_j r_{ij}]$.

**Paso 3:** Determinación de las soluciones ideal positiva $A^+$ y negativa $A^-$. Estas se obtienen eligiendo, entre todas las alternativas, los valores más favorables (mejores) y menos favorables (peores), respectivamente, para cada criterio.

**Paso 4:** Proximidad relativa: Se calcula

$$R_i = \frac{d(A_i, A^-)}{d(A_i, A^-) + d(A_i, A^+)}, \quad R_i \in [0,1], \quad 1 \leq i \leq n, \tag{6.2}$$

donde $d$ es la distancia euclidiana.

**Salida:** Ranking de las alternativas según los valores de $R_i$. Cuanto más cercano esté $R_i$ a 1, mayor será la similitud de la alternativa $A_i$ con la solución ideal y más lejos estará de la solución antiideal.

El método TOPSIS clásico, como se describió anteriormente, presenta una limitación relacionada con la asignación de los pesos de los criterios, los cuales suelen determinarse de forma subjetiva, según las preferencias de los tomadores de decisiones (Paso 2). Esta subjetividad puede afectar la confiabilidad y la validez de los resultados. Por ello, se han desarrollado diversos métodos para establecer los pesos, tanto de forma subjetiva como objetiva (Zeleny, 2011).

Siguiendo a Liern y Pérez-Gladish, 2022, proponemos el método uwTOPSIS como una alternativa que evita la asignación arbitraria de pesos, considerando estos como variables de decisión en un problema de optimización. El algoritmo uwTOPSIS se describe a continuación:

**Algoritmo 2.2:** uwTOPSIS

**Paso 1:** Defina la matriz de decisión $X = [x_{ij}]_{n \times m}$, donde el elemento $x_{ij}$ representa el desempeño de la alternativa $A_i$ respecto al criterio $C_j$.

**Paso 2:** Construya la matriz normalizada como:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}, \quad r_{ij} \in [0,1], \quad 1 \leq i \leq n, \quad 1 \leq j \leq m. \tag{6.3}$$

**Paso 3:** Determine la solución ideal positiva ($A^+$) y la solución ideal negativa ($A^-$):

$$A^+ = (A_1^+, \ldots, A_m^+), \quad A^- = (A_1^-, \ldots, A_m^-),$$

donde:

$$A_j^+ = \begin{cases} \text{máx}_{1\le i\le n}\, r_{ij}, & j \in J^+, \\ \text{mín}_{1\le i\le n}\, r_{ij}, & j \in J^-, \end{cases} \quad A_j^- = \begin{cases} \text{mín}_{1\le i\le n}\, r_{ij}, & j \in J^+, \\ \text{máx}_{1\le i\le n}\, r_{ij}, & j \in J^-, \end{cases} \tag{6.4}$$

siendo, $J^+$ y $J^-$ los conjuntos de índices correspondientes a los criterios de beneficio y de costo, respectivamente.

**Paso 4:** Sea $\Omega$ el conjunto:

$$\Omega = \left\{ w = (w_1, \ldots, w_m) \in \mathbb{R}^m \,\middle|\, 0 \le w_j \le 1, \sum_{j=1}^{m} w_j = 1 \right\}. \tag{6.5}$$

Defina las funciones $d_i^+ : \Omega \to [0,1]$ y $d_i^- : \Omega \to [0,1]$, $1 \le i \le n$, como:

$$d_i^+(w) = \sqrt{\sum_{j=1}^{m} \left( w_j r_{ij} - w_j A_j^+ \right)^2}, \quad d_i^-(w) = \sqrt{\sum_{j=1}^{m} \left( w_j r_{ij} - w_j A_j^- \right)^2}. \tag{6.6}$$

**Paso 5:** Para cada alternativa $A_i$, considere la función de proximidad relativa a la solución ideal:

$$R_i(w) = \frac{d_i^-(w)}{d_i^-(w) + d_i^+(w)}, \quad 1 \le i \le n. \tag{6.7}$$

**Paso 6:** Calcule los valores extremos de las funciones $R_i$ ($1 \le i \le n$):

$$R_i^- = \text{mín}\, R_i(w), \quad R_i^+ = \text{máx}\, R_i(w), \quad \text{sujeto a: } \sum_{j=1}^{m} w_j = 1, \quad l_j \le w_j \le u_j,\ 1 \le j \le m, \tag{6.8}$$

donde $l_j, u_j \in [0,1]$ son los límites inferior y superior para los pesos, respectivamente.

**Paso 7:** Construya los intervalos:

$$RI_i = \left[ R_i^-, R_i^+ \right], \quad 1 \le i \le n. \tag{6.9}$$

**Paso 8:** Clasifique las alternativas de acuerdo con los intervalos $RI_i$, $1 \le i \le n$.

Para comparar dos intervalos $A = [a_1, a_2]$ y $B = [b_1, b_2]$, se utiliza el criterio propuesto por Canós y Liern (2008): El intervalo $A$ es mayor que $B$ si:

$$A \succ B \iff \begin{cases} k_1 a_1 + k_2 a_2 > k_1 b_1 + k_2 b_2, & k_1 a_1 + k_2 a_2 \ne k_1 b_1 + k_2 b_2, \\ a_1 > b_1, & k_1 a_1 + k_2 a_2 = k_1 b_1 + k_2 b_2. \end{cases} \tag{6.10}$$

Aquí, $k_1$ y $k_2$ son constantes positivas predefinidas.

**Definición 6.1.** Dadas las alternativas $\{A_i\}_{i=1}^n$ y el conjunto de intervalos $RI_i$ definido en (6.9), se dice que la alternativa $A_i$ es preferida a la alternativa $A_k$ siempre que $RI_i \ge RI_k$.

Se denomina indicador uwTOPSIS ($R^{uw}$) al valor calculado como:

$$R_i^{uw} = k_1 R_i^- + k_2 R_i^+, \quad 1 \le i \le n. \tag{6.11}$$

En este trabajo, para simplificar, se considera $k_1 = k_2 = 0.5$ en la ecuación (6.11), utilizando los puntos medios de los intervalos $[R_i^-, R_i^+]$ como indicador uwTOPSIS.

## 6.3 METODOLOGÍA PROPUESTA

Tradicionalmente, el proceso de minería de datos se realiza considerando principalmente los umbrales de soporte y confianza para filtrar las reglas de asociación (R. Agrawal et al., 1993). Sin embargo, esto suele conducir a una explosión en la cantidad de reglas generadas, muchas de las cuales resultan redundantes. Aunque Z. Zhang y Guo (2009) propusieron el uso del método TOPSIS clásico para clasificar reglas interesantes, nuestra propuesta (uwTOPSIS) se aplica después de la extracción de las reglas, eliminando la necesidad de establecer pesos fijos, como ocurre en Z. Zhang y Guo (2009).

El interés de una regla de asociación depende fundamentalmente del contexto de aplicación (Srikant y R. Agrawal, 1997), y el conocimiento del dominio en áreas específicas puede proporcionar criterios valiosos para seleccionar reglas relevantes e importantes. Esta información puede aprovecharse para mejorar el proceso de elección de reglas (M.-C. Chen, 2007).

En esta sección, proponemos una nueva metodología para minimizar y clasificar reglas de asociación, considerando métricas como confianza, soporte, redundancia y conocimientos del dominio, sin requerir la definición explícita de pesos para los criterios.

El método propuesto puede ilustrarse mediante el diagrama de flujo de la Figura 6.1 y se describe a continuación:

**Etapa 1:** Se proporciona el conjunto de datos para la minería de reglas de asociación.

**Etapa 2:** Se generan las reglas de asociación utilizando el algoritmo Apriori (u otro), a partir de los umbrales mínimos de soporte y confianza definidos por el usuario.

**Etapa 3:** Los usuarios seleccionan los criterios que desean considerar (como soporte, confianza, lift, entre otros), sin necesidad de asignar pesos específicos a cada uno.

**Etapa 4:** Se aplica el método *uwTOPSIS* para clasificar las reglas de asociación generadas y minimizadas.

**Etapa 5:** Se seleccionan las reglas mejor posicionadas en el ranking para su análisis e implementación.

### 6.3.1 *COVID-19 en el Estado de Espírito Santo*

En esta sección se ilustra la metodología uwTOPSIS mediante un ejemplo práctico que destaca la relevancia del enfoque propuesto: la priorización de reglas de asociación derivadas de pacientes infectados por COVID-19 en el estado de Espírito Santo, Brasil.

La COVID-19 es una enfermedad viral perteneciente a la familia del coronavirus (SARS-CoV-2), caracterizada por su alta tasa de contagio entre seres humanos. Fue identificada a

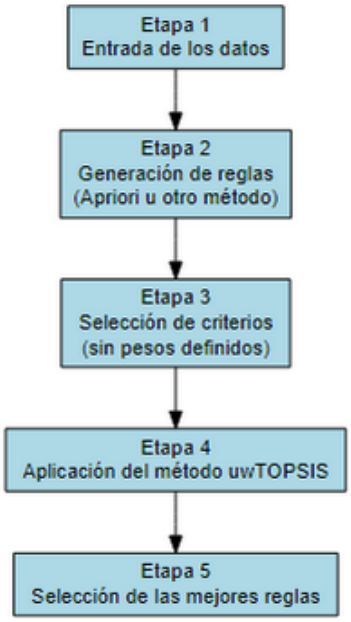


Figura 6.1: Diagrama de flujo de la metodología propuesta

finales de 2019 en la provincia china de Wuhan y, posteriormente, declarada pandemia global. La enfermedad puede manifestarse con síntomas leves o moderados, semejantes a los de una gripe común. Sin embargo, en individuos con comorbilidades o condiciones preexistentes, puede evolucionar hacia cuadros graves como el síndrome respiratorio agudo severo, complicaciones sistémicas e incluso la muerte.

La transmisión del virus se produce, principalmente, por el aire o por contacto con secreciones contaminadas, tales como gotas de saliva, estornudos, tos, mucosidad o contacto con superficies u objetos contaminados, seguido del contacto con ojos, boca o nariz.

El estudio del SARS-CoV-2 representa un componente esencial en las investigaciones científicas y de salud pública, ya que su mejor comprensión facilita la formulación de estrategias eficaces para mitigar sus efectos tanto sanitarios como socioeconómicos. En este contexto, la identificación y minería de reglas de asociación en datos de pacientes diagnosticados con COVID-19 contribuye significativamente al diseño de respuestas sanitarias basadas en evidencia empírica (Kissler et al., 2020).

Asimismo, el análisis de patrones de transmisión y factores de riesgo asociados permite a los gobiernos diseñar políticas públicas fundamentadas científicamente, optimizando medidas como el distanciamiento social, las campañas de vacunación y la implementación de restricciones específicas en actividades económicas (Kissler et al., 2020).

El método uwTOPSIS se propone como una herramienta eficaz para la priorización de reglas de asociación, facilitando la identificación de las más relevantes en el contexto del control epidemiológico. Esta metodología permite ordenar las reglas según su cercanía a una solución ideal, contribuyendo a una mejor interpretación de los patrones subyacentes y a la formulación de estrategias de intervención más precisas (Benítez y Liern, 2021).

Para la extracción de reglas de asociación se utilizaron datos disponibles en el portal oficial del gobierno del estado de Espírito Santo[1], en formato `csv`. Inicialmente, la base de datos contenía 1 048 142 registros. Tras el proceso de depuración y limpieza de microdatos, se obtuvieron 269 100 observaciones válidas. Las variables seleccionadas para la minería de reglas incluyeron: edad, fiebre, dificultad respiratoria, tos, dolor de garganta, cefalea, comorbilidades y el resultado del test de COVID-19 (positivo o negativo). Cabe resaltar que la variable correspondiente al diagnóstico de COVID-19 fue utilizada como variable de salida (outcome).

Aplicado el algoritmo de minería descrito en secciones anteriores, se definieron los umbrales de soporte mínimo en 0.0001 y confianza mínima en 0.85, resultando en un total de 288 reglas de asociación extraídas.

Posteriormente, las reglas fueron progresivamente minimizadas. La Tabla 6.1 presenta las 49 reglas resultantes tras este proceso. La etapa de minimización eliminó el 82.99 % de las reglas redundantes, lo que equivale a 239 reglas descartadas de las 288 iniciales.

### 6.3.2 *Aplicación de uwTOPSIS*

Para clasificar el conjunto de 49 reglas asociadas a la COVID-19 en el estado de Espírito Santo (Brasil), en función de cinco criterios, se empleó el método uwTOPSIS.

Tabla 6.1: Association rule solutions

| Rule | Support | Confidence | Coverage | Lift | Length | Rule | Support | Confidence | Coverage | Lift | Length |
|---|---|---|---|---|---|---|---|---|---|---|---|
| R1 | 0.2092 | 0.8636 | 0.2422 | 1.0864 | 2 | R26 | 0.0677 | 0.8747 | 0.0774 | 1.1004 | 4 |
| R2 | 0.2425 | 0.8576 | 0.2828 | 1.0789 | 2 | R27 | 0.0910 | 0.8767 | 0.1038 | 1.1029 | 4 |
| R3 | 0.0361 | 0.8610 | 0.0420 | 1.0832 | 3 | R28 | 0.0395 | 0.8690 | 0.0454 | 1.0932 | 4 |
| R4 | 0.0098 | 0.8587 | 0.0114 | 1.0802 | 3 | R29 | 0.0630 | 0.8711 | 0.0723 | 1.0959 | 4 |
| R5 | 0.0371 | 0.8599 | 0.0431 | 1.0817 | 3 | R30 | 0.0881 | 0.8651 | 0.1019 | 1.0884 | 4 |
| R6 | 0.0744 | 0.8511 | 0.0875 | 1.0707 | 3 | R31 | 0.1061 | 0.8685 | 0.1222 | 1.0926 | 4 |
| R7 | 0.1010 | 0.8735 | 0.1156 | 1.0989 | 3 | R32 | 0.0748 | 0.8668 | 0.0863 | 1.0904 | 4 |
| R8 | 0.1276 | 0.8663 | 0.1473 | 1.0899 | 3 | R33 | 0.1593 | 0.8674 | 0.1836 | 1.0912 | 4 |
| R9 | 0.1421 | 0.8618 | 0.1649 | 1.0842 | 3 | R34 | 0.0417 | 0.8614 | 0.0484 | 1.0836 | 4 |
| R10 | 0.1904 | 0.8674 | 0.2195 | 1.0912 | 3 | R35 | 0.1230 | 0.8582 | 0.1434 | 1.0796 | 4 |
| R11 | 0.1464 | 0.8584 | 0.1705 | 1.0799 | 3 | R36 | 0.0911 | 0.8606 | 0.1058 | 1.0827 | 4 |
| R12 | 0.2250 | 0.8597 | 0.2617 | 1.0815 | 3 | R37 | 0.0020 | 0.8542 | 0.0023 | 1.0746 | 5 |
| R13 | 0.0088 | 0.8690 | 0.0102 | 1.0932 | 4 | R38 | 0.0022 | 0.8588 | 0.0025 | 1.0804 | 5 |
| R14 | 0.0087 | 0.8578 | 0.0102 | 1.0791 | 4 | R39 | 0.0049 | 0.8534 | 0.0058 | 1.0735 | 5 |
| R15 | 0.0147 | 0.8686 | 0.0169 | 1.0928 | 4 | R40 | 0.0037 | 0.8705 | 0.0042 | 1.0951 | 5 |
| R16 | 0.0040 | 0.8707 | 0.0046 | 1.0954 | 4 | R41 | 0.0083 | 0.8808 | 0.0095 | 1.1081 | 5 |
| R17 | 0.0058 | 0.8571 | 0.0067 | 1.0783 | 4 | R42 | 0.0116 | 0.8589 | 0.0135 | 1.0805 | 5 |
| R18 | 0.0090 | 0.8761 | 0.0102 | 1.1022 | 4 | R43 | 0.0204 | 0.8613 | 0.0237 | 1.0835 | 5 |
| R19 | 0.0091 | 0.8656 | 0.0105 | 1.0890 | 4 | R44 | 0.0102 | 0.8559 | 0.0119 | 1.0768 | 5 |
| R20 | 0.0234 | 0.8598 | 0.0272 | 1.0817 | 4 | R45 | 0.0239 | 0.8588 | 0.0278 | 1.0804 | 5 |
| R21 | 0.0133 | 0.8510 | 0.0157 | 1.0705 | 4 | R46 | 0.0245 | 0.8848 | 0.0277 | 1.1131 | 5 |
| R22 | 0.0330 | 0.8630 | 0.0382 | 1.0856 | 4 | R47 | 0.0559 | 0.8777 | 0.0637 | 1.1041 | 5 |
| R23 | 0.0111 | 0.8529 | 0.0130 | 1.0730 | 4 | R48 | 0.0013 | 0.8592 | 0.0016 | 1.0809 | 6 |
| R24 | 0.0526 | 0.8532 | 0.0616 | 1.0734 | 4 | R49 | 0.0501 | 0.8593 | 0.0583 | 1.0810 | 6 |
| R25 | 0.0281 | 0.8850 | 0.0317 | 1.1134 | 4 | | | | | | |

1 https://coronavirus.es.gov.br/painel-covid-19-es

La Tabla 6.1 presenta las puntuaciones obtenidas por las 49 reglas minimizadas mediante el algoritmo propuesto, evaluadas según cinco criterios. Los cuatro primeros criterios, soporte, confianza, cobertura y lift, son considerados beneficios (criterios de maximización). El quinto criterio, por otro lado, es interpretado como un criterio de coste, ya que representa el número de variables en cada regla: a mayor número de variables, mayor complejidad y costo de implementación en el sistema de salud, debido al uso intensivo de recursos. Por lo tanto, este último es un criterio de minimización.

Aplicamos el método uwTOPSIS, descrito en la Sección 6.2, sobre los datos de la Tabla 6.1. Como se ha destacado previamente, una de las principales ventajas del método uwTOPSIS radica en su carácter completamente basado en datos: no requiere evaluaciones subjetivas de expertos ni asignaciones arbitrarias de calificaciones lingüísticas a números difusos triangulares (Benítez y Liern, 2021). El único punto crítico del método es la definición de los límites inferior y superior para los pesos de los criterios, representados por $l_j$ y $u_j$, respectivamente. En este estudio, se adoptaron los siguientes límites uniformes para todos los criterios:

$$l_j = 0.05; \quad u_j = 0.35; \quad 1 \leq j \leq 5$$

La decisión de asignar pesos uniformes se fundamenta en la suposición de que todos los criterios contribuyen de manera equivalente a la identificación de patrones relevantes para el control de la COVID-19.

La Figura 6.2 muestra las puntuaciones obtenidas por las reglas bajo dos enfoques: el método TOPSIS clásico (línea roja) con pesos fijos de 0.2 para cada criterio, y el método uwTOPSIS (línea azul) con intervalos definidos. A primera vista, ambos métodos producen rankings similares; sin embargo, un análisis más detallado revela diferencias significativas.

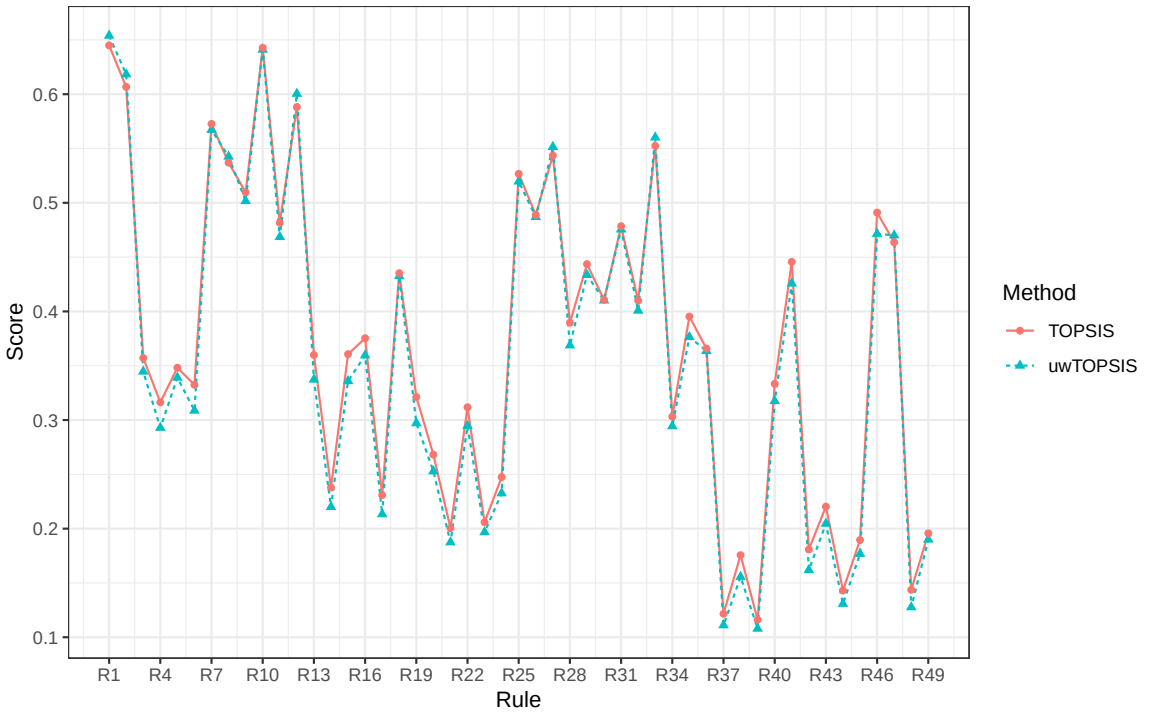


Figura 6.2: Resultados obtenidos mediante uwTOPSIS y TOPSIS clásico

Aunque las distribuciones de las puntuaciones de ambos métodos son aproximadamente normales, el método uwTOPSIS ofrece una distribución más suave y continua. Esto se observa claramente en la Figura 6.3, donde la distribución de densidad de los puntajes de TOPSIS tradicional presenta discontinuidades, en contraste con la distribución normal continua del uwTOPSIS. Esta aparente pequeña diferencia tiene un impacto importante en el ordenamiento final de las reglas.

La Tabla 6.2 muestra la clasificación completa de las 49 reglas según ambos métodos y las diferencias de posición entre ellos. Se observan variaciones significativas, especialmente entre las posiciones 11 a 18 y 23 a 28, lo que indica una sensibilidad estructural en el método tradicional de TOPSIS causada por sus discontinuidades. En cambio, el método uwTOPSIS proporciona una clasificación más estable y coherente, al basarse en una distribución continua de las puntuaciones.

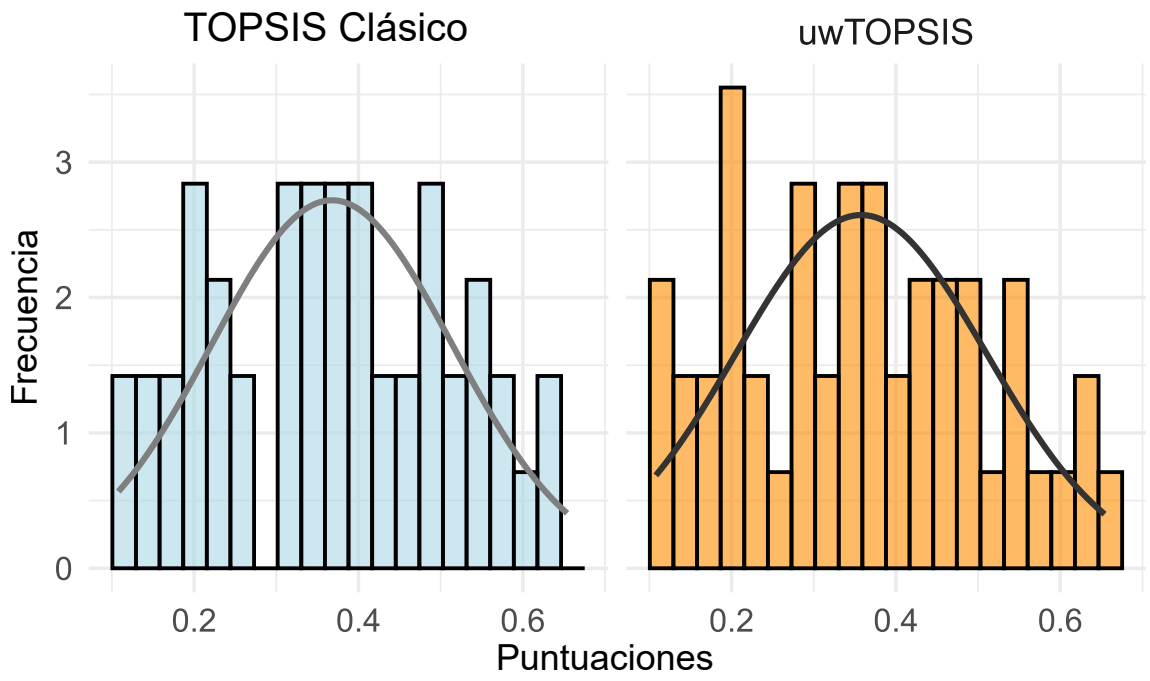


Figura 6.3: Histograma de frecuencias relativas entre uwTOPSIS y el TOPSIS clásico

A pesar de estas diferencias en los rankings individuales, las puntuaciones promedio de las reglas son similares en ambos métodos, como se resume también en la Tabla 6.2. Esta convergencia en las medias indica una concordancia global, pero oculta diferencias relevantes en posiciones específicas del ranking.

Una de las principales ventajas del método uwTOPSIS es que no requiere realizar análisis de escenarios por separado, ya que el propio enfoque considera los extremos del intervalo de pesos: los valores inferiores representan escenarios pesimistas y los superiores, escenarios optimistas. La Figura 6.4 presenta los valores de uwTOPSIS con sus respectivos máximos, mientras que la Figura 6.5 muestra los valores obtenidos por el método TOPSIS tradicional con pesos fijos.

Tabla 6.2: Comparacion de la ordenación de las 49 reglas por los métodos uwTOPSIS y TOPSIS clásico

| Regla | Sc_uwT | Rank_uwT | Sc_trad | Rank_Trad | Dif. | Regla | Sc_uwT | Rank_uwT | Sc_trad | Rank_Trad | Dif. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| R1 | 0.65 | 1 | 0.64 | 1 | 0 | R5 | 0.34 | 26 | 0.35 | 28 | -2 |
| R10 | 0.64 | 2 | 0.64 | 2 | 0 | R13 | 0.34 | 27 | 0.36 | 26 | 1 |
| R2 | 0.62 | 3 | 0.61 | 3 | 0 | R15 | 0.34 | 28 | 0.36 | 25 | 3 |
| R12 | 0.60 | 4 | 0.59 | 4 | 0 | R40 | 0.32 | 29 | 0.33 | 29 | 0 |
| R7 | 0.57 | 5 | 0.57 | 5 | 0 | R6 | 0.31 | 30 | 0.33 | 30 | 0 |
| R33 | 0.56 | 6 | 0.55 | 6 | 0 | R19 | 0.30 | 31 | 0.32 | 31 | 0 |
| R27 | 0.55 | 7 | 0.54 | 7 | 0 | R22 | 0.29 | 32 | 0.31 | 33 | -1 |
| R8 | 0.54 | 8 | 0.54 | 8 | 0 | R34 | 0.29 | 33 | 0.30 | 34 | -1 |
| R25 | 0.52 | 9 | 0.53 | 9 | 0 | R4 | 0.29 | 34 | 0.32 | 32 | 2 |
| R9 | 0.50 | 10 | 0.51 | 10 | 0 | R20 | 0.25 | 35 | 0.27 | 35 | 0 |
| R26 | 0.49 | 11 | 0.49 | 12 | -1 | R24 | 0.23 | 36 | 0.25 | 36 | 0 |
| R31 | 0.48 | 12 | 0.48 | 14 | -2 | R14 | 0.22 | 37 | 0.24 | 37 | 0 |
| R46 | 0.47 | 13 | 0.49 | 11 | 2 | R17 | 0.21 | 38 | 0.23 | 38 | 0 |
| R47 | 0.47 | 14 | 0.46 | 15 | -1 | R43 | 0.20 | 39 | 0.22 | 39 | 0 |
| R11 | 0.47 | 15 | 0.48 | 13 | 2 | R23 | 0.20 | 40 | 0.21 | 40 | 0 |
| R29 | 0.43 | 16 | 0.44 | 17 | -1 | R49 | 0.19 | 41 | 0.20 | 42 | -1 |
| R18 | 0.43 | 17 | 0.44 | 18 | -1 | R21 | 0.19 | 42 | 0.20 | 41 | 1 |
| R41 | 0.43 | 18 | 0.45 | 16 | 2 | R45 | 0.18 | 43 | 0.19 | 43 | 0 |
| R30 | 0.41 | 19 | 0.41 | 19 | 0 | R42 | 0.16 | 44 | 0.18 | 44 | 0 |
| R32 | 0.40 | 20 | 0.41 | 20 | 0 | R38 | 0.16 | 45 | 0.18 | 45 | 0 |
| R35 | 0.38 | 21 | 0.40 | 21 | 0 | R44 | 0.13 | 46 | 0.14 | 47 | -1 |
| R28 | 0.37 | 22 | 0.39 | 22 | 0 | R48 | 0.13 | 47 | 0.14 | 46 | 1 |
| R36 | 0.36 | 23 | 0.37 | 24 | -1 | R37 | 0.11 | 48 | 0.12 | 48 | 0 |
| R16 | 0.36 | 24 | 0.38 | 23 | 1 | R39 | 0.11 | 49 | 0.12 | 49 | 0 |
| R3 | 0.34 | 25 | 0.36 | 27 | -2 | | | | | | |

Como subraya Benítez y Liern (2021), uwTOPSIS es menos sensible a variaciones en los pesos que otros métodos de decisión multicriterio (MCDM). No obstante, su comportamiento sigue estando influenciado por los parámetros $k_1$, $k_2$, y los límites establecidos para los pesos $l_j$ y $u_j$.

## 6.4 ANÁLISIS DE SENSIBILIDAD

Con el objetivo de evaluar la estabilidad del método TOPSIS frente a variaciones en los pesos asignados a los criterios, se llevó a cabo un análisis de sensibilidad basado en simulaciones. Esta etapa metodológica es fundamental para verificar la robustez de los resultados derivados del método TOPSIS cuando se aplica al contexto de reglas de asociación, donde las ponderaciones asignadas a los distintos criterios pueden no estar completamente definidas o pueden contener cierto grado de incertidumbre.

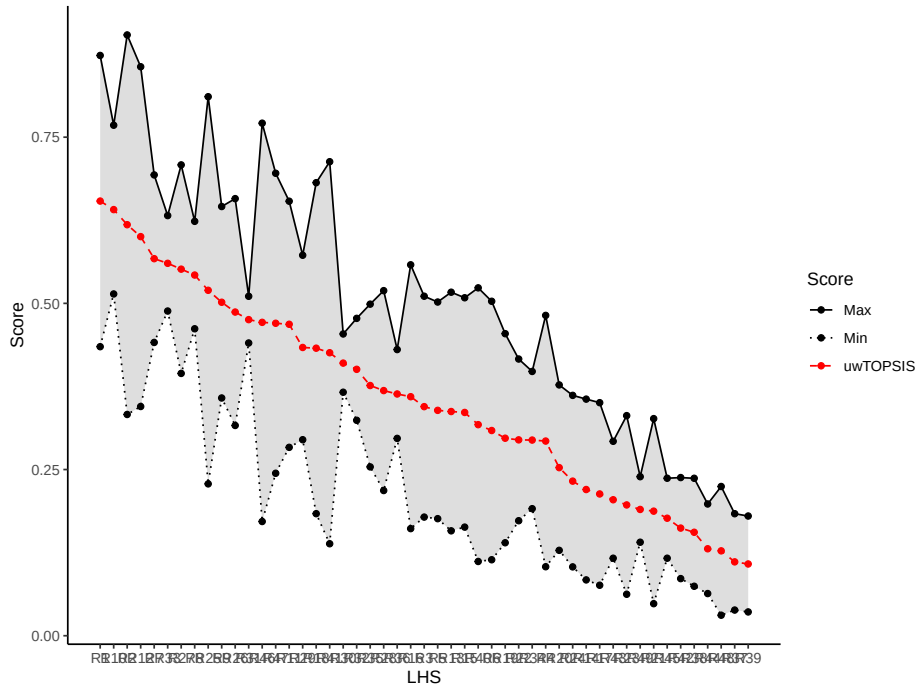


Figura 6.4: Ordenación uwTOPSIS

El procedimiento consistió en simular distintas configuraciones de pesos alrededor de un vector base, manteniendo la suma total igual a 1 (condición esencial en métodos de decisión multicriterio). Se definió inicialmente un vector de pesos uniforme, dado por $w_0 = (0.2, 0.2, 0.2, 0.2, 0.2)$, correspondiente a cinco criterios considerados igualmente importantes.

Posteriormente, se generaron vectores aleatorios distribuidos uniformemente sobre la superficie de una hiperesfera en cuatro dimensiones, ya que el quinto peso se ajustó automáticamente para que la suma total permaneciera constante. Estos vectores aleatorios fueron escalados por un parámetro $\delta$, que representa el nivel de perturbación o desviación respecto al vector base. Se consideraron ocho valores diferentes para $\delta$: 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35 y 0.40. Para cada valor de $\delta$, se generaron 1000 vectores de pesos simulados, totalizando 8000 simulaciones.

Cada vector de pesos fue utilizado para aplicar nuevamente el método TOPSIS sobre la misma base de datos de reglas de asociación, y los resultados obtenidos fueron comparados con la clasificación original generada a partir del vector base $w_0$. Para medir la sensibilidad del método, se calculó el cambio relativo promedio en las puntuaciones TOPSIS, definido como $\Delta R$, que representa la media de las diferencias relativas entre las clasificaciones simuladas y las originales.

En la Figura 6.6, se presentan los resultados de esta evaluación. El gráfico muestra cómo la media de $\Delta R$ se incrementa a medida que aumenta el valor de $\delta$. Esto indica que el método TOPSIS es sensible a variaciones en los pesos: perturbaciones mayores conducen a cambios más significativos en la clasificación de las reglas.

Además, se puede observar que para valores bajos de $\delta$ (hasta aproximadamente 0.15), el valor medio de $\Delta R$ permanece por debajo de 0.10, lo que sugiere una buena estabilidad

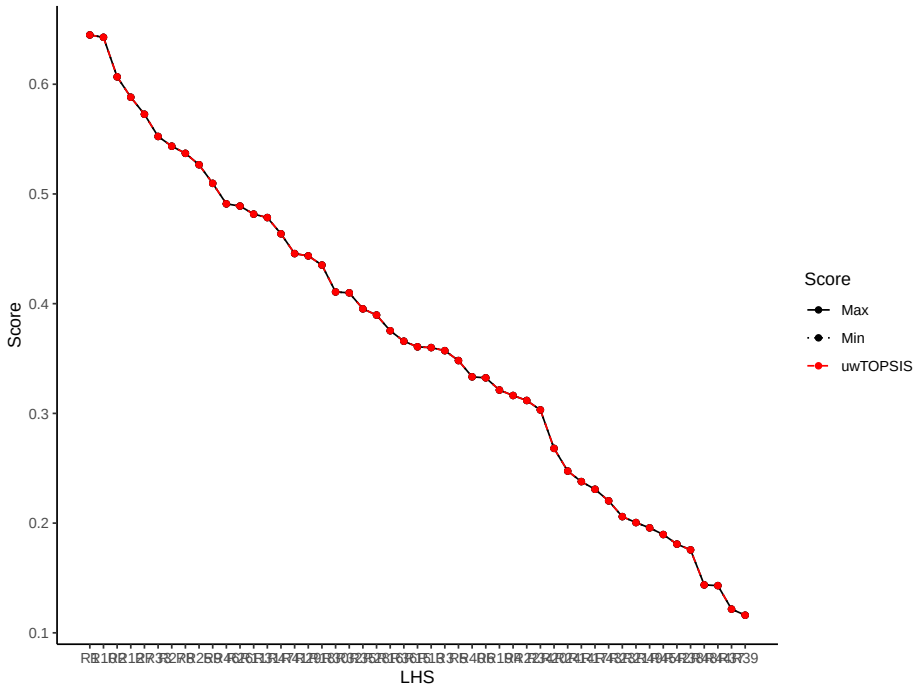


Figura 6.5: TOPSIS Clásico

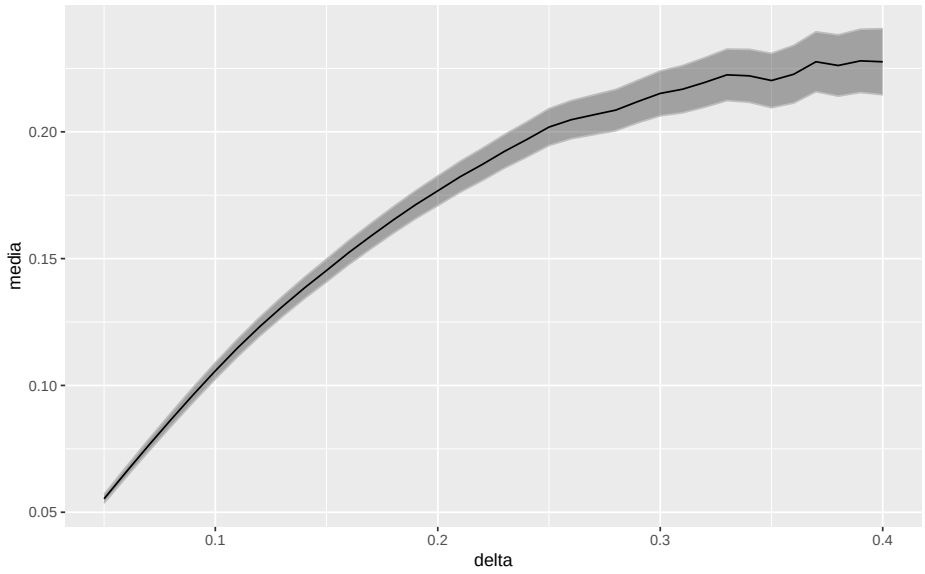


Figura 6.6: Análisis de sensibilidad en TOPSIS

del método frente a pequeñas fluctuaciones en las ponderaciones. Sin embargo, cuando $\delta$ supera 0.30, las diferencias relativas superan en promedio el 20 %, lo que revela una pérdida considerable de confiabilidad en los resultados del TOPSIS bajo condiciones de alta incertidumbre en los pesos.

Para cada nivel de $\delta$, también se estimaron los intervalos de confianza utilizando un test t, lo que permite visualizar la variabilidad de los resultados y la incertidumbre asociada a las simulaciones. La anchura de las bandas de confianza aumenta con $\delta$, reforzando la conclusión de que las decisiones basadas en TOPSIS pueden volverse menos previsibles cuando hay mayor imprecisión en los pesos.

En conclusión, el análisis de sensibilidad demuestra que el método TOPSIS es robusto frente a pequeñas perturbaciones en los pesos de los criterios, pero su estabilidad disminuye progresivamente con el aumento de la incertidumbre. Esta información es particularmente útil en el contexto de reglas de asociación, donde la asignación precisa de pesos puede ser un desafío, y resalta la importancia de validar la robustez de los métodos multicriterio antes de su aplicación en la toma de decisiones.

# 7

# ANÁLISIS DE LA SATISFACCIÓN DEL SISTEMA EDUCATIVO EN MOZAMBIQUE CON REGLAS DE ASOCIACIÓN FUZZY Y UWTOPSIS

## 7.1 INTRODUCCIÓN

Uno de los objetivos de esta tesis ha sido desarrollar una metodología que permitiese identificar reglas del tipo si-entonces propiamente difusas, reducir la redundancia entre las reglas generadas y ordenarlas mediante el método uwTOPSIS, todo ello como una extensión de la metodología fsQCA que tanta aplicación ha tenido en ciencias sociales. La validez de este enfoque ha sido contrastada a partir de datos extraídos de estudios previamente publicados por distintos autores en diferentes contextos disciplinarios.

En este capítulo se ilustra la metodología presentada en esta tesis con una aplicación a un caso empírico, en el cual se emplearon reglas de asociación fuzzy, un algoritmo de minimización lógica y el método uwTOPSIS para generar soluciones propiamente difusas dentro de un problema típico del campo de las ciencias sociales, como es el campo de la educación. Concretamente analizaremos qué condiciones son las conducentes a la satisfacción de los estudiantes de ciencias de la educación de la Universidad de Licungo en Mozambique.

La estructura de los datos utilizados en el análisis de la satisfacción estudiantil en el sistema educativo puede representarse mediante términos lingüísticos tanto en las variables de entrada como de salida. El fenómeno de la satisfacción en el contexto educativo posee una naturaleza compleja y multidimensional, por lo que su análisis requiere una diferenciación clara entre los factores que dependen directamente de los propios estudiantes y aquellos que son atribuibles a las condiciones ofrecidas por la institución educativa.

Los datos utilizados en este capítulo fueron recolectados durante una estancia en el Centro de Investigación de la Universidad Licungo, ubicada en Quelimane, Mozambique, África. Durante los meses de agosto, septiembre y octubre de 2024, se recopilaron datos en los 10 cursos de licenciatura de la Facultad de Educación, dedicada a la formación de profesores de enseñanza secundaria en distintas disciplinas. Los resultados obtenidos también fueron analizados utilizando el enfoque fuzzy-set Qualitative Comparative Analysis (fsQCA), cuyo estudio se encuentra actualmente en revisión para publicación en la revista científica Rect@ (Dom Luís, Benítez, M. d. C. Bas y Cuamba, 2025, mayo).

Las variables relacionadas con los estudiantes reflejan percepciones, actitudes, expectativas y niveles de involucramiento en el proceso educativo. Entre ellas se incluyen la motiva-

ción, el compromiso con los estudios, las expectativas personales con respecto al curso y a la carrera profesional, los hábitos de estudio, la participación en actividades extracurriculares, la satisfacción con el propio rendimiento, así como las condiciones socioeconómicas y el contexto familiar. Por otro lado, las variables asociadas a la institución se refieren a los aspectos estructurales y organizativos que configuran el entorno educativo. Estas comprenden la calidad del cuerpo docente, la infraestructura física y tecnológica, el acceso a bibliotecas y recursos digitales, la organización curricular, el acompañamiento académico y administrativo, el ambiente de aprendizaje y seguridad, y las políticas institucionales de inclusión y apoyo al estudiante. Estas últimas corresponden a una responsabilidad directa de la institución, que debe garantizar condiciones adecuadas para el aprendizaje y el bienestar de su comunidad estudiantil.

Es importante destacar que la calidad del proceso formativo de los estudiantes está fuertemente condicionada por las oportunidades, recursos y apoyos que la institución es capaz de ofrecer. En este sentido, el análisis de la satisfacción estudiantil puede orientar la mejora de políticas institucionales con base en evidencia empírica.

La aplicación del método propuesto tiene como finalidad contribuir a la mejora de las políticas institucionales mediante una comprensión más precisa y operativa de los factores que inciden en la satisfacción de los estudiantes.

### 7.1.1 *Características de la Población*

Mozambique, cuyo nombre oficial es República de Mozambique, se encuentra en el sureste de África, con una costa bañada por el Océano Índico. Limita al norte con Tanzania, al noreste con Malawi y Zambia, al oeste con Zimbabue, y al suroeste con Esuatini (antiguamente Suazilandia) y Sudáfrica. Su ubicación geográfica está representada en la Figura 7.1.

Con una extensión total de 801.590 $km^2$, el país está dividido administrativamente en 11 provincias. Para el año 2025, se estima que su población rondará los 33 millones de habitantes, con una distribución demográfica caracterizada por una alta proporción de jóvenes y un incremento en la urbanización.

### 7.1.2 *Estructura del sistema educativo en Mozambique*

El sistema educativo en Mozambique está compuesto principalmente por los subsistemas de enseñanza primaria, secundaria general, técnico-profesional y superior, siendo que la enseñanza primaria abarca desde el 1º hasta el 7º grado, normalmente iniciada por individuos de seis años de edad, estructurándose en tres ciclos, donde el primero corresponde al 1º y 2º grado, el segundo del 3º al 5º grado y el tercero incluye el 6º y 7º grado, mientras que la enseñanza secundaria general, que acoge a estudiantes con una edad media de catorce años, se compone de dos ciclos, siendo el primero del 8º al 10º grado y el segundo formado por el 11º y 12º grados, y al final de cada ciclo el alumno se somete a exámenes finales que determinan su progresión o no al siguiente ciclo (Dom Luís y Mulema, 2021).

Figura 7.1: Localización de Mozambique en el globo.

Paralelamente a la enseñanza secundaria general, tiene lugar la educación técnico-profesional con una duración de dos a tres años, que busca preparar a los ciudadanos con competencias técnicas adecuadas para enfrentar los desafíos del mercado laboral, ofreciendo una vía alternativa a la formación general, más orientada a la práctica y a la inserción profesional temprana.

En la cima del sistema educativo se encuentra la enseñanza superior, dividida en tres ciclos principales, siendo el primero correspondiente a la licenciatura, con una duración media de cuatro años, el segundo al máster, que generalmente se extiende por dos años, y el tercero al doctorado, con duración variable, que culmina con la producción de una tesis original y contribuciones científicas al avance del conocimiento, conformando así un recorrido académico vertical que va desde la formación inicial hasta la producción científica y el desarrollo de competencias avanzadas.

Después de la independencia nacional, y especialmente a partir de los años 90, la enseñanza superior en Mozambique pasó por un proceso de expansión tanto en número de instituciones como de estudiantes, aunque inicialmente esta expansión se concentró en la ciudad de Maputo y posteriormente se extendió por todo el territorio nacional, observándose también un desarrollo de la formación de posgrado que, en un primer momento, ocurría casi exclusivamente en el exterior, pero que con el tiempo pasó a incluir másteres y algunos programas de doctorado realizados en el propio país, generalmente en asociación con instituciones extranjeras.

## 7.2 SATISFACCIÓN EN LA EDUCACIÓN

La educación constituye la base fundamental para el desarrollo de cualquier nación. El progreso de un país comienza, esencialmente, con la capacidad del gobierno para proporcionar mejores condiciones educativas a sus ciudadanos. En todo el mundo, los gobiernos han demostrado una preocupación constante por priorizar la educación, con el objetivo principal de dotar a sus recursos humanos de excelencia en conocimientos y alta productividad. En este contexto, las instituciones educativas desempeñan un papel vital en la formación de graduados no solo en términos cuantitativos, sino también cualitativos, preparando individuos dinámicos, innovadores y proactivos que actúan como catalizadores del desarrollo. Estas instituciones también contribuyen a la creación de una fuerza laboral calificada, capaz de responder a diferentes niveles de especialización (Wibowo, 2022).

En general, todas las instituciones educativas, ya sean públicas o privadas, deben mejorar constantemente sus servicios y modernizar sus instalaciones para maximizar el potencial de sus usuarios. Los servicios educativos prestados por las escuelas comprenden actividades esenciales dentro del sistema educativo, destinadas a atender las necesidades de la comunidad (Dinh et al., 2021)

En el marco del sistema educativo, estos servicios incluyen tanto instalaciones físicas como infraestructura académica. Para mejorar tales servicios, las instituciones deben adoptar estrategias que incluyan, por un lado, la continua capacitación de sus recursos humanos y, por otro, el fortalecimiento de la infraestructura que sustenta el funcionamiento integral de los servicios académicos (Pasaribu et al., 2020; Santika et al., 2021).

Existe un reconocimiento generalizado sobre la necesidad de que las escuelas se adapten a diversos factores externos e internos que influyen en el desarrollo del ámbito educativo. Los factores externos hacen referencia a aquellos sobre los cuales las instituciones educativas no tienen control directo, mientras que los factores internos se relacionan con aspectos bajo la responsabilidad y gestión directa de las propias instituciones. En este contexto, resulta imperativo que los servicios educativos diversifiquen los productos que ofrecen, de modo que la satisfacción estudiantil, en tanto que consumidores, se convierta en un referente crucial en la elección de dichos servicios.

Los estudiantes, como principales usuarios de las escuelas, esperan beneficiarse de las instalaciones y recursos disponibles que respaldan y enriquecen el proceso de aprendizaje. Diversos estudios indican que la insatisfacción estudiantil con la experiencia educativa suele estar vinculada a la insuficiencia de la infraestructura ofrecida. Por ejemplo, los servicios administrativos son frecuentemente señalados como un punto crítico debido a la ausencia o ineficacia de procedimientos operativos estandarizados que podrían apoyar de manera más eficiente los programas educativos.

En consecuencia, las instituciones educativas tienen la responsabilidad de comprender y abordar cada dimensión de los indicadores que los estudiantes consideran esenciales. Existe una relación causal clara entre la satisfacción del estudiante y el nivel de calidad de los servicios ofrecidos. La satisfacción en el sistema educativo surge de la comparación entre las expectativas iniciales de los estudiantes (lo que esperan, simbólicamente guarda-

do “en la mano izquierda”) y las percepciones que desarrollan con el tiempo sobre los servicios prestados (lo que experimentan, guardado “en la mano derecha”). Este contraste entre expectativas y percepciones define el sentido de (in)satisfacción.

Así, corresponde a las instituciones educativas implementar políticas sólidas que minimicen la brecha entre lo esperado y lo entregado. La calidad de los servicios educativos debe estar alineada con las expectativas individuales de los estudiantes, abarcando aspectos como la calidad de las actividades de enseñanza y aprendizaje, la infraestructura escolar, el compromiso y la cultura laboral del profesorado, el acceso a internet y el uso de medios educativos. En resumen, la mejora continua de estos servicios no solo constituye un factor diferenciador, sino también una obligación para asegurar un entorno educativo que promueva la plena satisfacción y el desarrollo integral del estudiante.

### 7.2.1 *Calidad en la Educación Superior*

La calidad ha acompañado la historia de la civilización humana y ha evolucionado desde prácticas artesanales hasta complejos sistemas de gestión. En el ámbito de la educación superior, su aplicación se centra en la capacidad de las instituciones para satisfacer las expectativas académicas, sociales y personales de los estudiantes.

La satisfacción estudiantil se considera actualmente un componente esencial de los sistemas de aseguramiento de la calidad. Evaluarla implica identificar en qué medida los servicios y procesos ofrecidos por la universidad responden a las necesidades de su alumnado. Esta visión está en consonancia con la definición propuesta por la norma ISO 9000:2000, que define la calidad como el grado en que un conjunto de características cumple con requisitos específicos.

La gestión de la calidad educativa requiere un enfoque integrado que contemple los siguientes principios:

1. Enfoque en el estudiante: comprender las necesidades individuales y colectivas de los alumnos, tanto en el ámbito académico como en el bienestar general.

2. Mejora continua: adaptar programas de estudio, métodos de enseñanza y servicios de apoyo a partir de datos recogidos de manera sistemática.

3. Participación activa: fomentar el compromiso de docentes, estudiantes y personal administrativo en el desarrollo del proceso educativo.

4. Toma de decisiones basada en evidencias: fundamentar las acciones institucionales en datos concretos, evitando la improvisación y promoviendo la eficacia.

Instrumentos como los cuestionarios de satisfacción permiten analizar dimensiones clave del entorno universitario, tales como la calidad metodológica, la infraestructura física y tecnológica, el apoyo institucional y el compromiso docente.

### 7.2.2 *La Universidad Licungo*

La Universidad Licungo (UniLicungo) está ubicada en la zona centro de Mozambique. Fue creada mediante el Decreto n.º 3/2019, de 14 de febrero, como parte de una estrategia gubernamental para reestructurar el sistema de educación superior del país.

La institución surgió a partir de la transformación de las delegaciones de la antigua Universidad Pedagógica (específicamente las delegaciones de Quelimane y Beira), integrando sus recursos humanos, materiales y financieros. Desde entonces, la UniLicungo actúa en las provincias de Zambezia y Sofala, desempeñando un papel crucial en la formación de profesionales en la región central.

Con el paso del tiempo, la UniLicungo se ha consolidado como la mayor institución pública de educación superior en esta zona del país, destacándose por su enfoque en la formación de docentes, la investigación educativa y su compromiso con el desarrollo comunitario y regional.

## 7.3 METODOLOGÍA

### 7.3.1 *Instrumento de Recolección de Datos*

El objetivo principal de esta investigación fue analizar las condiciones causales de la insatisfacción global de los estudiantes de la UniLicungo, utilizando reglas de asociación fuzzy y el método uwTOPSIS. Para ello, se tomaron en cuenta diversos factores que influyen en la experiencia académica de los estudiantes.

En particular, se buscó evaluar cómo perciben los estudiantes la calidad de la enseñanza, los recursos disponibles, el apoyo institucional y otros aspectos del entorno universitario. Este análisis resultó fundamental para comprender las percepciones estudiantiles respecto a los distintos componentes del proceso educativo en esta institución de educación superior.

### 7.3.2 *Muestreo*

Para garantizar la representatividad de la muestra con respecto a la población de futuros docentes de la UniLicungo, se adoptó un muestreo estratificado, utilizando el curso como criterio de estratificación. Esta estrategia permitió captar la diversidad de experiencias estudiantiles en las distintas áreas de formación docente.

El estudio abarcó diez programas de la Facultad de Educación, todos orientados a la formación de profesores. Los cursos incluidos fueron: Enseñanza de Lengua Inglesa, Enseñanza de Geografía, Enseñanza de Química, Enseñanza de Biología, Enseñanza de Psicología, Enseñanza de Matemáticas, Enseñanza de Física, Enseñanza de Lengua Portuguesa, Enseñanza de Lengua Francesa y Enseñanza de Educación Visual.

En total, se seleccionaron 244 estudiantes, garantizando una representación adecuada de todos los cursos ofrecidos por la Facultad de Educación.

### 7.3.3 *Aplicación del Cuestionario*

La recolección de datos se llevó a cabo mediante un cuestionario en línea, elaborado a través de la plataforma Google Forms, que contenía 20 preguntas cerradas basadas en una escala de Likert del 1 al 10. Esta escala permitió a los participantes expresar detalladamente su nivel de satisfacción, desde 1 (totalmente insatisfecho) hasta 10 (totalmente satisfecho).

El cuestionario fue estructurado en seis dimensiones clave que abarcan aspectos fundamentales de la calidad educativa, las condiciones de aprendizaje y el bienestar estudiantil. Estas dimensiones permiten una evaluación integral de la experiencia universitaria desde la perspectiva del estudiante:

**Dimensión 1: Metodología de Enseñanza y Calidad Académica (MECA)** Evalúa el grado de satisfacción de los estudiantes en relación con las metodologías pedagógicas utilizadas por el profesorado, la claridad y profundidad de los contenidos impartidos, la adecuación y disponibilidad de materiales didácticos, así como la coherencia y objetividad de los sistemas de evaluación. Se fundamenta en la premisa de que una enseñanza de calidad es esencial para el desarrollo de competencias profesionales y pensamiento crítico.

**Dimensión 2: Accesibilidad Física y Recursos Institucionales (AFRI)** Analiza el acceso físico a las instalaciones universitarias tales como aulas, laboratorios, auditorios, áreas administrativas y espacios de estudio. Considera también la señalización, la adecuación para personas con discapacidad y la facilidad de circulación dentro del campus. Esta dimensión valora el entorno físico como facilitador del proceso educativo.

**Dimensión 3: Cortesía y Atención al Estudiante (CAR)** Evalúa la calidad del trato ofrecido por el personal docente, administrativo y de apoyo. Se enfoca en la amabilidad, el respeto, la disposición para atender las dudas y necesidades de los estudiantes, y la eficacia de los canales de comunicación. Esta dimensión reconoce la importancia de un entorno institucional empático y humanizado para el bienestar emocional y motivacional del estudiante.

**Dimensión 4: Puntualidad y Compromiso del Profesorado (PCP)** Mide la puntualidad de los docentes en el cumplimiento de sus horarios, la regularidad en la realización de clases, y su involucramiento activo en la formación académica. El compromiso docente es clave para garantizar continuidad, coherencia y responsabilidad en el proceso educativo, reforzando la confianza y el respeto del alumnado hacia la institución.

**Dimensión 5: Infraestructura y Saneamiento (IS)** Analiza las condiciones físicas de la universidad, incluyendo la limpieza de aulas, pasillos, servicios sanitarios y patios; el estado de conservación de los espacios; y el nivel de confort, seguridad e higiene que proporcionan. La infraestructura adecuada es un componente esencial para promover un entorno educativo saludable y funcional.

**Dimensión 6: Biblioteca e Internet (BI)** Evalúa la disponibilidad, accesibilidad y calidad de los servicios de biblioteca, tanto física como digital, así como la estabilidad, velocidad y cobertura del servicio de internet en el campus. Estos recursos son fundamentales para el desarrollo académico, el acceso a fuentes de información confiables y el fomento del aprendizaje autónomo y la investigación.

El cuestionario fue aplicado exclusivamente en formato digital, lo que ofreció mayor flexibilidad a los estudiantes y facilitó tanto la recolección como la organización y el análisis de los datos. Los resultados se organizaron por curso, y se calcularon los promedios de satisfacción correspondientes a cada dimensión.

### 7.3.4 *Calibración de los datos*

La calibración constituye una etapa fundamental en la extracción de reglas de asociación fuzzy, ya que los puntajes de pertenencia obtenidos durante este proceso orientan todas las fases posteriores del análisis. Este procedimiento transforma los datos numéricos brutos en puntajes de pertenencia difusa, asignando niveles como "totalmente dentro" (fuzzy = 1,0) y "totalmente fuera" (fuzzy = 0,0). La calibración se basa en umbrales empíricos y teóricos, lo que asegura que los valores calibrados reflejen niveles claros y conceptualmente significativos de pertenencia a los conjuntos.

En este estudio, se aplicó el método de calibración triangular propuesto por Zeedh, utilizando tres términos lingüísticos (baja satisfacción, satisfacción media y alta satisfacción). Estos términos fueron definidos a partir de umbrales tanto cualitativos como empíricos. Los datos originales estaban compuestos por puntuaciones ordinales categóricas, con valores que oscilaban entre 1 (muy insatisfecho) y 10 (muy satisfecho).

El proceso de calibración se fundamentó en medidas de posición para establecer los puntos de corte. Específicamente, se utilizaron los percentiles P25, P50 y P75 como anclas para la transformación, garantizando así un equilibrio entre la comprensión teórica del fenómeno y su representación empírica. Esta estrategia permitió que los puntajes fuzzy representaran fielmente el significado sustantivo de las condiciones analizadas.

A diferencia de la metodología fsQCA, las reglas de asociación difusa no presentan restricciones en torno al punto de cruce de 0,5 —que en QCA representa ambigüedad máxima—, lo que amplia su aplicabilidad en contextos empíricos complejos. Para el análisis de los datos, se empleó el software R, en particular el paquete `lfl`, diseñado para la extracción de reglas de asociación difusas, conforme a la propuesta de citeburda2014fuzzy.

Al adoptar este enfoque riguroso de calibración, se logró una interpretación más clara y precisa de las respuestas de satisfacción de los estudiantes, evitando ambigüedades y proporcionando un mapeo robusto de las percepciones. En esencia, el proceso de calibración —de naturaleza contextual— refleja tanto la calidad como el atributo de la información, al integrar conocimientos empíricos y conceptuales para obtener resultados optimizados.

## 7.4 RESULTADOS

### 7.4.1 *Extracción de Reglas de Asociación Difusas de Satisfacción Estudiantil*

Para la extracción de reglas de asociación difusas, se utilizó el algoritmo *Apriori*, adoptando un umbral de confianza de 0.75 y un soporte mínimo de 0.0001. Esta configuración tuvo como objetivo capturar reglas con bajo soporte pero con consistencia aceptable — lo cual es especialmente relevante en contextos educativos donde comportamientos significativos pueden ocurrir con baja frecuencia.

El proceso de extracción se dividió en dos etapas principales:

1. Identificación de reglas asociadas a la baja satisfacción de los estudiantes con el sistema educativo recibido;
2. Identificación de reglas asociadas a la alta satisfacción de los estudiantes con la institución.

### 7.4.2 *Análisis de la Baja Satisfacción de los Estudiantes con el Sistema Educativo*

En la primera etapa, se extrajeron 51 reglas asociadas a la baja satisfacción de los estudiantes de la UniLicungo. Debido a la gran cantidad de reglas generadas, se aplicó el método uwTOPSIS, conforme descrito en el capítulo anterior, con el objetivo de clasificar las reglas según múltiples criterios de evaluación: longitud de las reglas (a minimizar), confianza y soporte (a maximizar).

La principal ventaja del método uwTOPSIS es que está completamente orientado por datos, sin necesidad de evaluaciones subjetivas ni asignaciones lingüísticas arbitrarias. Para la clasificación, se definieron límites inferior y superior para los pesos de los criterios: 0.05 (inferior) y 0.45 (superior) para todos ellos. La elección de pesos uniformes se justifica por la suposición de que todos los criterios contribuyen de manera equilibrada a la calidad de las reglas.

La Figura 7.2 muestra los resultados de las puntuaciones obtenidas con el *uwTOPSIS* (línea roja), así como los valores máximos y mínimos para cada alternativa (representados por la franja gris).

La Figura 7.3 presenta las principales reglas ordenadas según las puntuaciones obtenidas.

Para facilitar el análisis de los patrones, se utilizó un mapa de calor. Este mapa (ver en la Figura 7.4) destaca las combinaciones de dimensiones de insatisfacción que aparecen con mayor frecuencia en los antecedentes de las reglas asociadas a la baja satisfacción general. Cada celda representa la frecuencia con que dos factores — por ejemplo, `Bajo.CAR` (baja cortesía y atención) y `Bajo.IS` (infraestructura y saneamiento) — ocurren juntos en las reglas.

Un ejemplo importante es la combinación frecuente entre baja cortesía y atención e infraestructura y saneamiento inadecuados, que aparecen juntas en cinco reglas. Esto sugiere

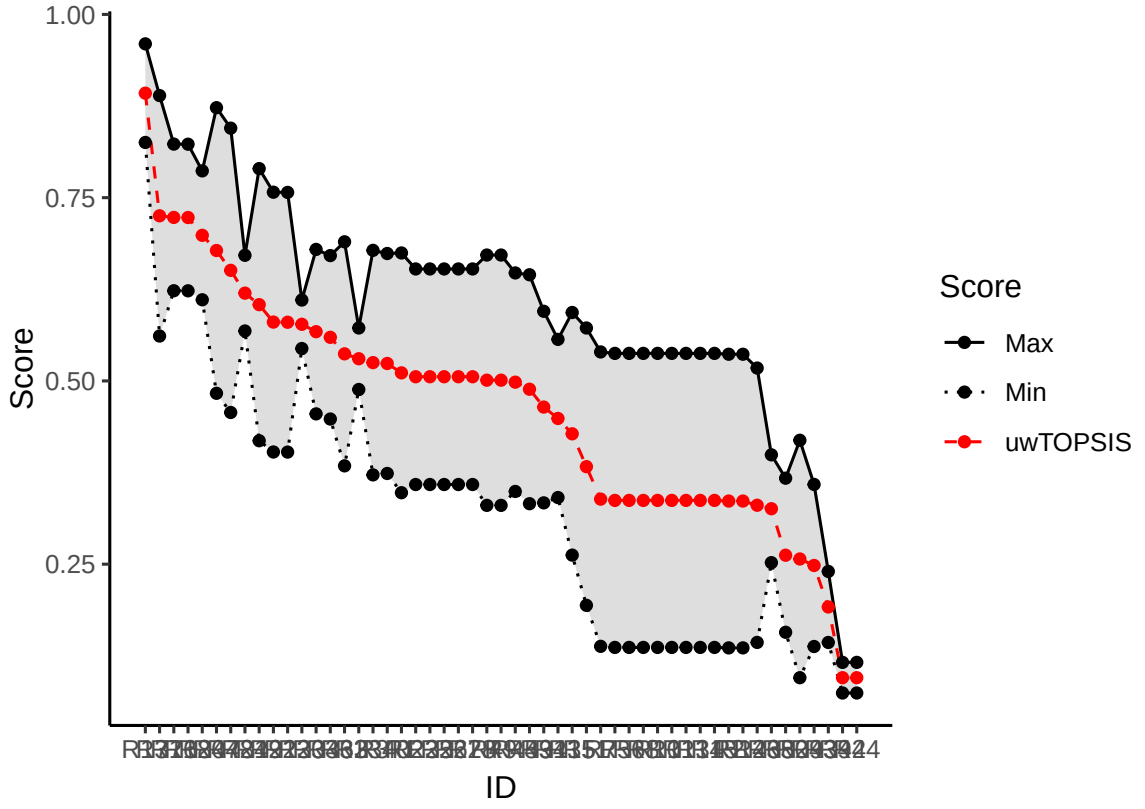


Figura 7.2: Ordenación uwTOPSIS según las puntuaciones

que, siempre que los estudiantes perciben ausencia de atención o acogida, junto con condiciones deficientes de los edificios, hay una tendencia acentuada hacia la insatisfacción general. Se trata de un efecto combinado negativo significativo, ya que tanto el entorno físico como la relación interpersonal impactan directamente en el clima institucional y la motivación de los estudiantes.

Otro patrón identificado muestra una ocurrencia moderada entre baja cortesía y atención y baja puntualidad docente. Además, la percepción de falta de compromiso docente y fallos metodológicos, especialmente cuando se asocian con la ausencia de acogida institucional, afecta negativamente la confianza de los estudiantes en el proceso de enseñanza-aprendizaje.

Tambiem, se observa que variables como accesibilidad intermedia o media satisfacción con la biblioteca y internet aparecen con poca frecuencia y generalmente de forma aislada. Esto indica que estas condiciones intermedias, por sí solas, no explican la baja satisfacción, pudiendo actuar como factores moduladores, pero no como causas directas.

El análisis de las reglas de asociación extraídas a partir de los datos de satisfacción de los estudiantes de UniLicungo revela patrones recurrentes que ayudan a comprender los factores más críticos asociados a la insatisfacción estudiantil. Las reglas destacadas presentan, en general, soportes entre 0.002 y 0.0206, valores relativamente bajos, lo que indica que cada regla se aplica solo a una pequeña parte de la muestra total. No obstante, la mayoría de las reglas presenta índices de confianza elevados, generalmente superiores a 0.90, lo que sugiere que, cuando se cumplen las condiciones indicadas en las reglas, existe una alta

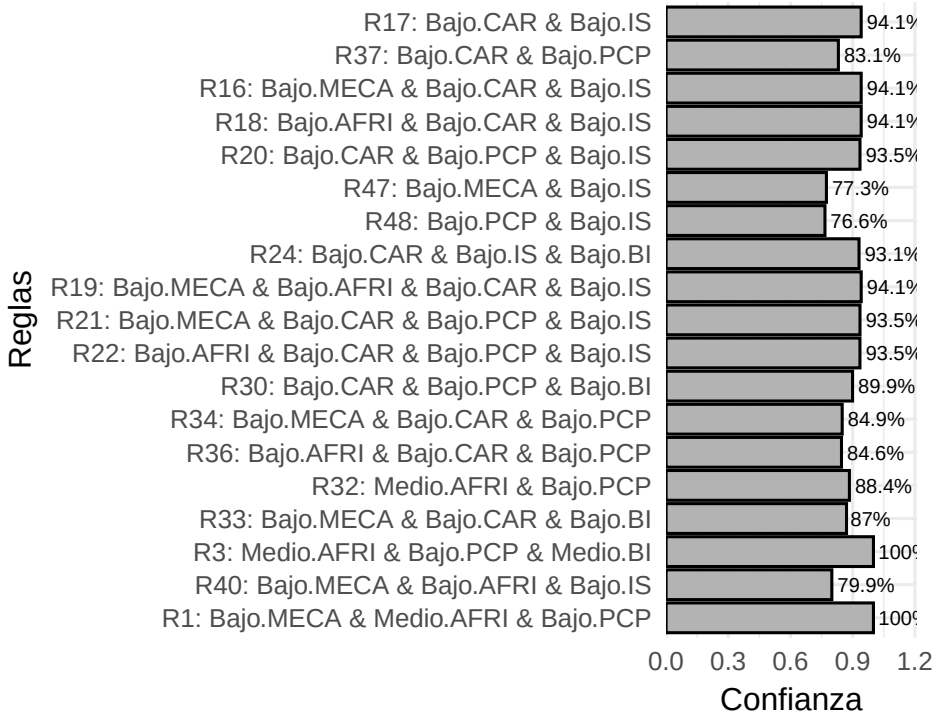


Figura 7.3: Top 20 de reglas ordenadas

probabilidad de que ocurra la consecuente baja satisfacción. Esto refuerza la relevancia de las combinaciones de factores identificadas, aunque se apliquen a subconjuntos reducidos de estudiantes.

Entre los factores que aparecen con mayor frecuencia en las combinaciones explicativas se encuentran la cortesía en la atención y las infraestructuras sanitarias. Ambas aparecen en más de la mitad de las reglas analizadas, ya sea de forma aislada o en conjunto con otras condiciones como la puntualidad y compromiso, la dimensión didáctico-científica y el acceso y disponibilidad (ver la Figura 7.5). Esto indica que la percepción de falta de respeto o de atención inadecuada, sumada a condiciones sanitarias insatisfactorias, son elementos clave del descontento estudiantil.

La combinación entre cortesía en la atención e infraestructura sanitaria aparece, por ejemplo, en las reglas R17, R16, R18, R20 y R24, todas con índices de confianza superiores a 0.93, lo que sugiere una fuerte asociación entre estas variables y la insatisfacción. Algunas reglas como la R47 y R48 demuestran que, incluso sin la presencia de cortesía en la atención, la combinación de factores como la infraestructura sanitaria con la dimensión didáctico-científica o la puntualidad también se relaciona fuertemente con la respuesta negativa de los estudiantes.

El análisis de las reglas de asociación en relación con las métricas de soporte y confianza permite obtener valiosos indicadores sobre qué condiciones son más relevantes para la satisfacción de los estudiantes, que en este caso se mide como “Bajo.Sat”. A través de un gráfico de dispersión que relaciona el soporte con la confianza de las reglas (7.6), se observa

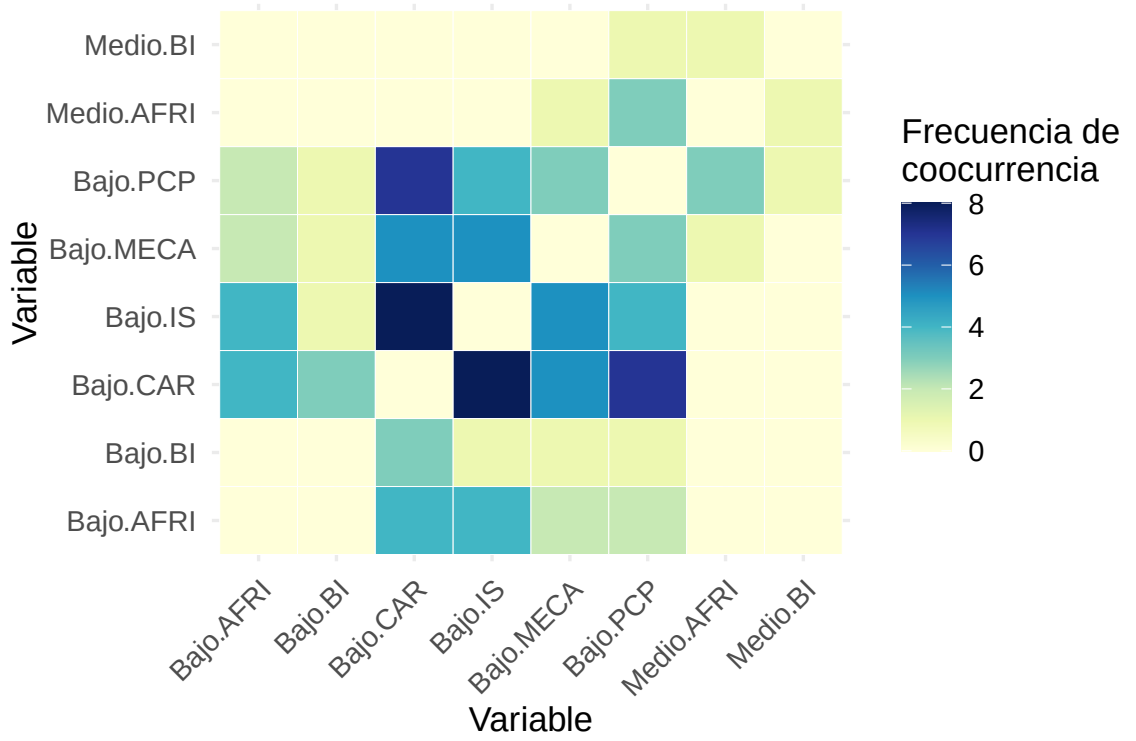


Figura 7.4: Mapa de calor de combinaciones de dimensiones de insatisfacción que aparecen con mayor frecuencia en los antecedentes de las reglas

que las reglas con mayor soporte no necesariamente presentan la máxima confianza. Sin embargo, muchas de estas se encuentran en regiones de alta confianza y con soportes medianos, lo que justifica su valor analítico para la formulación de políticas institucionales.

Las métricas de soporte y confianza, junto con las variables asociadas a cada dimensión de la universidad, pueden ofrecer conclusiones prácticas que impacten directamente las decisiones políticas para mejorar la calidad educativa. Por ejemplo, las reglas que presentan soporte medio a alto, como las reglas R17, R16, R18, R19 y R20, están asociadas con condiciones como "Bajo.CAR" (Cortesía y Atención al Estudiante) y "Bajo.IS" (Infraestructura y Saneamiento), lo que sugiere que la satisfacción de los estudiantes está fuertemente correlacionada con una atención adecuada y una infraestructura funcional. Las políticas dirigidas a mejorar estas áreas tienen el potencial de aumentar la satisfacción estudiantil de manera significativa.

Por otro lado, reglas como R32 y R3, que presentan un soporte muy bajo (alrededor de 0.0037) pero una confianza perfecta (1.0), indican que aunque estas condiciones son raras, cuando ocurren, tienen una relación directa y perfecta con una baja satisfacción. Esto resalta la importancia de evaluar las condiciones de accesibilidad física (Medio.AFRI) y los recursos de biblioteca e Internet (Medio.BI), incluso cuando estos factores no se encuentren en niveles extremos, para garantizar que no conduzcan a insatisfacción.

También se encuentran reglas con confianza baja y soporte moderado, como R47 y R48, que muestran que cuando existen variables como "Bajo.MECA" (Metodología de Enseñan-

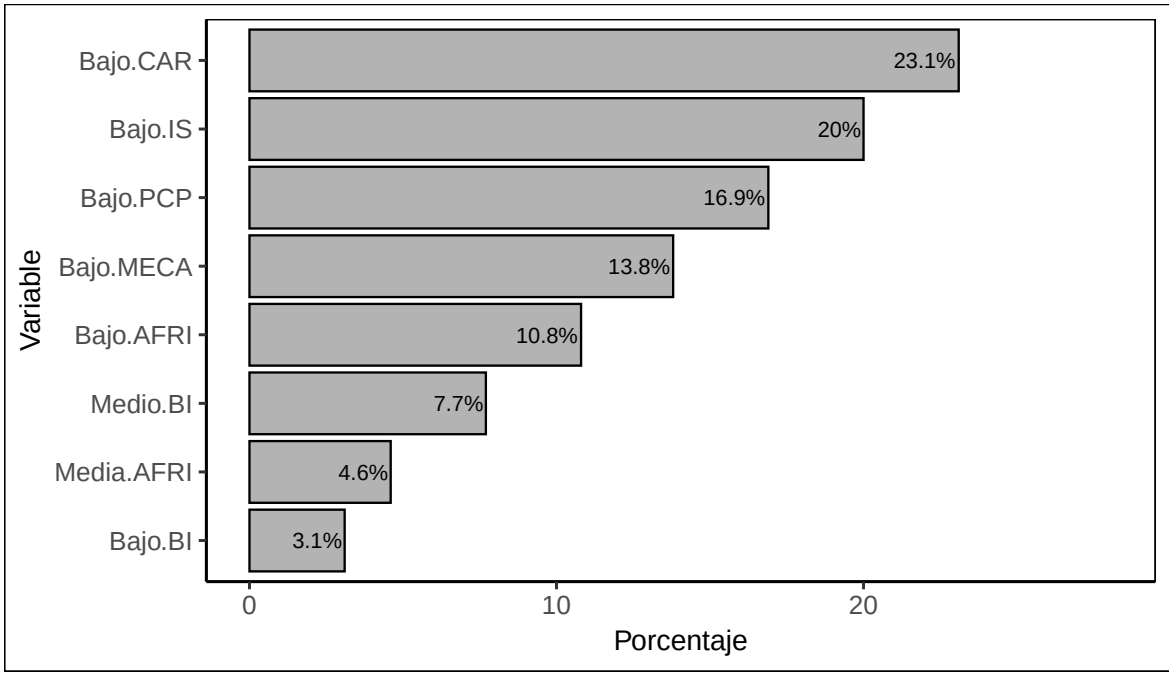


Figura 7.5: Porcentaje de las Variables en las Reglas de Baja Satisfacción

za) y "Bajo.PCP" (Puntualidad y Compromiso del Profesorado), la relación con la satisfacción baja es significativa, aunque menos pronunciada que en otras reglas. Esto sugiere que la metodología de enseñanza y el compromiso docente son áreas que necesitan atención para evitar que los estudiantes se sientan insatisfechos.

Finalmente, las reglas con soporte bajo y confianza moderada, como R24, R30 y R34, señalan que factores como "Bajo.CAR" (Cortesía) o "Bajo.BI" (Biblioteca e Internet) también pueden llevar a una baja satisfacción, aunque de manera menos crítica. Sin embargo, estos factores siguen siendo importantes y deben ser monitoreados y mejorados de forma continua para garantizar una experiencia estudiantil positiva.

A partir de este análisis, se pueden extraer consideraciones clave para políticas institucionales en UniLicungo. Es fundamental que se prioricen las áreas relacionadas con la atención al estudiante, infraestructura, y los recursos educativos como la biblioteca y la internet. Además, la metodología de enseñanza y el compromiso docente deben ser constantemente evaluados para asegurar que no afecten negativamente la satisfacción de los estudiantes. A través de estas acciones, UniLicungo puede trabajar para mejorar la experiencia educativa y aumentar la satisfacción de los estudiantes.

### 7.4.3 *Análisis de las reglas asociadas a la Alta Satisfacción Estudiantil*

En esta sección se presentan los principales resultados derivados del análisis de reglas de asociación que conducen a una alta satisfacción estudiantil. A partir de un total de 153 reglas identificadas, se seleccionaron las veinte más relevantes según el método de ordenamiento multicriterio *uwTOPSIS*, el cual consideró simultáneamente el soporte, la confianza y el numero de atributos en las regras. El gráfico 7.7 muestra esta ordenacion,

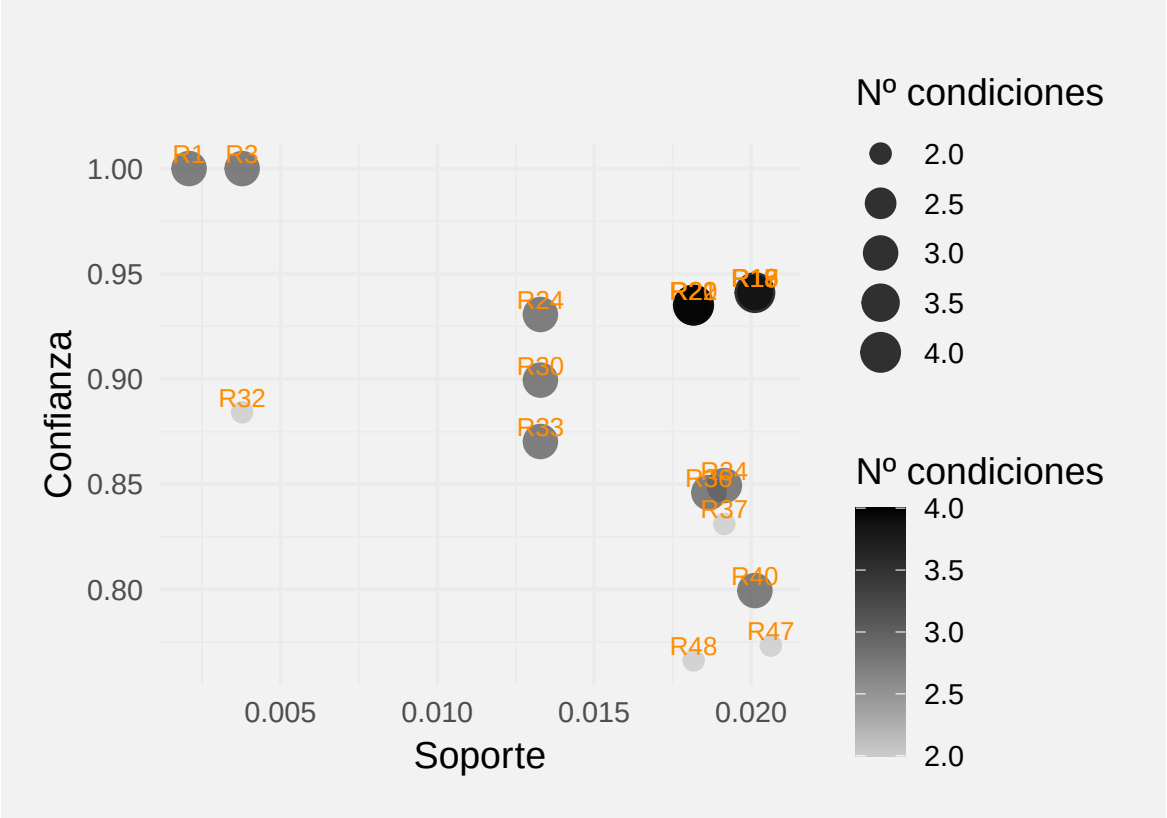


Figura 7.6: Dispersión que relacione el soporte con la confianza de las reglas

mientras que la figura 7.8 presenta las veinte reglas mejor posicionadas con sus confianzas y clasificación uwTOPSIS.

Una observación preliminar del gráfico de dispersión entre soporte y confianza 7.9 revela una distribución asimétrica, con varias reglas presentando niveles máximos de confianza (1.0000) pero soporte reducido (por debajo del 0.01), e outras com soporte moderado, mas leve descenso en la confianza. Esta dinámica sugiere la existencia de configuraciones causales altamente confiables pero poco frecuentes, junto a reglas más generalizables pero con leve pérdida de precisión.

### 7.4.4 *Patrones dominantes en las reglas de Alta Satisfacción*

El mapa de calor 7.10 revela la fuerte presencia de tres dimensiones clave entre las configuraciones causales que conducen a alta satisfacción: Biblioteca e Internet (BI), Metodología de Enseñanza y Calidad Académica (MECA) y Cortesía y Atención al Estudiante (CAR). Esto se confirma también por la figura de frecuencias de aparición de variables 7.11, donde estas dimensiones aparecen de forma reiterada entre los antecedentes de las reglas más significativas.

Entre las reglas de mayor soporte y confianza, destaca la Regla **R99** (*Alto MECA & Alto BI*), con soporte de 0.0545 y confianza de 0.9933. Esta configuración sugiere que la combinación de metodologías de enseñanza bien valoradas y un adecuado acceso a bibliotecas

Tabla 7.1: Reglas de Asociación Seleccionadas

| ID | Regla | Soporte | Confianza |
|---|---|---|---|
| R17 | Bajo.CAR & Bajo.IS ⇒ Bajo.Sat | 0.0201342031 | 0.9410544 |
| R37 | Bajo.CAR & Bajo.PCP ⇒ Bajo.Sat | 0.0191492513 | 0.8310926 |
| R16 | Bajo.MECA & Bajo.CAR & Bajo.IS ⇒ Bajo.Sat | 0.0201342031 | 0.9410544 |
| R18 | Bajo.AFRI & Bajo.CAR & Bajo.IS ⇒ Bajo.Sat | 0.0201218594 | 0.9410203 |
| R20 | Bajo.CAR & Bajo.PCP & Bajo.IS ⇒ Bajo.Sat | 0.0181713775 | 0.9351003 |
| R47 | Bajo.MECA & Bajo.IS ⇒ Bajo.Sat | 0.0206293128 | 0.7732762 |
| R48 | Bajo.PCP & Bajo.IS ⇒ Bajo.Sat | 0.0181713775 | 0.7662345 |
| R24 | Bajo.CAR & Bajo.IS & Bajo.BI ⇒ Bajo.Sat | 0.0132852988 | 0.9306353 |
| R19 | Bajo.MECA & Bajo.AFRI & Bajo.CAR & Bajo.IS ⇒ Bajo.Sat | 0.0201218594 | 0.9410203 |
| R21 | Bajo.MECA & Bajo.CAR & Bajo.PCP & Bajo.IS ⇒ Bajo.Sat | 0.0181713775 | 0.9351003 |
| R22 | Bajo.AFRI & Bajo.CAR & Bajo.PCP & Bajo.IS ⇒ Bajo.Sat | 0.0181590337 | 0.9350591 |
| R30 | Bajo.CAR & Bajo.PCP & Bajo.BI ⇒ Bajo.Sat | 0.0132852988 | 0.8994406 |
| R34 | Bajo.MECA & Bajo.CAR & Bajo.PCP ⇒ Bajo.Sat | 0.0191492513 | 0.8493433 |
| R36 | Bajo.AFRI & Bajo.CAR & Bajo.PCP ⇒ Bajo.Sat | 0.0186541416 | 0.8459605 |
| R32 | Medio.AFRI & Bajo.PCP ⇒ Bajo.Sat | 0.0037779142 | 0.8841316 |
| R33 | Bajo.MECA & Bajo.CAR & Bajo.BI ⇒ Bajo.Sat | 0.0132852988 | 0.8702694 |
| R3 | Medio.AFRI & Bajo.PCP & Medio.BI ⇒ Bajo.Sat | 0.0037779142 | 1.0000000 |
| R40 | Bajo.MECA & Bajo.AFRI & Bajo.IS ⇒ Bajo.Sat | 0.0201218594 | 0.7994211 |
| R1 | Bajo.MECA & Medio.AFRI & Bajo.PCP ⇒ Bajo.Sat | 0.0020860378 | 1.0000000 |

e internet es por sí sola un poderoso predictor de satisfacción, incluso sin necesidad de otras condiciones complementarias.

De modo similar, la Regla **R85** (*Alto MECA & Alto CAR & Alto BI*), con soporte de 0.0427 y confianza de 0.9989, refuerza la importancia de un entorno académico donde la calidad pedagógica se combine con un trato cordial y recursos digitales accesibles. La combinación de estas tres condiciones constituye, en términos prácticos, una tríada estructural de satisfacción: enseñanza sólida, servicios de apoyo tecnológico y clima institucional empático.

*Diversidad estructural y reglas raras pero perfectas*

Un subconjunto interesante está formado por reglas de confianza perfecta (1.0000) pero bajo soporte, como las siguientes:

- **R46**: Medio IS & Alto BI — soporte: 0.0111
- **R36**: Bajo AFRI & Alto BI — soporte: 0.0045
- **R11**: Bajo MECA & Alto CAR — soporte: 0.0040
- **R35**: Bajo MECA & Alto PCP — soporte: 0.0040
- **R77**: Bajo CAR & Alto BI — soporte: 0.0040
- **R50**: Alto PCP & Bajo IS — soporte: 0.0016

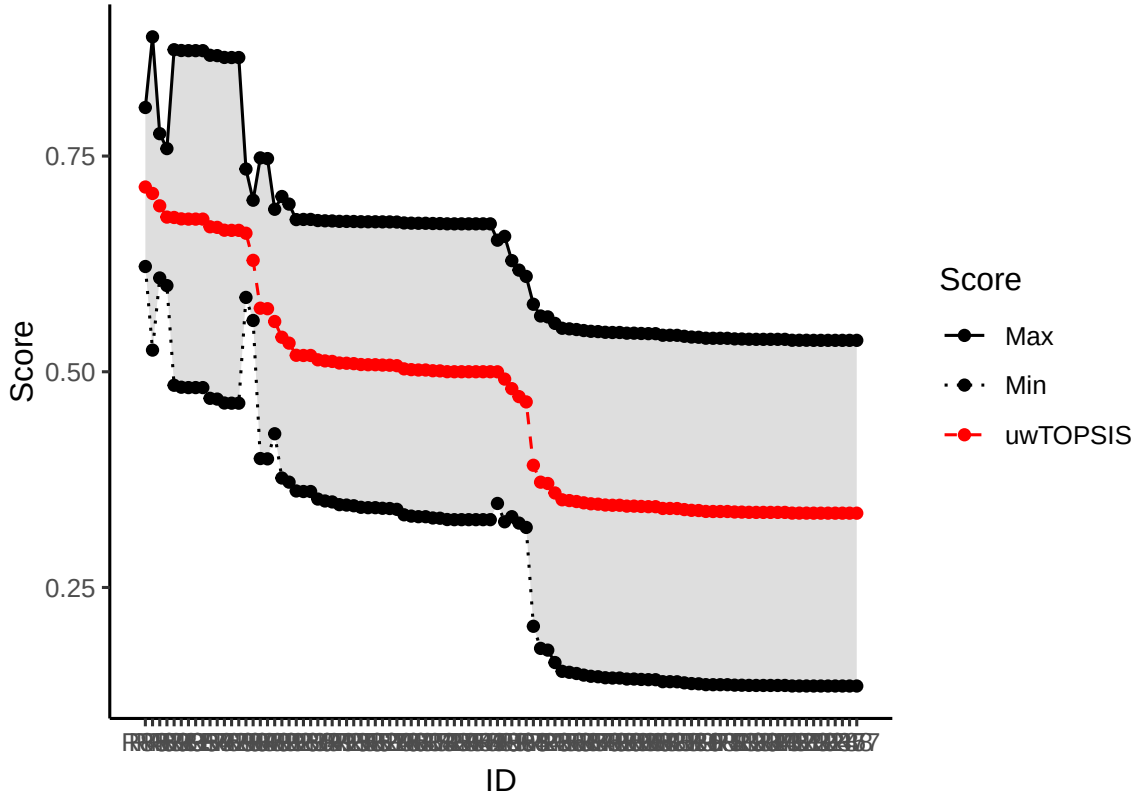


Figura 7.7: uwTOPSIS para Satisfacción posetiva

- **R7**: Alto CAR & Bajo IS — soporte: 0.0014
- **R42**: Alto AFRI & Bajo BI — soporte: 0.0005
- **R68**: Alto MECA & Bajo CAR — soporte: 0.0004
- **R78**: Alto MECA & Bajo IS — soporte: 0.0004

Estas reglas incorporan combinaciones menos frecuentes como, por ejemplo, la Regla R50 (*Alto PCP & Bajo IS*), que sugiere que incluso con infraestructura limitada, el compromiso docente puede ser determinante para lograr satisfacción, aunque en casos aislados. Otras, como R68 (*Alto MECA & Bajo CAR*), exploran la compensación entre dimensiones, donde fortalezas pedagógicas podrían mitigar carencias en atención al estudiante, aunque de manera excepcional.

La existencia de tales reglas plantea implicaciones importantes: mientras algunas condiciones son suficientes por sí solas o en combinaciones mínimas (e.g., MECA y BI), otras pueden ser sustitutas contextuales o incluso mostrar efectos de compensación estructural, desde que en contextos específicos.

*Configuraciones complejas y saturación de condiciones*

Otras reglas destacan por su complejidad combinatoria y alta confianza, como:

- **R86**: Alto MECA & Alto PCP & Alto IS & Alto BI — confianza: 0.9988

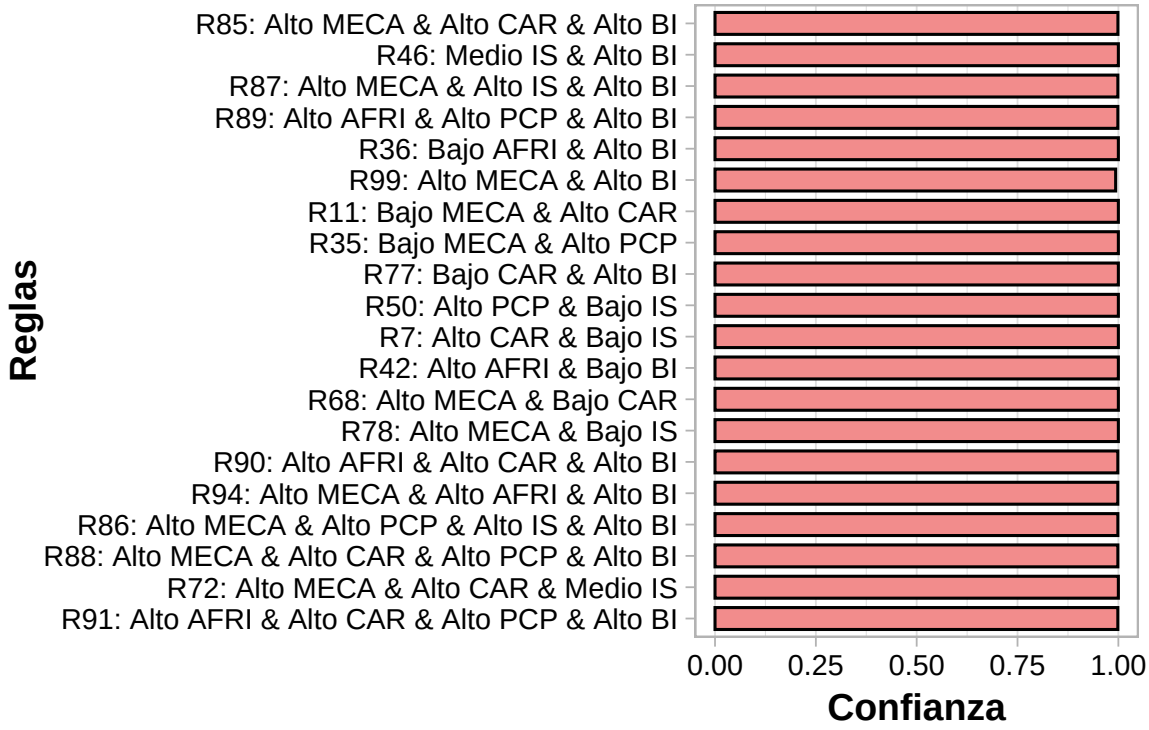


Figura 7.8: Top 20 de las reglas de alta satisfacción

- **R88**: Alto MECA & Alto CAR & Alto PCP & Alto BI — confianza: 0.9988
- **R91**: Alto AFRI & Alto CAR & Alto PCP & Alto BI — confianza: 0.9987
- **R87**: Alto MECA & Alto IS & Alto BI — confianza: 0.9988
- **R89**: Alto AFRI & Alto PCP & Alto BI — confianza: 0.9988
- **R90**: Alto AFRI & Alto CAR & Alto BI — confianza: 0.9987
- **R94**: Alto MECA & Alto AFRI & Alto BI — confianza: 0.9986
- **R72**: Alto MECA & Alto CAR & Medio IS — confianza: 1.0000

Estas configuraciones de alta cardinalidad reflejan escenarios de satisfacción sustentada en múltiples pilares del sistema educativo. Aunque su soporte es menor al de reglas más simples, su presencia reafirma que en contextos de excelencia institucional generalizada (docencia, puntualidad, infraestructura, servicios), la probabilidad de satisfacción plena tiende a ser extremadamente alta.

### 7.4.5 *Conclusión y Recomendaciones*

El análisis de las reglas de asociación difusas, complementado con el método *uwTOPSIS*, permitió identificar tanto los factores que generan alta satisfacción como aquellos que con-

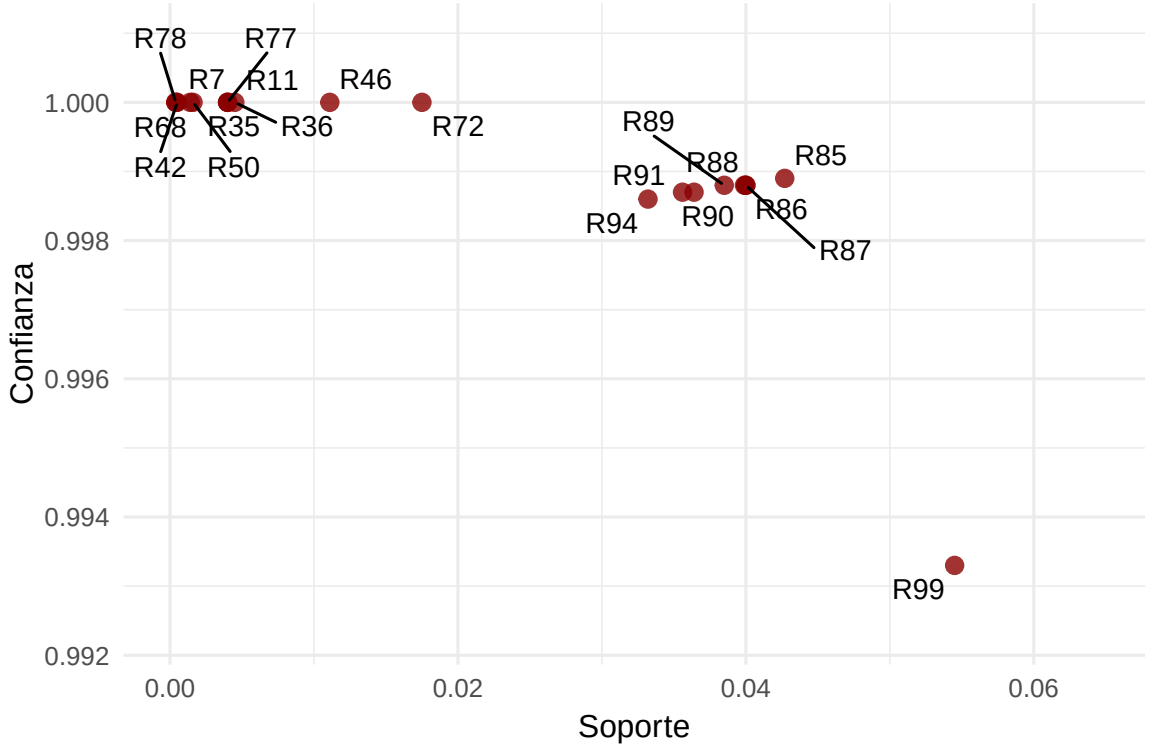


Figura 7.9: Dispersión de Reglas (alta satisfacción) según Soporte y Confianza

tribuyen significativamente a la insatisfacción estudiantil en UniLicungo. Esta doble perspectiva enriquece la interpretación de los resultados y permite trazar un mapa estratégico de intervención institucional.

Las reglas más fuertemente asociadas a la alta satisfacción incluyen la presencia simultánea de buena atención al estudiante, respeto, cumplimiento docente, acceso efectivo a recursos institucionales y un ambiente de acogida. Estas configuraciones no solo presentan alto soporte y confianza, sino que actúan como contrapeso a los factores negativos, demostrando que cuando estas condiciones se cumplen, la percepción general mejora notablemente, incluso cuando algunos aspectos estructurales aún presentan deficiencias.

Por otro lado, las reglas asociadas a la baja satisfacción señalan con recurrencia la falta de cortesía y atención, la impuntualidad docente, la carencia de acogida institucional y problemas en infraestructura y saneamiento. Estas condiciones, cuando combinadas, generan percepciones negativas consistentes en diferentes subgrupos de estudiantes, revelando zonas críticas que requieren intervención inmediata.

Esta lectura cruzada revela que la satisfacción estudiantil no depende de factores aislados, sino de combinaciones específicas que pueden reforzarse mutuamente (en el caso positivo) o agravarse (en el caso negativo). Así, la universidad puede utilizar estos hallazgos para modular su acción institucional, reforzando buenas prácticas existentes y corrigiendo deficiencias estructurales y actitudinales.

**Recomendaciones principales:**

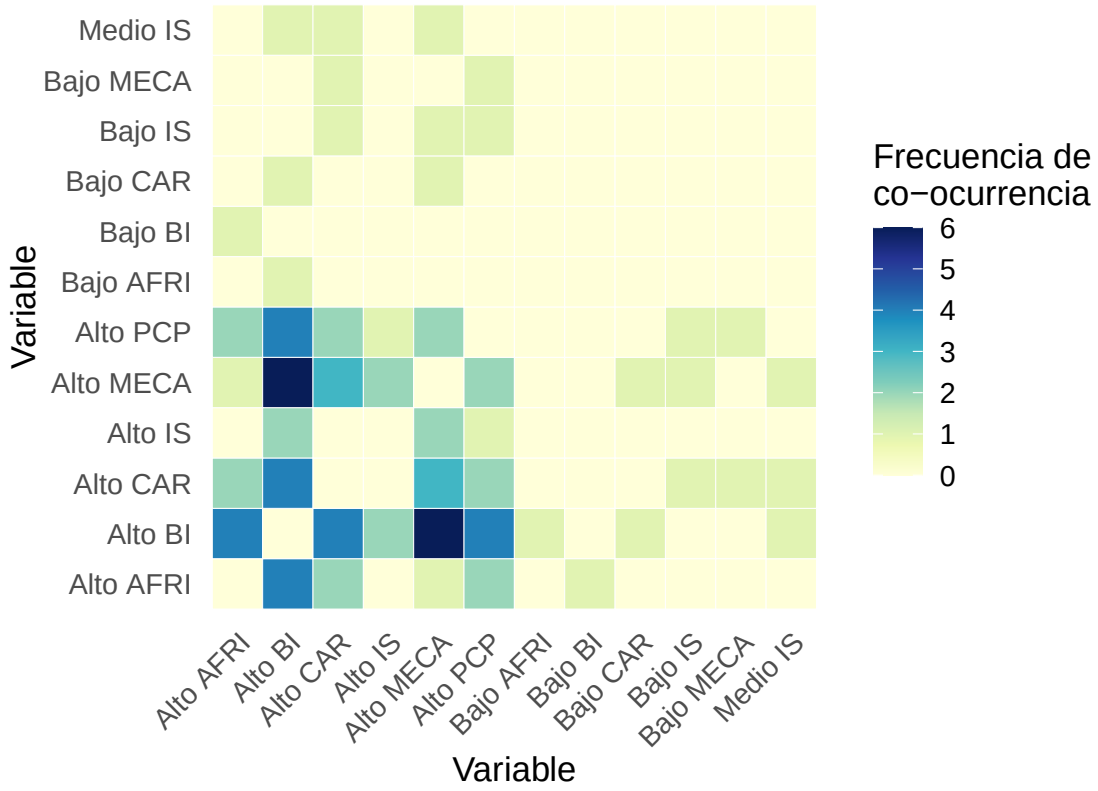


Figura 7.10: Mapa del calor Satisfacción alta

- Consolidar los factores que generan alta satisfacción, como el trato respetuoso, el cumplimiento docente y el acceso a recursos, promoviendo su replicación en todas las unidades académicas.
- Abordar de forma prioritaria los elementos críticos que alimentan la insatisfacción, particularmente la atención deficiente al estudiante, la falta de acogida y las condiciones físicas inadecuadas.
- Implementar formación continua para el personal docente y administrativo centrada en competencias relacionales, comunicación efectiva y compromiso institucional.
- Mejorar la infraestructura sanitaria y de apoyo académico, reconociendo su influencia directa en la percepción de calidad del entorno educativo.
- Diseñar un sistema dinámico de monitoreo de la satisfacción, basado en análisis de datos, que permita identificar configuraciones emergentes de satisfacción o insatisfacción y responder oportunamente.

En suma, el enfoque basado en reglas de asociación difusas y análisis multicriterio permite comprender que la experiencia estudiantil es multifacética y que su mejora exige una mirada sistémica. UniLicungo está en posición de avanzar hacia una gestión centrada en el bienestar del estudiante, guiada por evidencia empírica y orientada a la excelencia institucional.

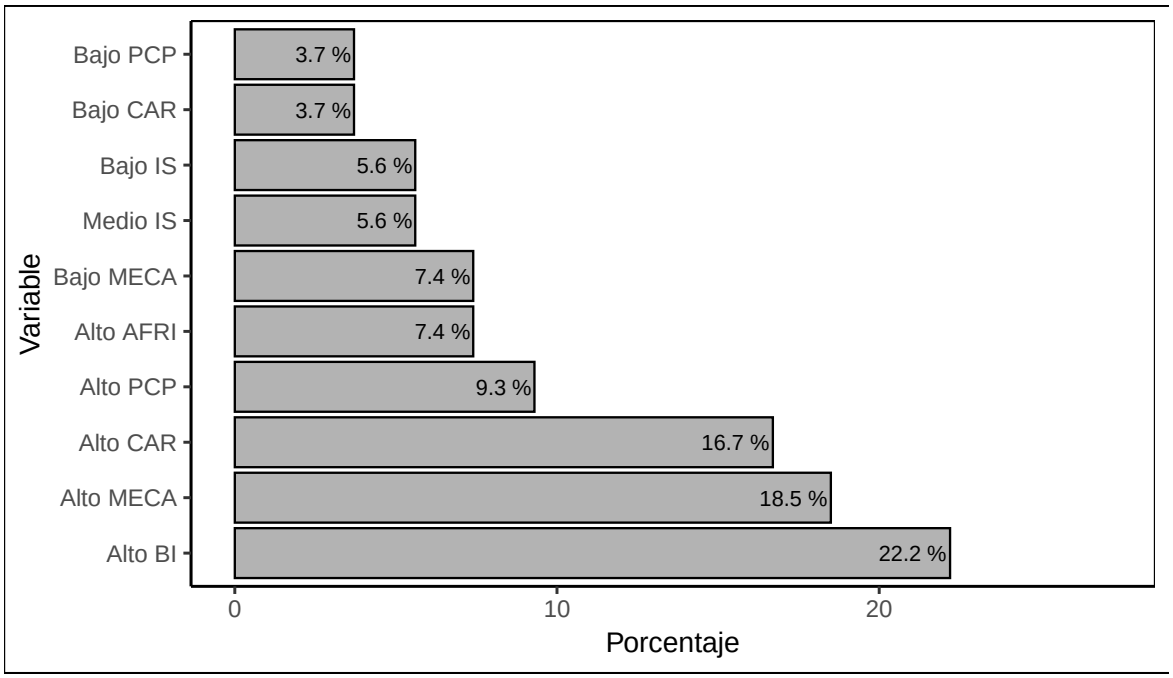


Figura 7.11: Porcentaje de condiciones que aparecen en la base de datos de soluciones alta satisfacción

Tabla 7.2: Regras de associação selecionadas (variáveis recodificadas)

| Regra | Antecedentes | Suporte | Confiança |
|---|---|---|---|
| R85 | Alto MECA & Alto CAR & Alto BI | 0.0427 | 0.9989 |
| R46 | Medio IS & Alto BI | 0.0111 | 1.0000 |
| R87 | Alto MECA & Alto IS & Alto BI | 0.0400 | 0.9988 |
| R89 | Alto AFRI & Alto PCP & Alto BI | 0.0385 | 0.9988 |
| R36 | Bajo AFRI & Alto BI | 0.0045 | 1.0000 |
| R99 | Alto MECA & Alto BI | 0.0545 | 0.9933 |
| R11 | Bajo MECA & Alto CAR | 0.0040 | 1.0000 |
| R35 | Bajo MECA & Alto PCP | 0.0040 | 1.0000 |
| R77 | Bajo CAR & Alto BI | 0.0040 | 1.0000 |
| R50 | Alto PCP & Bajo IS | 0.0016 | 1.0000 |
| R7 | Alto CAR & Bajo IS | 0.0014 | 1.0000 |
| R42 | Alto AFRI & Bajo BI | 0.0005 | 1.0000 |
| R68 | Alto MECA & Bajo CAR | 0.0004 | 1.0000 |
| R78 | Alto MECA & Bajo IS | 0.0004 | 1.0000 |
| R90 | Alto AFRI & Alto CAR & Alto BI | 0.0364 | 0.9987 |
| R94 | Alto MECA & Alto AFRI & Alto BI | 0.0332 | 0.9986 |
| R86 | Alto MECA & Alto PCP & Alto IS & Alto BI | 0.0400 | 0.9988 |
| R88 | Alto MECA & Alto CAR & Alto PCP & Alto BI | 0.0399 | 0.9988 |
| R72 | Alto MECA & Alto CAR & Medio IS | 0.0175 | 1.0000 |
| R91 | Alto AFRI & Alto CAR & Alto PCP & Alto BI | 0.0356 | 0.9987 |

# 8

# CONCLUSIONES Y LÍNEAS FUTURAS

La presente tesis doctoral propuso y validó un enfoque metodológico innovador que integra los principios del Análisis Cualitativo Comparativo (QCA), en sus variantes crisp y fuzzy, con técnicas de Minería de Reglas de Asociación (ARM), incorporando lógica difusa y métodos de ordenación multicriterio. Esta propuesta surgió ante las limitaciones teóricas, metodológicas y computacionales del QCA tradicional, especialmente cuando se aplica a conjuntos de datos complejos, ruidosos y de gran escala, como es común en las ciencias sociales contemporáneas. La investigación comenzó con la identificación de estas debilidades, destacando la limitada escalabilidad del QCA, su sensibilidad a los criterios de calibración y la dificultad para manejar variables lingüísticas de manera representativa. La tesis defendió la integración entre el rigor configuracional del QCA y la capacidad exploratoria del ARM, especialmente en su vertiente fuzzy, como una estrategia prometedora para profundizar en el análisis causal en contextos reales de los fenómenos sociales.

Se estableció la base epistemológica y crítica de las principales metodologías configuracionales (QCA, fsQCA) y de los algoritmos de minería de datos (como Apriori), además de una comparación con métodos estadísticos clásicos como la regresión. También se discutieron los fundamentos de la lógica booleana y difusa, así como las principales dificultades que enfrenta el QCA, como la ambigüedad en la binarización y la rigidez en la definición de patrones causales en datos contradictorios. Con base en ello, se justificó la adopción de modelos híbridos que permitieran mantener la estructura lógica del QCA, pero con una flexibilidad y eficiencia computacional superiores, como el ARM con discretización lingüística.

Se demostró que existe una equivalencia parcial entre el csQCA y la minería de reglas binarias. Fue posible probar matemáticamente que todas las soluciones generadas por el csQCA son subconjuntos de las soluciones obtenidas por el ARM bajo ciertas condiciones de soporte y confianza, lo que refuerza la base teórica de la propuesta. La tesis también introdujo el concepto de reglas de asociación minimizadas y propuso una alternativa al algoritmo de Quine-McCluskey, con el objetivo de reducir la explosión combinatoria típica del ARM. Este enfoque demostró mejoras en escalabilidad y rendimiento computacional, especialmente en la generación de soluciones más parsimoniosas e interpretables.

En relación con las variables continuas y lingüísticas, se mostró que el fsQCA, a pesar de su retórica difusa, opera de forma esencialmente binaria, ya que los valores calibrados son discretizados en la construcción de la tabla de verdad binaria. Para superar esta limitación, se desarrolló un modelo basado en el ARM cuantitativo, con discretización de múltiples intervalos y uso de categorías lingüísticas. Se demostró que este modelo preserva los ma-

tices de los datos sociales y permite la expresión de reglas como: "Si la Educación es Alta y el Ingreso es Medio, entonces la Satisfacción es Alta", algo no directamente viable con el fsQCA. Esta formalización reforzó la tesis de que el fsQCA es un subconjunto del ARM difuso, y que este último ofrece mayor poder expresivo e interpretativo.

La investigación también profundizó en la modelización con reglas difusas, utilizando el algoritmo Fuzzy Apriori. Se propuso un método de extracción y minimización de reglas tipo SI-ENTONCES, capaz de capturar relaciones causales graduales sin comprometer el contenido lingüístico de los datos. Este enfoque superó las ambigüedades del fsQCA y permitió la construcción de modelos más compatibles con la complejidad de las variables sociales. Para manejar la gran cantidad de reglas generadas, se introdujo el método uwTOPSIS como estrategia de ordenación multicriterio, utilizando métricas como cumplimiento, soporte, confianza, lift y cobertura de las reglas. Esta técnica resultó eficaz en la selección de las reglas más relevantes, manteniendo la interpretabilidad de los resultados y ampliando la aplicabilidad práctica del modelo.

La metodología desarrollada se aplicó al contexto real del sistema educativo mozambiqueño, con foco en la Universidad Licungo. A partir de datos de satisfacción estudiantil, se extrajeron reglas difusas que revelaron patrones claros de insatisfacción, asociados a la falta de cortesía, condiciones sanitarias deficientes e infraestructura precaria. Por otro lado, la satisfacción se asoció a la puntualidad del profesorado, la accesibilidad a los servicios y la acogida institucional. Esta aplicación empírica validó los modelos teóricos propuestos y demostró su potencial para orientar decisiones institucionales y políticas públicas orientadas a la mejora de la calidad de la educación superior en Mozambique.

Se concluye, por tanto, que el csQCA es un subconjunto de las soluciones extraídas por el ARM binario y que el fsQCA, aunque pretende operar con lógica difusa, produce soluciones esencialmente discretizadas y binarias. El ARM difuso, por su parte, se reveló como una alternativa más representativa, flexible y escalable para la modelización de la causalidad en las ciencias sociales. La introducción del algoritmo de minimización y del mecanismo de ordenación mediante uwTOPSIS proporcionó una aplicación práctica robusta e interpretable, superando las limitaciones de la explosión combinatoria y del exceso de redundancia en las soluciones generadas. Finalmente, la aplicación práctica en el contexto mozambiqueño no solo validó el modelo, sino que también generó conocimiento útil para la gestión universitaria, reforzando la tesis de que la combinación entre lógica configuracional y minería de datos ofrece un camino sólido e innovador para comprender fenómenos sociales complejos.

## 8.1 LÍNEAS FUTURAS

Como ocurre en toda investigación que abre nuevas perspectivas metodológicas, esta tesis no agota las posibilidades analíticas del enfoque propuesto. Por el contrario, sus resultados abren un camino para futuras líneas de investigación que podrían profundizar o ampliar sus aplicaciones.

Una posible línea de avance consiste en incorporar técnicas avanzadas de aprendizaje automático en el análisis configuracional. En particular, sería de gran interés científico explorar la integración sistemática de algoritmos como árboles de decisión (Decision Trees), Random Forests y redes neuronales (Deep Learning) con metodologías como fsQCA, de forma que se potencie tanto la capacidad predictiva como la interpretabilidad causal. Estas técnicas, al ser capaces de capturar relaciones no lineales, interacciones complejas y estructuras jerárquicas de decisión, pueden complementar el enfoque basado en reglas al proporcionar una visión más amplia de los patrones subyacentes en los datos.

Asimismo, la comparación formal entre fsQCA y métodos basados en ensembles, como Random Forest, podría esclarecer en qué medida las configuraciones causales obtenidas por QCA se corresponden con las rutas de decisión construidas por los modelos supervisados. La aplicación de técnicas como la importancia de variables (feature importance) y la generación de reglas explícitas a partir de árboles también podría facilitar la interpretación de modelos tradicionalmente considerados como "cajas negras".

Desde una perspectiva metodológica, resulta prometedor examinar la posibilidad de construir árboles de decisión orientados a configuraciones causales, que respeten los principios de necesidad, suficiencia y cobertura del QCA, combinando de forma novedosa la lógica booleana con heurísticas del machine learning.

Finalmente, se alienta la investigación futura a establecer puentes entre estas herramientas y el enfoque configuracional clásico, no solo desde el punto de vista computacional, sino también epistemológico, consolidando así un marco híbrido que reúna lo mejor de la lógica cualitativa comparativa con el poder exploratorio del aprendizaje automático contemporáneo.

# BIBLIOGRAFÍA


Aggarwal, C. C., & Yu, P. S. (1998). Online generation of association rules. *Proceedings 14th International Conference on Data Engineering*, 402-411. https://doi.org/10.1109/ICDE.1998.655803

Agrawal, A., & Knoeber, C. R. (1996). Firm Performance and Mechanisms to Control Agency Problems between Managers and Shareholders. *The Journal of Financial and Quantitative Analysis*, *31*(3), 377-397. Consultado el 12 de julio de 2023, desde http://www.jstor.org/stable/2331397

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207-216. https://doi.org/10.1145/170035.170072

Albers, S., Wohlgezogen, F., & Zajac, E. J. (2013). Strategic Alliance Structures: An Organization Design Perspective. *Journal of Management*, *42*(3), 582-614. https://doi.org/10.1177/0149206313488209

Ashrafi, M. Z., Taniar, D., & Smith, K. (2004). A New Approach of Eliminating Redundant Association Rules. En F. Galindo, M. Takizawa & R. Traunmüller (Eds.), *Database and Expert Systems Applications* (pp. 465-474). Springer Berlin Heidelberg.

Ashrafi, M. Z., Taniar, D., & Smith, K. (2005). Redundant Association Rules Reduction Techniques. En S. Zhang & R. Jarvis (Eds.), *AI 2005: Advances in Artificial Intelligence* (pp. 254-263). Springer Berlin Heidelberg.

Bas, M. C., Bolós, V. J., Prieto, Á. E., Rodríguez-Echeverría, R., & Sánchez-Figueroa, F. (2024). A multi-criteria decision support system to evaluate the effectiveness of training courses on citizens' employability. *Applied Intelligence*, *55*(1), 57. https://doi.org/10.1007/s10489-024-05967-0

Becerril-Elías, J., & Merritt, H. (2021). Alianzas para la innovación en organizaciones intensivas en conocimiento: el caso de México. *Revista CEA*, *7*, e1780. https://doi.org/10.22430/24223182.1780

Bedhouche, L., Koalal, M. A., Ben Ayed, A., & Biskri, I. (2023). Efficient Association Rules Minimization Using a Double-Stage Quine-McCluskey-Based Approach. En N. T. Nguyen, J. Botzheim, L. Gulyás, M. Núñez, J. Treur, G. Vossen & A. Kozierkiewicz (Eds.), *Computational Collective Intelligence* (pp. 245-255). Springer Nature Switzerland.

Benítez, R., & Liern, V. (2021). Unweighted TOPSIS: a new multi-criteria tool for sustainability analysis. *International Journal of Sustainable Development & World Ecology*, *28*(1), 36-48. https://doi.org/10.1080/13504509.2020.1778583

Blanca Pérez-Gladish, F. A. F. F., & Zopounidis, C. (2021). MCDM/A studies for economic development, social cohesion and environmental sustainability: introduction. *International Journal of Sustainable Development & World Ecology*, *28*(1), 1-3. https://doi.org/10.1080/13504509.2020.1821257

Buxton, E. K., Vohra, S., Guo, Y., Fogleman, A., & Patel, R. (2019). Pediatric population health analysis of southern and central Illinois region: A cross sectional retrospective study using association rule mining and multiple logistic regression. *Computer Methods and Programs in Biomedicine*, *178*, 145-153. https://doi.org/https://doi.org/10.1016/j.cmpb.2019.06.020

Camarillo-Cisneros, J., Ramirez-Alonso, G., Arzate-Quintana, C., Varela-Rodríguez, H., & Guzman-Pando, A. (2024). MolGC: molecular geometry comparator algorithm for bond length mean absolute error computation on molecules. *Molecular Diversity*, *28*(4), 1925-1945. https://doi.org/10.1007/s11030-024-10945-2

Caren, N., & Panofsky, A. (2005). TQCA: A Technique for Adding Temporality to Qualitative Comparative Analysis. *Sociological Methods and Research*, *34*, 147-172. https://doi.org/10.1177/0049124105277197

Carvalho, L., Almeida, D., Loures, A., Ferreira, P., & Rebola, F. (2024). Quality Education for All: A Fuzzy Set Analysis of Sustainable Development Goal Compliance. *Sustainability*, *16*(12). https://doi.org/10.3390/su16125218

Changpetch, P., & Lin, D. K. (2013). Model selection for logistic regression via association rules analysis. *Journal of Statistical Computation and Simulation*, *83*(8), 1415-1428. https://doi.org/10.1080/00949655.2012.662231

Chen, M.-C. (2007). Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. *Expert Systems with Applications*, *33*(4), 1110-1116. https://doi.org/https://doi.org/10.1016/j.eswa.2006.08.007

Chen, S.-C., Ho, H.-L., Liu, C.-A., Hung, Y.-P., Chiang, N.-J., Chen, M.-H., Chao, Y., & Yang, M.-H. (2025). PIVKA-II as a surrogate biomarker for therapeutic response in Non-AFP-secreting hepatocellular carcinoma. *BMC Cancer*, *25*(1), 199. https://doi.org/10.1186/s12885-025-13568-4

Chiu, A., & Xu, Y. (2023). Bayesian Rule Set: A Quantitative Alternative to Qualitative Comparative Analysis. *The Journal of Politics*, *85*(1), 280-295. https://doi.org/10.1086/720791

Choi, D. H., Ahn, B. S., & Kim, S. H. (2005). Prioritization of association rules in data mining: Multiple criteria decision approach. *Expert Systems with Applications*, *29*(4), 867-878. https://doi.org/https://doi.org/10.1016/j.eswa.2005.06.006

Cornelis, C., Yan, P., Zhang, X., & Chen, G. (2006). Mining Positive and Negative Association Rules from Large Databases. *2006 IEEE Conference on Cybernetics and Intelligent Systems*, 1-6. https://doi.org/10.1109/ICCIS.2006.252251

Cronqvist, L. (2004). *Presentation of TOSMANA: Adding Multi-Value Variables and Visual Aids to QCA* (inf. téc.) (Accessed February 23, 2009). COMPASSS Working Paper no. 2004-20. http://www.compasss.org/files/wpfiles/Cronqvist2004.pdf

Cronqvist, L., & Berg-Schlosser, D. (2008). Multi-Value QCA (mvQCA). En B. Rihoux & C. C. Ragin (Eds.), *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques* (pp. 69-86). Sage.

Cruz, J. (2023). A fuzzy-set qualitative comparative analysis of how corruption, education, inequality and trust in parliament affect voter-turnout. *Crime, Law and Social Change*, *80*, 1-21. https://doi.org/10.1007/s10611-023-10102-0

Das, A. [Amitabha], Ng, W.-K., & Woon, Y.-K. (2001). Rapid association rule mining. *Tenth International Conference on Information and Knowledge Management*, 474-481. https://doi.org/10.1145/502585.502665

Das, A. [Ankita], Debnath, B., Sengupta, A., Das, A., & De, D. (2024). Sustainability analysis of FarmFox IoT device towards Agriculture 5.0. *Environment, Development and Sustainability*. https://doi.org/10.1007/s10668-024-05356-0

Deng, H., Yeh, C.-H., & Willis, R. J. (2000). Inter-company comparison using modified TOPSIS with objective weights. *Computers & Operations Research*, *27*(10), 963-973. https://doi.org/https://doi.org/10.1016/S0305-0548(99)00069-6

Desai, A. (2023). Machine Learning for Economics Research: When What and How? https://arxiv.org/abs/2304.00086

Dinh, H.-V. T., Nguyen, Q. A. T., Phan, M.-H. T., Pham, K. T., Nguyen, T., & Nguyen, H. T. (2021). Vietnamese Students' Satisfaction toward Higher Education Service: The Relationship between Education Service Quality and Educational Outcomes. *European Journal of Educational Research*, *10*(3), 1397-1410. https://doi.org/10.12973/eu-jer.10.3.1397

Dissaneewate, P., Thanavirun, P., Tangjaroenpaisan, Y., & Dissaneewate, K. (2025). External validation and revision of the Lafontaine criteria for unstable distal radius fractures: a retrospective study. *Journal of Orthopaedic Surgery and Research*, *20*(1), 199. https://doi.org/10.1186/s13018-025-03678-9

Dom Luís, A., Benítez, R., & Bas, M. d. C. (2025). Bridging Crisp-Set Qualitative Comparative Analysis and Association Rule Mining: A Formal and Computational Integration. *Mathematics*, *13*(12). https://doi.org/10.3390/math13121939

Dom Luís, A., Benítez, R., Bas, M. d. C., & Cuamba, E. (2025, mayo). *Higher Education and Student Satisfaction: A fsQCA Causal Analysis in an African Context* [Article submitted to Rect@].

Dom Luís, A., & Mulema, S. A. (2021). Avaliação do desempenho dos professores pelos seus alunos. Um Estudo comparativo entre professores de Matemática e de Português da Escola Secundária Geral de Namacurra Sede, Moçambique. *Bolema: Boletim de Educação Matemática*, *35*(70), 1073-1085. https://doi.org/10.1590/1980-4415v35n70a24

Duşa, A. (2019). *QCA with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-75668-4

Elgin, D. J., Erickson, E., Crews, M., Kahwati, L. C., & Kane, H. L. (2024). Applying qualitative comparative analysis in large-N studies: a scoping review of good practices before, during, and after the analytic moment. *Quality & Quantity*, *58*(5), 4241-4256. https://doi.org/10.1007/s11135-024-01849-2

Elmasri, R., & Navathe, S. B. (2017). *Fundamentals of Database Systems* (6th). Pearson Education Ltd. https://asolanki.co.in/wp-content/uploads/2019/02/Fundamentals_of_Database_Systems_6th_Edition-1.pdf

Emmenegger, P., & Klemmensen, R. (2013). What Motivates You? The Relationship between Preferences for Redistribution and Attitudes toward Immigration. *Comparative Politics*, *45*, 227-246. https://doi.org/10.2307/41714184

Emmenegger, P., Schraff, D., & Walter, A. (2014). QCA, the Truth Table Analysis and Large-N Survey Data: The Benefits of Calibration and the Importance of Robustness Tests. Consultado el 9 de octubre de 2024, desde http://www.compasss.org/wpseries/EmmeneggerSchraffWalter2014.pdf

Ermatita, E., & Febry, F. (2017). Implementation of association rule method and TOPSIS method to decision support system for determining epidemic dengue based on risk factors association. *Journal of Theoretical and Applied Information Technology*, *95*(7), 1418-1424.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Fernández, A. C. (2016). El Sentido de la Investigación Cualitativa. *Escuela Abierta*, *19*, 33-48.

Fernández-Portillo, L. A., Demir, G., Sianes, A., & Santos-Carrillo, F. (2024). Setting a shared development agenda: prioritizing the sustainable development goals in the Dominican Republic with fuzzy-LMAW. *Scientific Reports*, *14*(1), 12146. https://doi.org/10.1038/s41598-024-62790-w

Finn, V. (2022). A qualitative assessment of QCA: method stretching in large-N studies and temporality. *Quality & Quantity*, *56*, 3815-3830. https://doi.org/10.1007/S11135-021-01278-5/FIGURES/1

Fiss, P. C. (2007). A Set-Theoretic Approach to Organizational Configurations. *The Academy of Management Review*, *32*(4), 1180-1198. Consultado el 12 de julio de 2023, desde http://www.jstor.org/stable/20159362

Fiss, P. C., Sharapov, D., & Cronqvist, L. (2013). Opposites Attract? Opportunities and Challenges for Integrating Large-N QCA and Econometric Analysis. *Political Research Quarterly*, *66*(1), 191-198. Consultado el 5 de diciembre de 2024, desde http://www.jstor.org/stable/23563602

Freyburg, T., & Garbe, L. (2018). Authoritarian Practices in the Digital Age | Blocking the Bottleneck: Internet Shutdowns and Ownership at Election Times in Sub-Saharan Africa. *International Journal of Communication*, *12*(0). https://ijoc.org/index.php/ijoc/article/view/8546

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, *38*(3), 9. https://doi.org/http://doi.acm.org/10.1145/1132960.1132963

Greckhamer, T., Misangyi, V. F., & Fiss, P. C. (2013). The Two QCAs: From a Small-N to a Large-N Set Theoretic Approach. En *Configurational Theory and Methods in Organizational Research* (pp. 49-75, Vol. 38). Emerald Group Publishing Limited. https://doi.org/10.1108/S0733-558X(2013)0000038007

Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Datasets. *Journal of Machine Learning Research*, *12*, 1977-1981. https://jmlr.csail.mit.edu/papers/v12/hahsler11a.html

Hahsler, M., Gruen, B., & Hornik, K. (2005). arules – A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, *14*(15), 1-25. https://doi.org/10.18637/jss.v014.i15

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *2000 ACM SIGMOD International Conference on Management of Data*, 1-12. https://doi.org/10.1145/342009.335372

Hong, T.-P., Kuo, C.-S., & Chi, S.-C. (2001). Trade-off Between Computation Time and Number of Rules for Fuzzy Mining from Quantitative Data. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *9*, 587-604. https://doi.org/10.1142/S0218488501001071

Hwang, C.-L., & Yoon, K. (1981). Methods for Multiple Attribute Decision Making. En *Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey* (pp. 58-191). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-48318-9_3

Jacquet-Lagreze, E., & Siskos, J. (1982). Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. *European Journal of Operational Research*, *10*(2), 151-164. https://doi.org/https://doi.org/10.1016/0377-2217(82)90155-2

Jang, J.-S. R., & Gulley, N. (1995). *The Fuzzy Logic Toolbox for use with MATLAB*. The MathWorks, Inc. Natick, Massachusetts.

Khedr, A., Ramadan, R., & Abdel-Mageid, S. (2012). QMR: QUINE-MCCLUSKEY FOR RULE MINIMIZATION IN RULE-BASED SYSTEMS. *International Journal of Intelligent Computing and Information Science*.

Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., & Lipsitch, M. (2020). Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period [Epub 2020 Apr 14]. *Science*, *368*(6493), 860-868. https://doi.org/10.1126/science.abb5793

Kraus, S., Ribeiro-Soriano, D., & Schüssler, M. (2018). Fuzzy-set qualitative comparative analysis (fsQCA) in entrepreneurship and innovation research – the rise of a method. *International Entrepreneurship and Management Journal*, *14*(1), 15-33. https://doi.org/10.1007/s11365-017-0461-8

Kumar, S., Sahoo, S., Lim, W. M., Kraus, S., & Bamel, U. (2022). Fuzzy-set qualitative comparative analysis (fsQCA) in business and management research: A contemporary overview. *Technological Forecasting and Social Change*, *178*, 121599. https://doi.org/https://doi.org/10.1016/j.techfore.2022.121599

Kuok, C. M., Fu, A., & Wong, M. H. (1998). Mining fuzzy association rules in databases. *SIGMOD Rec.*, *27*(1), 41-46. https://doi.org/10.1145/273244.273257

Liao, H. (2025). E-commerce Live-Streaming Platform and Decision Support System Based on Fuzzy Association Rule Mining. *International Journal of Computational Intelligence Systems*, *18*(1), 41. https://doi.org/10.1007/s44196-025-00744-4

Liern, V., & Pérez-Gladish, B. (2022). Multiple criteria ranking method based on functional proximity index: unweighted TOPSIS. *Annals of Operations Research*, *311*(2), 1099-1121. https://doi.org/10.1007/s10479-020-03718-1

Liu, B., Hu, M., & Hsu, W. (2000). Multi-level organization and summarization of the discovered rules. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 208-217. https://doi.org/10.1145/347090.347128

Liu, Y., Mezei, J., Kostakos, V., & Li, H. (2017). Applying configurational analysis to IS behavioural research: a methodological alternative for modelling combinatorial complexities. *Information Systems Journal*, *27*, 59-89. https://doi.org/10.1111/isj.12094

Mahoney, J., & Rueschemeyer, D. (Eds.). (2003). *Comparative Historical Analysis in the Social Sciences*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511803963

McCluskey, E. J. (1956). Minimization of Boolean functions. *The Bell System Technical Journal*, *35*(6), 1417-1444. https://doi.org/10.1002/j.1538-7305.1956.tb03835.x

Mendel, J. M., & Korjani, M. M. (2013). Theoretical aspects of Fuzzy Set Qualitative Comparative Analysis (fsQCA) [Prediction, Control and Diagnosis using Advanced Neural Computations]. *Information Sciences*, *237*, 137-161. https://doi.org/https://doi.org/10.1016/j.ins.2013.02.048

Mendel, J. M., & Korjani, M. M. (2018). A new method for calibrating the fuzzy sets used in fsQCA. *Information Sciences*, *468*, 155-171. https://doi.org/https://doi.org/10.1016/j.ins.2018.07.050

Mguirris, I., Amdouni, H., & Gammoudi, M. M. (2015). A new validation method of fuzzy association rules based on the Structural Equation Modeling. *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 624-633. https://doi.org/10.1109/FSKD.2015.7382015

Mill, J. S. (1843). *A System of Logic, Ratiocinative and Inductive* [First published in 1843]. Longmans, Green, Reader, Dyer. https://doi.org/10.1017/CBO9781139149839

Müller, A. C., Gil-Lafuente, J., & Ferrer-Comalat, J. C. (2024). Colour Choice as a Strategic Instrument in Neuromarketing. *Mathematics*, *12*(14). https://doi.org/10.3390/math12142212

Oana, I.-E., Schneider, C. Q., & Thomann, E. (2021). *Qualitative Comparative Analysis Using R: A Beginner's Guide*. Cambridge University Press. https://doi.org/10.1017/9781009006781

Omar Elfarouk, K. Y. W., & Wong, W. P. (2022). Multi-objective optimization for multi-echelon, multi-product, stochastic sustainable closed-loop supply chain. *Journal of Industrial and Production Engineering*, *39*(2), 109-127. https://doi.org/10.1080/21681015.2021.1963338

Pasaribu, I. M., Gultom, A., & Pasaribu, N. M. (2020). School Facilities and Infrastructure Management System to Comply the National Standar for Education. *Proceedings of the 5th Annual International Seminar on Transformative Education and Educational Leadership (AISTEEL 2020)*, 447-453. https://doi.org/10.2991/assehr.k.201124.091

Pennings, P. (2000). The impact of democratic quality on the link between political participation and political trust. *Political Studies*, *48*(1), 49-64. https://doi.org/10.1111/1467-9248.00251

Quine, W. V. (1952). The Problem of Simplifying Truth Functions. *The American Mathematical Monthly*, *59*(8), 521-531. https://doi.org/10.1080/00029890.1952.11988183

Quine, W. V. (1955). A Way to Simplify Truth Functions. *The American Mathematical Monthly*, *62*(9), 627-631. Consultado el 20 de diciembre de 2024, desde http://www.jstor.org/stable/2307285

Ragin, C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. University of California Press.

Ragin, C. (2000). *Fuzzy-Set Social Science*. University of Chicago Press.

Ragin, C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. University of Chicago Press. https://books.google.es/books?hl=es&lr=&id=WUj9yT5zAiIC

Ragin, C., & Rihoux, B. (2004). Qualitative comparative analysis (CQA): State of the art and prospects. *Qualitative & Multi-Method Research*, *2*(2), 3-13. https://doi.org/10.5281/zenodo.998222

Ragin, C., Shulman, D., Weinberg, A. S., & Gran, B. K. (2003). Complexity, Generality, and Qualitative Comparative Analysis. *Field Methods*, *15*, 323-340. https://doi.org/10.1177/1525822x03257689

Ragin, C., & Sonnett, J. (2005). Between Complexity and Parsimony: Limited Diversity, Counterfactual Cases, and Comparative Analysis. En S. Kropp & M. Minkenberg (Eds.), *Vergleichen in der Politikwissenschaft* (pp. 180-197). VS Verlag für Sozialwissenschaften. https://doi.org/{10.1007/978-3-322-80441-9_9}

Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: insights from moodle log data. *Scientific Reports*, *15*(1), 9840. https://doi.org/10.1038/s41598-025-93918-1

Remo-Diez, N., Mendaña-Cuervo, C., & Arenas-Parra, M. (2024). A Fuzzy-Set Qualitative Comparative Analysis for Understanding the Interactive Effects of Good Governance Practices and CEO Profiles on ESG Performance. *Mathematics*, *12*(17). https://doi.org/10.3390/math12172726

Rihoux, B. (2006). Qualitative Comparative Analysis (QCA) and Related Systematic Comparative Methods: Recent Advances and Remaining Challenges for Social Science Research. *International Sociology*, *21*(5), 679-706. https://doi.org/10.1177/0268580906067836

Rihoux, B., & Ragin, C. (2009). *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. SAGE Publications, Inc. https://doi.org/https://doi.org/10.4135/9781452226569

Roig-Tierno, N., González-Cruz, T., & Llopis, J. (2017). An overview of qualitative comparative analysis: A bibliometric analysis. *Journal of Innovation and Knowledge*, *2*. https://doi.org/10.1016/j.jik.2016.12.002

Roy, B. (1996). *Multicriteria Methodology for Decision Aiding* (Vol. 12). Kluwer Academic Publishers. https://doi.org/10.1007/978-1-4757-2500-1

Rubinson, C. (2013). Contradictions in fsQCA. *Quality & Quantity*, *47*(5), 2847-2867. https://doi.org/10.1007/s11135-012-9694-3

Santika, F., Sowiyah, Pangestu, U., & Nurahlaini, M. (2021). School Facilities and Infrastructure Management in Improving Education Quality. *International Journal of Research and Innovation in Social Science*, *5*(6), 280-285. https://doi.org/10.47772/ijriss.2021.5612

Saunders, A., & Luukkanen, J. (2022). Sustainable development in Cuba assessed with sustainability window and doughnut economy approaches. *International Journal of Sustainable Development & World Ecology*, *29*(2), 176-186. https://doi.org/10.1080/13504509.2021.1941391

Schneider, C. Q., & Wagemann, C. (2012). *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9781139004244

Schwellnus, G. (2013). Eliminating the influence of irrelevant cases on the consistency and coverage of necessary and sufficient conditions in fuzzy-set QCA. *7th ECPR General Conference, Bordeaux 2013*.

Seawright, J. (2005a). Assumptions, causal inference, and the goals of QCA [Accessed: 2025-03-28]. *Studies in Comparative International Development*, *40*(1), 39-42. https://doi.org/10.1007/BF02686287

Seawright, J. (2005b). Qualitative comparative analysis vis-à-vis regression. *Studies in Comparative International Development*, *40*(1), 3-26. https://doi.org/10.1007/BF02686284

Sinthupundaja, J., Chiadamrong, N., & and, Y. K. (2019). Internal capabilities, external cooperation and proactive CSR on financial performance. *The Service Industries Journal*, *39*(15-16), 1099-1122. https://doi.org/10.1080/02642069.2018.1508459

Srikant, R., & Agrawal, R. (1997). Mining generalized association rules [Data Mining]. *Future Generation Computer Systems*, *13*(2), 161-180. https://doi.org/https://doi.org/10.1016/S0167-739X(97)00019-8

Tan, P., Steinbach, M., & Kumar, V. (2014). *Introduction to Data Mining*. Pearson. https://books.google.es/books?id=BExxngEACAAJ

Toloo, M., Sohrabi, B., & Nalchigar, S. (2009). A new method for ranking discovered rules from data mining by DEA. *Expert Systems with Applications*, *36*(4), 8503-8508. https://doi.org/https://doi.org/10.1016/j.eswa.2008.10.038

Valero-Moreno, S., Castillo Corullón, S., Montoya-Castilla, I., & Perez-Marin, M. (2020). Primary ciliary dyskinesia and psychological well-being in adolescence. *PLOS ONE*, *15*, e0227888. https://doi.org/10.1371/journal.pone.0227888

Vassinen, A. (2012). *Configurational Explanation of Marketing Outcomes: A Fuzzy-Set Qualitative Comparative Analysis Approach* [PhD thesis]. School of Economics, Aalto Univeristy [Available at https://urn.fi/URN:ISBN:978-952-60-4574-0].

Verlinde, H., De Cock, M., & Boute, R. (2006). Fuzzy versus quantitative association rules: A fair data-driven comparison. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, *36*, 679-84. https://doi.org/10.1109/TSMCB.2005.860134

Vis, B. (2012). The Comparative Advantages of fsQCA and Regression Analysis for Moderately Large-N Analyses. *Sociological Methods and Research*, *40*, 168-198. https://doi.org/10.1177/0049124112442142

Wang, L., Meng, J., Xu, P., & Peng, K. (2018). Mining temporal association rules with frequent itemsets tree. *Applied Soft Computing*, *62*, 817-829. https://doi.org/https://doi.org/10.1016/j.asoc.2017.09.013

Wibowo, Y. (2022). Literature review customer satisfaction determination and level of complaint: Product quality and service quality. *Dinasti International Journal of Digital Business Management*, *3*(4), 683-692. https://doi.org/10.31933/dijdbm.v3i4.1268

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd). Morgan Kaufmann Publishers Inc.

Woodside, A. G. (2013). Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory. *Journal of Business Research*, *66*(4), 463-472. https://doi.org/https://doi.org/10.1016/j.jbusres.2012.12.021

Woodside, A. G., & Zhang, M. (2012). Identifying X-Consumers Using Causal Recipes: "Whales.ªnd "Jumbo ShrimpsÇasino Gamblers. *Journal of Gambling Studies*, *28*, 13-26. https://doi.org/10.1007/S10899-011-9241-5

Yacoubi, S., Manita, G., & Korbaa, O. (2022). A Multiobjective Crystal Optimization-based association rule mining enhanced with TOPSIS for predictive maintenance analysis [Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022]. *Procedia Computer Science*, *207*, 2782-2793. https://doi.org/https://doi.org/10.1016/j.procs.2022.09.336

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, *8*(3), 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zaki, M. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, *12*(3), 372-390. https://doi.org/10.1109/69.846291

Zaki, M. J. (2000). Generating non-redundant association rules. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 34-43. https://doi.org/10.1145/347090.347101

Zaki, M. J., & Hsiao, C.-J. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. *2002 SIAM International Conference on Data Mining (SDM)*, 457-473. https://doi.org/10.1137/1.9781611972726.27

Zavyalova, E., Sokolov, D., & Lisovskaya, A. (2020). Agile vs traditional project management approaches. *International Journal of Organizational Analysis*, *28*(5), 1095-1112. https://doi.org/10.1108/IJOA-08-2019-1857

Zeleny, M. (2011). Multiple Criteria Decision Making (MCDM): From Paradigm Lost to Paradigm Regained? *Journal of Multi-Criteria Decision Analysis*, *18*(1-2), 77-89. https://doi.org/https://doi.org/10.1002/mcda.473

Zende, R. V., & Pawade, R. S. (2024). A novel Zende's-TOPSIS method towards estimation of measurement uncertainty in hole diameters. *Engineering Research Express*, *6*(2), 025415. https://doi.org/10.1088/2631-8695/ad45b7

Zhang, Z., & Guo, C. (2009). Ranking association rules by TOPSIS method with combination weights. *Knowledge and Systems Sciences. Proceedings of KSS2009*.

Zhao, J., & Yan, C. (2020). User acceptance of information feed advertising: A hybrid method based on sem and qca. *Future Internet*, *12*, 1-17. https://doi.org/10.3390/fi12120209

Zhao, Y., Wang, L., Tang, H., & Zhang, Y. (2020). Electronic Word-of-Mouth and Consumer Purchase Intentions in Social E-Commerce. *Electronic Commerce Research and Applications*, *41*, 100980. https://doi.org/10.1016/j.elerap.2020.100980

Zientara, P., Jażdżewska-Gutta, M., Bąk, M., & Zamojska, A. (2024). What drives tourists' sustainable mobility at city destinations? Insights from ten European capital cities. *Journal of Destination Marketing & Management*, *33*, 100931. https://doi.org/https://doi.org/10.1016/j.jdmm.2024.100931

# A

# SCRIPTS DE IMPLEMENTACIÓN DE QCA Y REGLAS DE ASOCIACIÓN

## A.1 SCRIPT DEL PRIMER EJEMPLO COMPARATIVO ENTRE CSQCA Y REGLAS DE ASOCIACIÓN BINARIAS

En este apartado contiene el código desarrollado en lenguaje R que fue utilizado en el primer ejemplo comparativo de esta tesis entre el método csQCA y las Reglas de Asociación Binarias, tanto positivas como negativas, para el ejemplo de bloqueos de internet en periodos electorales.

Este ejemplo tiene como base un conjunto de datos compuesto por tres condiciones causales y una condición resultado, todas con valores binarios (0 o 1). El objetivo es explorar cómo cada método identifica configuraciones o reglas que explican la presencia del resultado IS = 1, y comparar la capacidad explicativa, la redundancia y la parsimonia de ambos enfoques.

El código está dividido en tres grandes bloques funcionales: Primero, se aplican los procedimientos del csQCA.

Segundo, se realiza la transformación de los datos a un formato de transacciones, compatible con el algoritmo apriori, que se utiliza comúnmente para minería de reglas de asociación y se extraen las reglas de asociación que implican la presencia del resultado IS

Código 1: Ejemplo práctico de análisis comparativo entre csQCA y Reglas de Asociación Binarias mediante R.

```r
# Load required libraries
library(haven)
library(arules)
library(QCA)
library(dplyr)
library(readxl)
library(tidyr)
library(xtable)
library(tibble)

# Select relevant binary variables
Ds <- rs[, c("ISP", "AT", "EV", "IS")]
Ds <- as.data.frame(Ds)

# Check superset conditions for consistency and coverage
superSubset(Ds, outcome = "IS", incl.cut = 0.9)

# Build the truth table for QCA
Tr <- truthTable(Ds, outcome = "IS", complete = TRUE, sort.by = "incl, n+")
Tr

# ---------- Solution Minimization (QCA) ----------

```

```r
# Parsimonious solution
PG <- minimize(Tr, include = "?", details = TRUE, show.cases = FALSE)
PG

# Intermediate solution (specifying expected direction)
IG <- minimize(Tr, include= "?", dir.exp = "1,1,1", details = TRUE, show.cases =
    FALSE)
IG

# Conservative solution (no assumptions)
CL <- minimize(Tr, details = TRUE, show.cases = TRUE)
CL

# ---------- Association Rule Mining and Redundancy Elimination ----------

# Add an ID column for transformation
ID <- 1:8
NV <- cbind(ID = ID, Ds)

# Convert dataset to long format (tidy format)
datos_NV <- pivot_longer(NV, cols = -1, names_to = "Descripcion", values_to = "valor"
    )

# Replace variable names with lowercase when value is 0
datos_NV$Descripcion[datos_NV$valor == 0] <- tolower(datos_NV$Descripcion[datos_NV$
    valor == 0])

# Keep only the ID and condition names
datos_NV <- datos_NV %>% select(c(ID, Descripcion))

# Group variables by observation ID for transaction conversion
data_lists1 <- split(datos_NV$Descripcion, datos_NV$ID)

# Convert the data to a transactional dataset (class 'transactions')
datos_trxx_mig1 <- as(data_lists1, "transactions")

# Summary of the transactional dataset
summary(datos_trxx_mig1)

# Extract frequent itemsets with minimum support
support_RR1 <- apriori(
  datos_trxx_mig1,
  parameter = list(
    target = "frequent itemsets",
    supp = 0.1
  ),
  appearance = list(
    items = c("IS")
  )
)

# Show summary of frequent itemsets
summary(support_RR1)

# Inspect rules with right-hand side (rhs) fixed to outcome "IS"
inspect(
  apriori(
    datos_trxx_mig1,
```

```
    appearance = list(rhs = "IS"),
    parameter = list(support = 0.0001, confidence = 1)
  )
)

# Generate rules (without redundant ones if implicitly minimized)
reglas <- apriori(
  datos_trxx_mig,
  appearance = list(rhs = "IS"),
  parameter = list(support = 0.0001, confidence = 1)
)

# Print and inspect the rules
print(reglas)
reglas
```

## A.2 SCRIPT PARA EL ANÁLISIS COMPARATIVO: CSQCA VS. REGLAS DE ASOCIACIÓN BINARIAS (EJEMPLO 2)

El siguiente script corresponde al segundo ejemplo presentado en el Capítulo 3 de esta tesis, el cual aborda el caso de la oposición a la inmigración en Europa. Utilizando una base de datos con condiciones causales calibradas de forma binaria, este ejemplo tiene como objetivo comparar empíricamente los resultados obtenidos mediante el Análisis Cualitativo Comparativo de conjunto nítido (csQCA) y aquellos generados a través de la minería de reglas de asociación binarias, incluyendo tanto reglas positivas como negativas. La implementación computacional muestra paso a paso cómo se procesan, analizan y minimizan los datos en cada enfoque, lo que permite identificar similitudes y el poder explicativo de las soluciones obtenidas.

Código 2: Código R aplicado al ejemplo de oposición a la inmigración en Europa.

```
# Load required libraries
library(haven)     # For reading SPSS, Stata, and SAS files
library(arules)    # For mining association rules and frequent itemsets
library(QCA)       # For Qualitative Comparative Analysis (QCA)
library(dplyr)     # For data manipulation
library(readxl)    # For reading Excel files
library(tidyr)     # For reshaping data
library(xtable)    # For LaTeX table output

# Read the dataset containing calibrated crisp-set values
mig <- read_excel("C:/Users/Dom Luis JR/Desktop/NOVOS Dados CRISP ARTIGO - 1/
    immigration.xlsx")

# Rename columns for better interpretation
colnames(mig) <- c("LOWEDU","UNITY","NOFRIEND","THREAT","MAN","OPPOSITION")

mig <- as.data.frame(mig)  # Ensure data is in data.frame format

# --- QCA Section ---

# Test necessity relations: Is OPPOSITION necessary for other conditions?
pof("LOWEDU <= OPPOSITION", data = mig)
pof("UNITY <= OPPOSITION", data = mig)
pof("NOFRIEND <= OPPOSITION", data = mig)
pof("THREAT <= OPPOSITION", data = mig)
```

```r
pof("MAN <= OPPOSITION", data = mig)

# Identify potential sufficient condition sets using superSubset
superSubset(mig, outcome = "OPPOSITION", incl.cut = 0.9)

# Build the truth table
TVmig <- truthTable(mig, outcome = "OPPOSITION", complete = TRUE, sort.by = "incl, n+
    ")
TVmig

# Minimize the truth table to get simplified csQCA solutions
MigSol <- minimize(TVmig, datails = TRUE)


# --- Binary Association Rules Section ---

# Create an ID column for transaction identification
ID <- 1:2223
Nueva_mig <- cbind(ID = ID, mig)

# Convert data to long format for transaction transformation
datos_en_long_mig <- pivot_longer(Nueva_mig, cols = -1, names_to = "Descripcion",
    values_to = "valor")

# Convert 0-values to lowercase to distinguish presence/absence in rules
datos_en_long_mig$Descripcion[datos_en_long_mig$valor == 0] <- tolower(datos_en_long_
    mig$Descripcion[datos_en_long_mig$valor==0])
datos_en_long_mig <- datos_en_long_mig %>% select(c(ID, Descripcion))

# Split data by ID for transaction creation
data_lists = split(datos_en_long_mig$Descripcion, datos_en_long_mig$ID)

# Convert the list into a transaction object
datos_trxx_mig = as(data_lists, "transactions")

# Summary of transactions (number of items, etc.)
summary(datos_trxx_mig)

# Generate frequent itemsets with minimum support of 10%
support_RR = apriori(datos_trxx_mig,
                     parameter = list(target = "frequent itemsets", supp = 0.1),
                     appearance = list(items = c("OPPOSITION"))
)

# Generate rules with 100% confidence and very low support
reglas <- apriori(datos_trxx_mig,
                  appearance = list(rhs = "OPPOSITION"),
                  parameter = list(support = 0.0001, confidence = 1)
)

# Convert rules to a readable data frame
rulesdf_mig <- DATAFRAME(reglas)
view(rulesdf_mig)

xtable(rulesdf_mig)  # Optional: Export rules as LaTeX table


# --- Rule Minimization Section ---
```

```r

reglas <- list()   # Initialize list to store binary vectors of rules

for(i in 1:nrow(rulesdf_mig)){
  lhs_end <- rulesdf_mig$LHS[i]
  regla1 <- as.character(lhs_end)

  # Split items and clean braces
  regla_temp1 <- unlist(strsplit(regla1, ","))
  regla_temp2 <- gsub("\\{","",regla_temp1)
  regla_temp3 <- gsub("\\}","",regla_temp2)

  # Convert rule into binary vector: 1 = positive (UPPER), 0 = negative (lower)
  regla_bin <- as.numeric(sapply(regla_temp3, function(x) identical(x, toupper(x))))
  names(regla_bin) <- toupper(regla_temp3)
  reglas[[i]] <- regla_bin
}

# Function to group rules by number of conditions (length)
group_vectors_by_length <- function(vector_list) {
  grouped_vectors <- split(vector_list, sapply(vector_list, length))
  return(grouped_vectors)
}

reglas_by_length <- group_vectors_by_length(reglas)

# Initialize final minimized rule list
sizes <- names(reglas_by_length)
N <- length(sizes)
reglas_def <- vector("list", length = N)
names(reglas_def) <- sizes

# Begin minimization process by group size
for (i in N:1){
  keys <- unique(c(names(reglas_by_length[i]), names(reglas_def)))

  lista_ev <- setNames(mapply(c, reglas_by_length[i][keys], reglas_def[keys]), keys)
  lista_ev <- lapply(lista_ev, function(x){ x[!duplicated(x)] })
  lista_ev <- lista_ev[[sizes[i]]]

  # Combine rules with same variable structure
  NMs <- unique(lapply(lista_ev, names))
  res <- list()
  for (k in seq_along(NMs)){
    tmp <- NULL
    tmplist <- lapply(lista_ev,  function(x){
      if(all(names(x) == NMs[[k]])){
        append(tmp, x)
      }
    })
    res[[k]] <- do.call(rbind, tmplist)
  }

  # Transform each subgroup into truth tables and minimize
  tablas <- lapply(res, function(x) cbind(x, "SG"= 1))
  tablastt <- lapply(tablas, function(x) truthTable(x, outcome = "SG", inc.cut = 1))
  tablas_min <- lapply(tablastt, function(x) minimize(x))

```

```
  # Extract unique solutions
  sol_unicas <- lapply(tablas_min, function(x) unique(unlist(x$solution)))
  sol_unicas <- unlist(sol_unicas)
  sol_unicas_sep <- lapply(sol_unicas, function(x) unlist(strsplit(x, "\\*")))
  sol_unicas_by_length <- group_vectors_by_length(sol_unicas_sep)

  # Convert simplified solutions into binary form
  sol_unicas_bin <- lapply(sol_unicas_by_length, function(ls){
    lapply(ls,  function(x){
      res <- as.numeric(!grepl("~", x))
      names(res) <- gsub(pattern = "~", "", x)
      return(res)
    })
  } )

  # Merge into final simplified rule list
  keys <- unique(c(names(sol_unicas_bin), names(reglas_def)))
  merged_list <- setNames(mapply(c, sol_unicas_bin[keys], reglas_def[keys]), keys)
  reglas_def <- lapply(merged_list, function(x){ x[!duplicated(x)] })
}

# View final list of simplified (non-redundant) rules
aaa <- DATAFRAME(reglas_def)
view(aaa)
```

## A.3 GENERACIÓN Y DEPURACIÓN DE REGLAS DE ASOCIACIÓN BINARIAS NO REDUNDANTES PARA EL ANÁLISIS CAUSAL INTERPRETATIVO

Este apéndice presenta el código desarrollado en lenguaje R, utilizado en el proceso de minería de reglas de asociación binarias, con énfasis especial en la eliminación de reglas redundantes, con el objetivo de obtener un conjunto mínimo y no redundante de implicaciones entre las condiciones causales y el resultado. Este enfoque busca aumentar la interpretabilidad lógica y reducir la complejidad del modelo, en consonancia con los principios de parsimonia también valorados en el Análisis Comparativo Cualitativo (QCA).

Se ha utilizado el algoritmo Apriori para generar un conjunto inicial de reglas de la forma:

$$A_1 \wedge A_2 \wedge \cdots \wedge A_k \Rightarrow B$$

donde $A_i \in \{0,1\}$ representa las variables causales (condiciones) y $B \in \{0,1\}$ es la variable de resultado.

Cada regla es evaluada mediante dos métricas clásicas:

- Soporte (s):
$$s = \frac{\text{soporte}(\text{LHS} \wedge \text{RHS})}{N}$$
- Confianza (c):
$$c = \frac{\text{soporte}(\text{LHS} \wedge \text{RHS})}{\text{soporte}(\text{LHS})}$$

Después de la generación inicial, el código implementa un proceso de minimización lógica de las reglas a través de la eliminación de reglas *redundantes*. Una regla $R_j$ se considera redundante si existe otra regla $R_i$ tal que:

- $LHS(R_i) \subseteq LHS(R_j)$,
- $RHS(R_i) = RHS(R_j)$,
- y $conf(R_i) \geq conf(R_j)$.

En otras palabras, la regla $R_j$ se elimina si otra regla más simple y al menos igual de confiable la vuelve innecesaria. Esta lógica se basa en el concepto de cobertura subsuntiva y se puede expresar formalmente como:

$$\text{Redundancia:} \quad R_j \text{ es redundante si } \exists R_i : \quad LHS(R_i) \subset LHS(R_j) \wedge RHS(R_i) = RHS(R_j)$$

El resultado de este proceso es un subconjunto de reglas no redundantes, denotado como:

$$\mathcal{R}^* = \{R \in \mathcal{R} \mid \nexists R' \in \mathcal{R} : R' \text{ hace redundante a } R\}$$

Este conjunto final $\mathcal{R}^*$ representa la forma minimizada de las reglas de asociación, facilitando la interpretación causal y su comparación directa con soluciones obtenidas mediante métodos como el csQCA.

Código 3: Generación y depuración de reglas de asociación binarias no redundantes para el análisis causal interpretativo.

```r
# Loading libraries -------------------------------------------------------

library(QCA)       # For QCA analysis
library(arules)    # For association rule mining
library(ggplot2)
library(dplyr)
library(tidyr)
library(Rcpp)
library(stringr)
library(purrr)

# Function to generate random binary data matrix
gendat <- function(ncols, nrows, seed = 12345) {
  set.seed(seed)
  dat <- unique(matrix(sample(0:1, ncols * nrows, replace = TRUE), ncol = ncols))
  colnames(dat) <- LETTERS[seq(ncols)]
  return(as.data.frame(dat[order(dat[, ncols], decreasing = TRUE), ]))
}

# C++ Functions -----------------------------------------------------------

cppFunction("
#include <Rcpp.h>
#include <unordered_map>
#include <algorithm>
using namespace Rcpp;
using namespace std;

// Extract LHS vars from {var1,var2,...} => {something}
vector<string> extract_lhs_vars(const string& line) {
  size_t start = line.find('{');
  size_t end = line.find('}');
  if (start == string::npos || end == string::npos || end <= start)
```

```
    return {};

  string inner = line.substr(start + 1, end - start - 1);
  vector<string> vars;
  string token;
  for (size_t i = 0; i <= inner.size(); ++i) {
    if (i == inner.size() || inner[i] == ',') {
      if (!token.empty()) {
        vars.push_back(token);
        token.clear();
      }
    } else {
      token += inner[i];
    }
  }
  return vars;
}

// [[Rcpp::export]]
CharacterVector filter_bitwise_isolated_from_rules(CharacterVector lines) {
  unordered_map<int, vector<string>> by_length;

  for (int i = 0; i < lines.size(); ++i) {
    string line = as<string>(lines[i]);
    vector<string> vars = extract_lhs_vars(line);
    int len = vars.size();
    by_length[len].push_back(line);
  }

  vector<string> result;

  for (auto& kv : by_length) {
    const vector<string>& group = kv.second;
    unordered_map<string, vector<pair<string, vector<int>>>> var_groups;

    for (const string& line : group) {
      vector<string> vars = extract_lhs_vars(line);
      vector<string> vars_upper;
      vector<int> bitvec;

      for (const string& v : vars) {
        string vu = v;
        transform(vu.begin(), vu.end(), vu.begin(), ::toupper);
        vars_upper.push_back(vu);
        bitvec.push_back(isupper(v[0]) ? 1 : 0);
      }

      sort(vars_upper.begin(), vars_upper.end());
      string key = vars_upper[0];
      for (size_t k = 1; k < vars_upper.size(); ++k)
        key += '*' + vars_upper[k];

      var_groups[key].push_back({line, bitvec});
    }

    for (auto& g : var_groups) {
      const auto& entries = g.second;
      int n = entries.size();
```

```
92        vector<bool> keep(n, true);
93
94        for (int i = 0; i < n - 1; ++i) {
95          for (int j = i + 1; j < n; ++j) {
96            int diff = 0;
97            for (size_t k = 0; k < entries[i].second.size(); ++k) {
98              if (entries[i].second[k] != entries[j].second[k]) ++diff;
99              if (diff > 1) break;
100           }
101           if (diff == 1) {
102             keep[i] = false;
103             keep[j] = false;
104           }
105         }
106       }
107
108       for (int i = 0; i < n; ++i)
109         if (keep[i]) result.push_back(entries[i].first);
110     }
111   }
112
113   return wrap(result);
114 }
115 ")
116
117 cppFunction("
118 #include <Rcpp.h>
119 #include <unordered_set>
120 #include <algorithm>
121 #include <string>
122
123 using namespace Rcpp;
124 using std::string;
125 using std::vector;
126 using std::unordered_set;
127
128 string tolower_str(const string& s) {
129   string out = s;
130   std::transform(out.begin(), out.end(), out.begin(), ::tolower);
131   return out;
132 }
133
134 bool is_upper(const string& s) {
135   return s != tolower_str(s);
136 }
137
138 bool is_lower(const string& s) {
139   return s == tolower_str(s);
140 }
141
142 // [[Rcpp::export]]
143 CharacterVector find_redundant_rules(List rule_vars) {
144   int n = rule_vars.size();
145   unordered_set<string> redundant_set;
146
147   for (int i = 0; i < n - 1; ++i) {
148     CharacterVector lhs1 = rule_vars[i];
149     unordered_set<string> upper1, lower1;
```

```
150
151     for (int k = 0; k < lhs1.size(); ++k) {
152       string var = as<string>(lhs1[k]);
153       (is_upper(var) ? upper1 : lower1).insert(var);
154     }
155
156     for (int j = i + 1; j < n; ++j) {
157       CharacterVector lhs2 = rule_vars[j];
158       unordered_set<string> upper2, lower2;
159
160       for (int k = 0; k < lhs2.size(); ++k) {
161         string var = as<string>(lhs2[k]);
162         (is_upper(var) ? upper2 : lower2).insert(var);
163       }
164
165       for (const string& up : upper1) {
166         string up_lower = tolower_str(up);
167         if (lower2.count(up_lower)) {
168           vector<string> new_vars;
169
170           for (int k = 0; k < lhs1.size(); ++k) {
171             string var = as<string>(lhs1[k]);
172             if (!is_upper(var)) new_vars.push_back(var);
173           }
174           for (int k = 0; k < lhs2.size(); ++k) {
175             string var = as<string>(lhs2[k]);
176             if (!is_lower(var)) new_vars.push_back(var);
177           }
178
179           unordered_set<string> seen;
180           bool conflict = false;
181           for (const string& var : new_vars) {
182             string low = tolower_str(var);
183             if (seen.count(low)) {
184               conflict = true;
185               break;
186             }
187             seen.insert(low);
188           }
189
190           if (!conflict) {
191             sort(new_vars.begin(), new_vars.end());
192             string implied;
193             for (size_t k = 0; k < new_vars.size(); ++k) {
194               implied += new_vars[k];
195               if (k != new_vars.size() - 1) implied += '*';
196             }
197             redundant_set.insert(implied);
198           }
199
200           break;
201         }
202       }
203
204       for (const string& up : upper2) {
205         string up_lower = tolower_str(up);
206         if (lower1.count(up_lower)) {
207           vector<string> new_vars;
```

```
208
209           for (int k = 0; k < lhs2.size(); ++k) {
210             string var = as<string>(lhs2[k]);
211             if (!is_upper(var)) new_vars.push_back(var);
212           }
213           for (int k = 0; k < lhs1.size(); ++k) {
214             string var = as<string>(lhs1[k]);
215             if (!is_lower(var)) new_vars.push_back(var);
216           }
217
218           unordered_set<string> seen;
219           bool conflict = false;
220           for (const string& var : new_vars) {
221             string low = tolower_str(var);
222             if (seen.count(low)) {
223               conflict = true;
224               break;
225             }
226             seen.insert(low);
227           }
228
229           if (!conflict) {
230             sort(new_vars.begin(), new_vars.end());
231             string implied;
232             for (size_t k = 0; k < new_vars.size(); ++k) {
233               implied += new_vars[k];
234               if (k != new_vars.size() - 1) implied += '*';
235             }
236             redundant_set.insert(implied);
237           }
238
239           break;
240         }
241       }
242     }
243   }
244
245   CharacterVector out;
246   for (const auto& s : redundant_set) out.push_back(s);
247   return out;
248 }
249 ")
250
251 # R functions to generate and process data -------------------------------
252
253 generate_random_causal_configs <- function(n_configs, n_conditions, seed = NULL) {
254   if (!is.null(seed)) set.seed(seed)
255   configs <- vector("list", n_configs)
256   for (i in seq_len(n_configs)) {
257     cond_indices <- sort(sample(seq_len(n_conditions), sample(1:n_conditions, 1)))
258     values <- sample(0:1, length(cond_indices), replace = TRUE)
259     cfg <- setNames(as.list(values), paste0("C", cond_indices))
260     configs[[i]] <- cfg
261   }
262   return(configs)
263 }
264
```

```r
generate_data_with_noise <- function(n_cases, n_conditions, causal_configs,
    consistency = 1.0, seed = NULL) {
  if (!is.null(seed)) set.seed(seed)
  data <- as.data.frame(matrix(sample(0:1, n_cases * n_conditions, replace = TRUE),
                               nrow = n_cases, ncol = n_conditions))
  names(data) <- paste0("C", 1:n_conditions)
  data$Y <- 0
  for (cfg in causal_configs) {
    vars <- names(cfg)
    match_rows <- apply(data[, vars, drop = FALSE], 1, function(row) all(row ==
        unlist(cfg)))
    flip <- runif(sum(match_rows)) < consistency
    data$Y[which(match_rows)[flip]] <- 1
  }
  return(data)
}

# Analysis functions ------------------------------------------------------

# csQCA function
processar_csQCA <- function(data, outcome, conditions) {
  tempo_csQCA <- system.time({
    truth_table <- truthTable(data, outcome = outcome, conditions = conditions)
    solucao_csQCA <- minimize(truth_table, max.comb = 0)
  })
  return(list(time = tempo_csQCA["elapsed"], sol = solucao_csQCA))
}

# Apriori rule mining function
processar_AR <- function(dados, suporte_min = 0.01, confianca_min = 0.98) {
  dados_validos <- dados
  # Convert columns to factors
  dados_validos[] <- lapply(dados_validos, factor)

  # Generate rules using Apriori
  regras <- apriori(dados_validos,
                    parameter = list(support = suporte_min,
                                     confidence = confianca_min,
                                     target = "rules"))

  # Convert to readable format
  regras_df <- as(regras, "data.frame")

  return(list(rules = regras, rules_df = regras_df))
}

## Function to extract LHS variables from rules ----
extract_lhs_from_rules <- function(rules) {
  lhs_vars <- labels(lhs(rules))
  lhs_vars <- gsub("\\{", "", lhs_vars)
  lhs_vars <- gsub("\\}", "", lhs_vars)
  lhs_list <- strsplit(lhs_vars, ",")
  lhs_list <- lapply(lhs_list, function(x) gsub("^\\s+|\\s+$", "", x))
  return(lhs_list)
}
```

## A.4 EJEMPLO 1 – CAPÍTULO 4: LA CAÍDA DE LA DEMOCRACIA EN EUROPA DURANTE LAS GRANDES GUERRAS

El código presentado a continuación corresponde al Ejemplo 1 del Capítulo 4 de esta tesis, cuyo objetivo es analizar los factores asociados con la caída de la democracia en países europeos durante la Primera y Segunda Guerra Mundial. Utilizando datos cuantitativos representativos de condiciones estructurales (como desarrollo, urbanización, industrialización, inestabilidad y alfabetización), se aplicaron técnicas de análisis comparativo cualitativo de conjuntos difusos (fsQCA) y minería de reglas de asociación binarizadas para identificar patrones causales.

Este ejemplo demuestra, de manera empírica, que las soluciones obtenidas mediante la técnica fsQCA constituyen un subconjunto de las soluciones derivadas de la minería de reglas de asociación cuantitativas, cuando estas se binarizan adecuadamente en dos intervalos. Esto refuerza la consistencia de las soluciones fsQCA frente a métodos alternativos para el descubrimiento de patrones causales, destacando su validez lógica y empírica.

Código 4: Análisis del Ejemplo 1 del Capítulo 4: comparación entre las soluciones obtenidas mediante fsQCA y las reglas de asociación binarizadas para explicar la caída de la democracia en Europa durante las guerras mundiales.

```
# Load required libraries
library(openxlsx)
library(dplyr)
library(writexl)
library(QCA)
library(haven)
library(arules)
library(readxl)
library(tidyr)

# Read the dataset
data <- data_caida.Demo

# Select variables of interest
vars <- c("Survived", "Developed", "Urban", "Literate", "Industrial", "Unstable")

# Compute quantiles for each variable
result <- sapply(data[vars], function(x) {
  quantile(x, probs = c(0, 0.33, 0.50, 0.66, 1), na.rm = TRUE)
})

# Transpose for better readability
result <- t(result)

# Print quantile results
print(result)

# Display first rows
head(data)

# Check min and max of variable
minmax(Dev)

# Extract variables
Cont <- data$Country
Sur <- data$Survived
Dev <- data$Developed
```

```r
Ur <- data$Urban
Lit <- data$Literate
Ind <- data$Industrial
Uns <- data$Unstable

# Create data frame for QCA
dadosQCA <- data.frame(Dev, Ur, Lit, Ind, Uns, Sur)

# Calibrate fuzzy sets
FsDev <- calibrate(dadosQCA$Dev, type = "fuzzy", thresholds = c(398.5, 552, 871.5))
FsUr <- calibrate(dadosQCA$Ur, type = "fuzzy", thresholds = c(25.87, 33.7, 53.125))
FsLit <- calibrate(dadosQCA$Lit, type = "fuzzy", thresholds = c(73.3, 89.7, 98))
FsSur <- calibrate(dadosQCA$Sur, type = "fuzzy", thresholds = c(-8.0, 1.5, 9.5))

# Create fuzzy dataset
dadosfsqca <- data.frame(FsDev, FsUr, FsLit, FsSur)

# Build truth table
tvSE <- truthTable(dadosfsqca, outcome = "~FsSur", incl.cut = 0.75, show.cases = TRUE
    , dcc = TRUE, sort.by = "OUT, n")
tvSE

# Compute complex solution
compxSE <- minimize(tvSE, datails = TRUE, show.cases = TRUE)
compxSE

# Compute parsimonious solution
ParcSE <- minimize(tvSE, include = "?")
ParcSE

# Compute intermediate solution
IntermedSE <- minimize(tvSE, include = "?", details = TRUE, show.cases = TRUE)
IntermedSE

# Show simplified expressions
IntermedSE$SA

# Binary association rules

# Binarize variables based on thresholds
dados_bin <- as.data.frame(ifelse(dadosQCA > c(552, 33.7, 89.7, 1.5), 1, 0))

# Binarize each variable individually
Dev_bin <- ifelse(dadosQCA$Dev > 552, 1, 0)
Ur_bin  <- ifelse(dadosQCA$Ur  > 33.7, 1, 0)
Lit_bin <- ifelse(dadosQCA$Lit > 89.7, 1, 0)

# Invert binarization logic for Sur
Sur_bin <- ifelse(dadosQCA$Sur <= 1.5, 1, 0)

# Create new binary data frame
dados_bin <- data.frame(Dev = Dev_bin, Ur = Ur_bin, Lit = Lit_bin, Sur = Sur_bin)

# Add ID column
ID <- 1:18
ID

# Combine ID and binary data
```

```r
Nueva_AIs <- cbind(ID = ID, dados_bin)
Nueva_AIs

# Convert data to long format
datos_en_long_mig <- pivot_longer(Nueva_AIs, cols = -1, names_to = "Descripcion",
    values_to = "valor")

# Make labels lowercase for 0s
datos_en_long_mig$Descripcion[datos_en_long_mig$valor == 0] <- tolower(datos_en_long_
    mig$Descripcion[datos_en_long_mig$valor == 0])

# Keep only ID and description
datos_en_long_mig <- datos_en_long_mig %>% select(c(ID, Descripcion))

# Split data into list by ID
data_lists = split(datos_en_long_mig$Descripcion, datos_en_long_mig$ID)

# Convert to transactions object
datos_trxx_mig = as(data_lists, "transactions")

# Show summary of transactions
summary(datos_trxx_mig)

# Find frequent itemsets
support_RR = apriori(datos_trxx_mig,
                     parameter = list(target = "frequent itemsets", supp = 0.1),
                     appearance = list(items = c("Sur"))
)

# Summary of frequent itemsets
summary(support_RR)

# Extract rules with Sur on RHS
inspect(apriori(datos_trxx_mig,
                appearance = list(rhs = "Sur"),
                parameter = list(support = 0.0001, confidence = 0.75)
))

# Generate rules
reglas <- apriori(datos_trxx_mig,
                  appearance = list(rhs = "Sur"),
                  parameter = list(support = 0.0001, confidence = 0.75)
)

# Print rules
print(reglas)
reglas

# Convert rules to data frame
reglasrulesdf_mig <- DATAFRAME(reglas)

# View rules (case sensitive fix: change view() to View() if needed)
view(rulesdf_mig)
```

## A.5 DISCRETIZACIÓN

A continuación se presenta el script correspondiente a la generación de reglas de asociación a partir de variables cuantitativas discretizadas en tres términos lingüísticos *bajo*, *medio* y *alto* aplicadas a los datos sobre la oposición a la inmigración en Europa durante la Primera y la Segunda Guerra Mundial.

Código 5: Script para la generación de reglas de asociación a partir de variables cuantitativas discretizadas en términos lingüísticos *bajo, medio, alto*.

```r
library(openxlsx)
library(arules)
library(dplyr)
library(writexl)

# Load the dataset (replace with your actual data)
# Example: data <- data_caida.Demo
data <- data_caida.Demo

# Select relevant columns
dadosQCA <- data %>%
  select(Country, Survived, Developed, Urban, Literate, Industrial, Unstable)

# Rename columns for shorter names
colnames(dadosQCA) <- c("Country", "Sur", "Dev", "Ur", "Lit", "Ind", "Uns")

# General Discretization function with 3 zones: Low, Mid, High
calibrar_fuzzy_geral <- function(df, columns) {
  for (col in columns) {
    values <- df[[col]]

    q1   <- quantile(values, 0.25, na.rm = TRUE)
    q1_5 <- quantile(values, 0.375, na.rm = TRUE)
    q2_5 <- quantile(values, 0.625, na.rm = TRUE)
    q3   <- quantile(values, 0.75, na.rm = TRUE)

    low_col    <- paste0("Low_", col)
    mid_col    <- paste0("Mid_", col)
    high_col   <- paste0("High_", col)

    df[[low_col]] <- df[[mid_col]] <- df[[high_col]] <- 0

    for (i in seq_along(values)) {
      v <- values[i]
      if (v <= q1) {
        df[[low_col]][i] <- 1
      } else if (v <= q1_5) {
        df[[low_col]][i] <- (q1_5 - v) / (q1_5 - q1)
        df[[mid_col]][i] <- 1 - df[[low_col]][i]
      } else if (v <= q2_5) {
        df[[mid_col]][i] <- 1
      } else if (v <= q3) {
        df[[high_col]][i] <- (v - q2_5) / (q3 - q2_5)
        df[[mid_col]][i] <- 1 - df[[high_col]][i]
      } else {
        df[[high_col]][i] <- 1
      }
    }
```

```
  }
  return(df)
}

# Apply the Discretization to continuous variables
continuous_vars <- c("Dev", "Ur", "Lit", "Ind", "Uns")
dados_fuzzy <- calibrar_fuzzy_geral(dadosQCA, continuous_vars)

# Convert membership degrees to binary (threshold > 0.5)
dados_binary <- dados_fuzzy %>%
  mutate(across(starts_with("Low_"):starts_with("High_"), ~ ifelse(. > 0.5, 1, 0)))

# Add the response variable (binary version of Survived)
dados_binary$Sur <- ifelse(dadosQCA$Sur > 0.5, 1, 0)

# Convert the dataset to transaction format for arules
trans_data <- as(dados_binary %>% select(-Country), "transactions")

# Generate association rules with Sur=1 as the target (consequent)
rules <- apriori(
  trans_data,
  parameter = list(supp = 0.1, conf = 0.75, target = "rules"),
  appearance = list(rhs = c("Sur=1"), default = "lhs")
)

# Show the rules sorted by confidence
inspect(sort(rules, by = "confidence"))
```

## A.6 EXTRACCIÓN DE REGLAS DIFUSAS SOBRE LA PARTICIPACIÓN ELECTORAL

El siguiente script corresponde al Ejemplo 1 del Capítulo 5, en el que se lleva a cabo la extracción de reglas difusas a partir de un conjunto de datos que analiza cómo diferentes factores sociales e institucionales afectan la participación electoral. En particular, se estudia el impacto de la percepción de la corrupción (NOCORRUPTION), el nivel educativo (EDUCATION), la desigualdad social (INEQUALITY) y la confianza en el parlamento (NOTRUST) sobre la intención de voto (VOTE).

A través de una lógica difusa, los datos se calibran utilizando contextos lingüísticos específicos para cada variable, permitiendo transformar valores cuantitativos en categorías como "bajo", "medio" y "alto". Luego, se aplican técnicas de minería de reglas de asociación difusas para identificar patrones y relaciones causales relevantes entre las condiciones consideradas y el resultado electoral. Estas reglas ofrecen una interpretación más flexible y realista de los datos sociales, capturando la ambigüedad e incertidumbre propias de fenómenos como el comportamiento político.

Código 6: Script para la generación de reglas de asociación difusas sobre la participación electoral.

```
library(openxlsx)     # For reading and writing Excel files
library(lfl)          # For fuzzy logic and linguistic variables
library(dplyr)        # For data manipulation
library(writexl)      # For exporting data to Excel

# Load the dataset
data1 <- data_CORRUPTION

# View the dataset
```

```
print(data1)

# Get the minimum of all numeric variables
sapply(data1, min, na.rm = TRUE)

# Get the maximum of all numeric variables
sapply(data1, max, na.rm = TRUE)

# Create a subset dataframe (if needed for inspection)
EuRag <- data.frame(VOTE, NOCORRUPTION, EDUCATION, INEQUALITY, NOTRUST)

# Fuzzify the data using linguistic context (calibration)
employees1 <- lcut(data1,
                   context = list(
                     VOTE = ctx3(low = 49.2, high = 86.495),
                     NOCORRUPTION = ctx3(low = 46.0, high = 88.0),
                     EDUCATION = ctx3(low = 45.367, high = 95.864),
                     INEQUALITY = ctx3(low = 23.000, high = 45.800),
                     NOTRUST = ctx3(low = 3.500, high = 38.500)
                   ),
                   atomic = list(VOTE = c("sm", "me", "bi")),
                   hedges = list(
                     VOTE = c("-"),
                     NOCORRUPTION = c("-"),
                     EDUCATION = c("-"),
                     INEQUALITY = c("-"),
                     NOTRUST = c("-")
                   ))

# Round the calibrated values to 3 decimal places
clibrod <- round(employees1, 3)


# Get column-wise sums of fuzzy memberships
colSums(employees1)

# Search for association rules with VOTE as consequent
rb <- searchrules(employees1,
                  lhs = grep("VOTE", colnames(employees1), invert = TRUE),  #
                      Antecedents
                  rhs = grep("sm.VOTE", colnames(employees1)),              #
                      Consequent
                  minSupport = 0.0001,     # Minimum support threshold
                  minConfidence = 0.75,    # Minimum confidence threshold
                  maxLength = 5)           # Maximum length of the rule
```

## A.7 IMPLEMENTACIÓN DEL CAPÍTULO 5: MÉTODO UWTOPSIS APLICADO AL RANKING DE REGLAS DE ASOCIACIÓN

Este script implementa la metodología propuesta en el Capítulo 5 de esta tesis doctoral, en la que se desarrolla y evalúa el método uwTOPSIS (TOPSIS con pesos no uniformes) para el ranking de reglas de asociación. El proceso comienza con la extracción de las reglas de asociación, que representan patrones relevantes encontrados en los datos. A continuación, estas reglas son minimizadas con el objetivo de reducir la complejidad del modelo y evitar la explosión combinatoria típica de grandes conjuntos de reglas, manteniendo únicamente aquellas que son lógicamente no redundantes e informativamente relevantes.

Tras esta etapa de preprocesamiento, se aplica el método uwTOPSIS para ordenar las reglas en función de múltiples criterios, considerando tanto la maximización como la minimización de determinados indicadores, e incorporando pesos personalizados. Seguidamente, se realiza un análisis de sensibilidad mediante simulaciones de Monte Carlo, variando los pesos asignados a los criterios de decisión. Este análisis tiene como objetivo evaluar la robustez del ranking frente a incertidumbres o perturbaciones en los pesos de los criterios.

Código 7: Método uwTOPSIS aplicado al ranking de reglas de asociación.

```
library(readxl)
library(ggplot2)
library(haven)
library(readr)
library(arules)
library(QCA)
library(dplyr)
library(tidyr)
library(xtable)

# Load custom TOPSIS script
source("Base de dados/TOPSIS.R")

# Load dataset from Excel file
MICRODADOS1 <- read_excel("C:/Users/Principal/OneDrive - Universitat de Valencia/
    Escritorio/Base de dados/MICRODADOS1.xlsx")

# Assign dataset to a working variable
pima_data1 <- MICRODADOS1

# View first few rows of the dataset
head(pima_data1)

# Select relevant columns for analysis
nova_base <- pima_data1 %>%
  select(
    Classificacao, FaixaEtaria, Sexo, Febre, DificuldadeRespiratoria, Tosse,
        DorGarganta, Cefaleia,
    ComorbidadePulmao, ComorbidadeCardio, ComorbidadeRenal, ComorbidadeDiabetes,
    ComorbidadeTabagismo, ComorbidadeObesidade, FicouInternado, ViagemBrasil,
        ViagemInternacional
  )

# Remove rows where any column contains "-" (missing or invalid data)
nova_base_limpa <- nova_base %>%
  filter(!apply(., 1, function(row) any(row == "-")))

# Check if any "-" values remain
sapply(nova_base_limpa, function(col) sum(col == "-"))

# Remove rows with any NA values
nova_base_limpa <- na.omit(nova_base_limpa)

# Print number of remaining records
print(nrow(nova_base_limpa))

# Filter dataset to include only rows where Classificacao is "Confirmados" or "
    Suspeito"
nova_base_limpa <- nova_base_limpa %>%
```

```
  filter(Classificacao %in% c("Confirmados", "Suspeito"))

# Plot bar chart of Classificacao counts with labels
ggplot(nova_base_limpa, aes(x = Classificacao)) +
  geom_bar(fill = "skyblue", color = "black") +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  labs(title = "Frequency of COVID Classification", x = "Classification", y = "
      Frequency") +
  theme_minimal()

# Convert categorical variables into binary numeric variables (0/1)
nova_base_limpa <- nova_base_limpa %>%
  mutate(
    Classificacao_Binaria = ifelse(Classificacao == "Confirmados", 1, 0),
    EDAD_ADULT = ifelse(FaixaEtaria %in% c("0 a 4 anos", "05 a 9 anos", "10 a 19 anos
        ", "20 a 29 anos"), 0, 1),
    HOMBRE = ifelse(Sexo %in% c("F", "I"), 0, 1),
    FEBRE = ifelse(Febre == "Nao", 0, 1),
    DIF_RESP = ifelse(DificuldadeRespiratoria == "Nao", 0, 1),
    TOSSE = ifelse(Tosse == "Nao", 0, 1),
    DORGARGANTA = ifelse(DorGarganta == "Nao", 0, 1),
    CEFALIA = ifelse(Cefaleia == "Nao", 0, 1),
    COM_PUL = ifelse(ComorbidadePulmao == "Nao", 0, 1),
    COM_CAR = ifelse(ComorbidadeCardio == "Nao", 0, 1),
    COM_REN = ifelse(ComorbidadeRenal == "Nao", 0, 1),
    COM_DB = ifelse(ComorbidadeDiabetes == "Nao", 0, 1),
    COM_TB = ifelse(ComorbidadeTabagismo == "Nao", 0, 1),
    COM_OB = ifelse(ComorbidadeObesidade == "Nao", 0, 1),
    VIAG = ifelse(ViagemBrasil %in% c("Nao", "Ignorado"), 0, 1)
  )

# Select only the binary variables for further analysis
novo_data_frame <- nova_base_limpa %>%
  select(
    Classificacao_Binaria,
    EDAD_ADULT,
    HOMBRE,
    FEBRE,
    DIF_RESP,
    TOSSE,
    DORGARGANTA,
    CEFALIA,
    COM_PUL,
    COM_CAR,
    COM_REN,
    COM_DB,
    COM_TB,
    COM_OB,
    VIAG
  )

head(novo_data_frame)

# Create a combined comorbidity variable: 1 if any comorbidity present, else 0
novo_data_frame <- novo_data_frame %>%
  mutate(COM = ifelse(rowSums(select(., COM_PUL:COM_OB)) > 0, 1, 0))

# Rename variables for clarity
```

```r
novo_data_frame <- novo_data_frame %>%
  rename(
    COVID = Classificacao_Binaria,
    DG = DORGARGANTA
  )

# Select final variables for model/data analysis
pima_data_bin <- novo_data_frame %>%
  select(
    EDAD_ADULT,
    FEBRE,
    DIF_RESP,
    TOSSE,
    DG,
    CEFALIA,
    COM,
    COVID
  )

# Check unique values of COVID binary variable
unique(pima_data_bin$COVID)

# Plot bar chart for COVID binary variable counts
ggplot(pima_data_bin, aes(x = factor(COVID))) +
  geom_bar(fill = "skyblue", color = "black") +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  labs(title = "Frequency of COVID (Binary)", x = "COVID", y = "Frequency") +
  theme_minimal()


# Association Rule Mining

# Add an ID column
nrow(pima_data_bin)

ID <- seq_len(nrow(pima_data_bin))

# Combine ID with dataset
Nueva_pima_data_bin <- cbind(ID = ID, pima_data_bin)
Nueva_pima_data_bin

# Transform the data into long format
datos_en_long_pima <- pivot_longer(
  Nueva_pima_data_bin,
  cols = -1,
  names_to = "Descripcion",
  values_to = "valor"
)

# Encode positive variables (value == 1) in uppercase and negative ones (value == 0)
    in lowercase
datos_en_long_pima$Descripcion[datos_en_long_pima$valor == 0] <- tolower(datos_en_
    long_pima$Descripcion[datos_en_long_pima$valor == 0])
datos_en_long_pima$Descripcion[datos_en_long_pima$valor == 1] <- toupper(datos_en_
    long_pima$Descripcion[datos_en_long_pima$valor == 1])

# Keep only the ID and Descripcion columns
datos_en_long_pima <- datos_en_long_pima %>% select(c(ID, Descripcion))
```

```r

# Split descriptions by ID into a list
data_lists <- split(datos_en_long_pima$Descripcion, datos_en_long_pima$ID)

# Convert the list to a transaction dataset
datos_trxx_pima <- as(data_lists, "transactions")

# Show summary and head of transactions
summary(datos_trxx_pima)
head(datos_trxx_pima)

# Discover frequent itemsets with minimum support of 0.1 that include "COVID"
support_RR <- apriori(
  datos_trxx_pima,
  parameter = list(target = "frequent itemsets", supp = 0.1),
  appearance = list(items = c("COVID"))
)

# Summary of frequent itemsets
summary(support_RR)

# Generate association rules with "COVID" as RHS (consequent)
inspect(
  apriori(
    datos_trxx_pima,
    appearance = list(rhs = "COVID"),
    parameter = list(support = 0.0001, confidence = 0.85)
  )
)

# Save the rules to an object
reglas <- apriori(
  datos_trxx_pima,
  appearance = list(rhs = "COVID"),
  parameter = list(support = 0.0001, confidence = 0.85)
)

# Print the rules
print(reglas)

# Convert rules to a data frame for analysis or export
rulesdf_pima <- DATAFRAME(reglas)

# View and format rules in a LaTeX-friendly table
View(rulesdf_pima)
xtable(rulesdf_pima)


# Quine-McCluskey Minimization Method

# Convert rules to data frame
rulesdf <- DATAFRAME(reglas)

# Initialize list of binary rules
reglas <- list()

# Loop through each rule
for(i in 1:nrow(rulesdf_pima)) {
```

```
  cat("Rule ", i, "/", nrow(rulesdf_pima), "\n")

  lhs_end <- rulesdf_pima$LHS[i]
  regla1 <- as.character(lhs_end)

  # Split the rule by commas and remove braces
  regla_temp1 <- unlist(strsplit(regla1, ","))
  regla_temp2 <- gsub("\\{", "", regla_temp1)
  regla_temp3 <- gsub("\\}", "", regla_temp2)

  # Convert variables to binary: 1 if uppercase (positive), 0 if lowercase (negative)
  regla_bin <- as.numeric(sapply(regla_temp3, function(x) identical(x, toupper(x))))

  # Assign variable names (all in uppercase)
  names(regla_bin) <- toupper(regla_temp3)

  # Store binary rule
  reglas[[i]] <- regla_bin
}

# Function to group rules by number of variables
group_vectors_by_length <- function(vector_list) {
  grouped_vectors <- split(vector_list, sapply(vector_list, length))
  return(grouped_vectors)
}

# Group rules by length
reglas_by_length <- group_vectors_by_length(reglas)

# Initialize minimized rules by size
sizes <- names(reglas_by_length)
N <- length(sizes)
reglas_def <- vector("list", length = N)
names(reglas_def) <- sizes

# Apply minimization from largest to smallest rule size
for (i in N:1) {
  cat(i, "\n")

  # Combine current and previous rules
  keys <- unique(c(names(reglas_by_length[i]), names(reglas_def)))
  lista_ev <- setNames(mapply(c, reglas_by_length[i][keys], reglas_def[keys]), keys)

  # Remove duplicates
  lista_ev <- lapply(lista_ev, function(x) {
    x[!duplicated(x)]
  })

  lista_ev <- lista_ev[[sizes[i]]]

  # Group by variable names
  NMs <- unique(lapply(lista_ev, names))
  res <- list()

  for (k in seq_along(NMs)) {
    tmp <- NULL
    tmplist <- lapply(lista_ev, function(x) {
      if (all(names(x) == NMs[[k]])) {
```

```
273         append(tmp, x)
274       }
275     })
276     res[[k]] <- do.call(rbind, tmplist)
277   }
278
279   # Create truth tables for each group
280   tablas <- lapply(res, function(x) cbind(x, "OUT" = 1))
281   tablastt <- lapply(tablas, function(x) truthTable(x, outcome = "OUT", inc.cut = 1))
282
283   # Minimize truth tables
284   tablas_min <- lapply(tablastt, function(x) minimize(x))
285
286   # Extract unique solutions
287   sol_unicas <- lapply(tablas_min, function(x) unique(unlist(x$solution)))
288   sol_unicas <- unlist(sol_unicas)
289   sol_unicas_sep <- lapply(sol_unicas, function(x) unlist(strsplit(x, "\\*")))
290
291   # Group solutions by length
292   sol_unicas_by_length <- group_vectors_by_length(sol_unicas_sep)
293
294   # Convert minimized solutions to binary format
295   sol_unicas_bin <- lapply(sol_unicas_by_length, function(ls) {
296     lapply(ls, function(x) {
297       res <- as.numeric(!grepl("~", x))
298       names(res) <- gsub(pattern = "~", "", x)
299       return(res)
300     })
301   })
302
303   # Merge with previously minimized rules
304   keys <- unique(c(names(sol_unicas_bin), names(reglas_def)))
305   merged_list <- setNames(mapply(c, sol_unicas_bin[keys], reglas_def[keys]), keys)
306
307   reglas_def <- lapply(merged_list, function(x) {
308     x[!duplicated(x)]
309   })
310 }
311
312 # Final minimized rules
313 reglas_def
314
315
316 # Convert binary rule vectors into QCA-style solutions (e.g., {A,b,C})
317 as_qca_sol <- function(reg){
318   var_pos <- reg == 1             # Identify variables present in the rule
319   var_neg <- reg == 0             # Identify variables absent (negated) in the rule
320   vars <- names(reg)
321   vars[var_neg] <- tolower(vars[var_neg])  # Convert negated variables to lowercase
322   solution <- paste(vars, collapse = ",")  # Join variables with commas
323   solution <- paste0("{",solution,"}")     # Wrap in curly braces to form a solution
        string
324   return(solution)
325 }
326
327 # Apply transformation to all rules to obtain QCA-style representations
328 reglas_defintivas <- lapply(reglas_def, function(x) lapply(x, as_qca_sol))
329 reglas_defintivas_vector <- unlist(reglas_defintivas)  # Flatten the list to a vector
```

```r

# Filter the original rule dataframe to keep only those with QCA-formatted LHS
rulesdf_minim <- rulesdf_pima %>% filter(LHS %in% reglas_defintivas_vector)

# Convert LHS to character type and calculate rule length
rulesdf_minim$LHS <- as.character(rulesdf_minim$LHS)
rulesdf_minim$Length <- sapply(strsplit(rulesdf_minim$LHS, ","), length)

# Keep only relevant columns (exclude RHS and count)
rulesdf_TOPSIS <- rulesdf_minim %>% select(-c(RHS, count))

# Rename rules as R1 to R49 for easier identification
rulesdf_TOPSIS$LHS <- paste0("R", seq_len(49))

# Preview the transformed dataset
head(rulesdf_TOPSIS)


# ------------------------------
# Perform uwTOPSIS analysis
# ------------------------------
library(uwTOPSIS)

# First uwTOPSIS run using different preference thresholds (L and U)
uwt1 <- uwTOPSIS(
  rulesdf_TOPSIS,
  directions = c("max", "max", "max", "max", "min"),  # Benefit or cost criteria
  L = c(0.05, 0.05, 0.05, 0.05, 0.05),                # Lower preference thresholds
  U = c(0.35, 0.35, 0.35, 0.35, 0.35),                # Upper preference thresholds
  norm.method = "minmax",
  ordered = TRUE,
  makefigure = FALSE
)

# Second uwTOPSIS run with equal weights for all indicators
uwt2 <- uwTOPSIS(
  rulesdf_TOPSIS,
  directions = c("max", "max", "max", "max", "min"),
  L = c(0.2, 0.2, 0.2, 0.2, 0.2),
  U = c(0.2, 0.2, 0.2, 0.2, 0.2),
  norm.method = "minmax",
  ordered = TRUE,
  makefigure = FALSE
)

# Check the structure of uwTOPSIS result object
str(uwt1$scores)


# ---------------------------------------------
# Compare scores and rankings between methods
# ---------------------------------------------

# Extract scores for each method
df1 <- uwt1$scores[, c("LHS", "uwTOPSIS")]
df2 <- uwt2$scores[, c("LHS", "uwTOPSIS")]

# Rename columns to identify the method
```

```r
colnames(df1) <- c("Regra", "Score_uwTOPSIS_Method1")
colnames(df2) <- c("Regra", "Score_uwTOPSIS_Method2")

# Calculate ranks based on scores (higher is better)
df1$Rank_uwTOPSIS_Method1 <- rank(-df1$Score_uwTOPSIS_Method1)
df2$Rank_uwTOPSIS_Method2 <- rank(-df2$Score_uwTOPSIS_Method2)

# Merge the two dataframes on Regra
comparacao_LHS <- merge(
  df1[, c("Regra", "Score_uwTOPSIS_Method1", "Rank_uwTOPSIS_Method1")],
  df2[, c("Regra", "Score_uwTOPSIS_Method2", "Rank_uwTOPSIS_Method2")],
  by = "Regra", all = TRUE
)

# Compute the difference in ranks between the two methods
comparacao_LHS$Rank_Difference <- comparacao_LHS$Rank_uwTOPSIS_Method1 - comparacao_
    LHS$Rank_uwTOPSIS_Method2

# Print the comparison dataframe
print(comparacao_LHS)


# Comparison of scores and rankings between two uwTOPSIS configurations

# 1. Retrieve scores and compute rankings for both methods
scores1 <- uwt1$scores$uwTOPSIS  # Scores from method with L=0.05, U=0.35
scores2 <- uwt2$scores$uwTOPSIS  # Scores from method with L=U=0.2

# Compute rankings (higher score = better rank)
rank1 <- rank(-scores1)
rank2 <- rank(-scores2)

# Create a comparative dataframe with rule identifiers
comparacao <- data.frame(
  ID = uwt1$scores$LHS,  # Rule identifiers (LHS of association rules)
  Score_L1 = round(scores1, 4),
  Rank_L1 = rank1,
  Score_L2 = round(scores2, 4),
  Rank_L2 = rank2,
  Dif_Rank = rank1 - rank2,                # Difference in rankings
  Dif_Score = round(scores1 - scores2, 4)  # Difference in scores
)

# View first rows of the comparison table
head(comparacao)


# 2. Scatter plot comparing rankings between the two methods
library(ggplot2)

ggplot(comparacao, aes(x = Rank_L1, y = Rank_L2, label = ID)) +
  geom_point(color = "steelblue", size = 3) +
  geom_text(vjust = -0.5, size = 3) +  # Show rule IDs above points
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "gray40") +
  theme_minimal() +
  labs(
    title = "Ranking Comparison: uwTOPSIS (L1 vs L2)",
    x = "Ranking with L = 0.05, U = 0.35",
```

```
    y = "Ranking with L = U = 0.2"
  )


# 3. Histogram + density plot of scores for both methods
scores_long <- data.frame(
  Score = c(scores1, scores2),
  Method = rep(c("uwTOPSIS", "Standard TOPSIS"), each = length(scores1))
)

ggplot(scores_long, aes(x = Score, fill = Method)) +
  geom_histogram(aes(y = ..density..), bins = 20, color = "black", alpha = 0.6) +

  # Fit normal curve for uwTOPSIS
  stat_function(
    fun = dnorm,
    args = list(mean = mean(scores1), sd = sd(scores1)),
    color = "gray20", size = 0.8, linetype = "solid",
    data = subset(scores_long, Method == "uwTOPSIS")
  ) +

  # Fit normal curve for Standard TOPSIS
  stat_function(
    fun = dnorm,
    args = list(mean = mean(scores2), sd = sd(scores2)),
    color = "gray50", size = 0.8, linetype = "solid",
    data = subset(scores_long, Method == "Standard TOPSIS")
  ) +

  facet_wrap(~ Method, nrow = 1) +
  scale_fill_manual(values = c("lightblue", "darkorange")) +
  theme_minimal(base_size = 10) +
  labs(
    title = "Score Distribution by Method",
    x = "Scores",
    y = "Density"
  ) +
  theme(
    legend.position = "none",
    strip.text = element_text(size = 10),
    plot.title = element_text(size = 10, hjust = 0.5),
    axis.title = element_text(size = 9),
    axis.text = element_text(size = 9)
  )


# 4. Create a new comparison dataframe (optional, duplicate)
comparacao <- data.frame(
  ID = seq_along(scores1),  # Numeric identifiers
  Score_L1 = round(scores1, 4),
  Rank_L1 = rank1,
  Score_L2 = round(scores2, 4),
  Rank_L2 = rank2,
  Dif_Rank = rank1 - rank2,
  Dif_Score = round(scores1 - scores2, 4)
)

# 5. Calculate correlation between scores and ranks
```

```r
cor_pearson <- cor(scores1, scores2, method = "pearson")
cor_spearman <- cor(rank1, rank2, method = "spearman")

# 6. Scatter plot comparing scores directly
ggplot(comparacao, aes(x = Score_L1, y = Score_L2)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "red") +
  labs(
    title = "Score Comparison: uwTOPSIS L1 vs L2",
    x = "Scores with L = 0.05",
    y = "Scores with L = 0.2"
  ) +
  theme_minimal()

# 7. Optional: Heatmap of ranking differences
library(pheatmap)

pheatmap(as.matrix(comparacao$Dif_Rank),
         cluster_rows = FALSE, cluster_cols = FALSE,
         main = "Ranking Difference (L1 − L2)",
         labels_row = comparacao$ID)

# 8. Print correlation coefficients
print(paste("Pearson correlation of scores:", round(cor_pearson, 4)))
print(paste("Spearman correlation of ranks:", round(cor_spearman, 4)))



# Number of Monte Carlo simulations
N <- 50000

# Number of criteria minus 1 (since the last weight is computed to sum to 1)
dimension <- 4

# Generate a matrix of standard normal random numbers
X <- matrix(rnorm(N * dimension), nrow = dimension)

# Normalize the columns of X to unit vectors
norms <- apply(X, MARGIN = 2, function(x) sqrt(sum(x^2)))
Xnorm <- apply(X, MARGIN = 1, function(x) x / norms)

# Generate random radii within the unit hypersphere
radios <- runif(N) ^(1/dimension)

# Delta values to test the sensitivity of the weights
delta  = seq(from = 0.05, to = 0.4, by = 0.01)

# Initialize vectors to store results
media <- numeric(length(delta))
min_ci <- numeric(length(delta))
max_ci <- numeric(length(delta))

# Loop over delta values
for(j in seq_along(delta)){
  cat("Radius ", j, " of ", length(delta), "\n")

  # Generate random weight perturbations within radius delta[j]
  Xfinal <- delta[j] * (Xnorm * radios)
```

```r

  # Original weights (first n-1 criteria)
  w0 <- c(0.2, 0.2, 0.2, 0.2, 0.2)

  # Add perturbation to original weights
  W <- apply(Xfinal, MARGIN = 2, function(x) x + w0)

  # Calculate last weight to ensure they sum to 1
  C4 <- apply(W, MARGIN = 1, function(x) 1 - sum(x))

  # Final weight matrix with all criteria
  Pesos <- cbind(W, C4)

  # Number of simulations for each delta
  N <- 1000
  DeltaR <- numeric(N)  # Vector to store ranking differences

  # Monte Carlo loop
  for (i in 1:N) {
    cat("Iteration ", i, " / ", N, "\n")

    # Apply TOPSIS using perturbed weights
    resMC <- tryCatch(TOPSIS(rulesdf_TOPSIS,
                             directions = c("max", "max", "max", "max", "min"),
                             norm.method = "minmax",
                             w = Pesos[i,], tol = 1e-3,
                             ordered = FALSE, makefigure = FALSE),
                      error = function(e) return(NA))

    # Compute mean relative difference in scores
    if(!is.na(resMC)){
      difer <- abs(resMC$scores$TOPSIS - uwt2$scores$TOPSIS) / uwt2$scores$TOPSIS
      DeltaR[i] <- mean(difer)
    } else {
      DeltaR[i] <- NA
    }
  }

  # Compute confidence interval for mean variation
  confint <- t.test(DeltaR)
  media[j] <- confint$estimate
  min_ci[j] <- confint$conf.int[1]
  max_ci[j] <- confint$conf.int[2]
}

# Save results in a data frame
resultados <- data.frame(delta = delta,
                         media = media,
                         min_ci = min_ci,
                         max_ci = max_ci)

# Boxplot of last DeltaR sample
boxplot(DeltaR)

# Plot with confidence ribbon for mean variation
library(ggplot2)

p <- resultados %>%
```

```
  ggplot(aes(x = delta, y = media)) +
  geom_ribbon(aes(ymin = min_ci, ymax = max_ci), col = "gray", alpha = 0.4) +
  geom_line()

# Save plot to PDF
ggsave("sensibilidade.TOPSIS.PDF", plot = p, device = "pdf", width = 8, height = 5)

# Final analysis of the ranking variation
summary(DeltaR)           # Descriptive statistics
hist(DeltaR, breaks = 10, main = "Ranking Variation Distribution", xlab = "Mean
    Variation")
range(DeltaR)             # Minimum and maximum values
table(DeltaR)             # Frequency table
boxplot(DeltaR, main = "Boxplot of Ranking Variation", ylab ="DeltaR")

# Close PDF device if open (uncomment if needed)
# dev.off()

# Compare scores from two methods: uwTOPSIS and standard TOPSIS
res1_df <- data.frame(Rule = 1:49,
                      Score = uwt1$scores$uwTOPSIS,
                      Method = "uwTOPSIS")

res2_df <- data.frame(Rule = 1:49,
                      Score = uwt2$scores$TOPSIS,
                      Method = "TOPSIS")

res_df <- rbind(res1_df, res2_df)

# Line plot comparing scores
res_df %>% ggplot(aes(x = Rule, y = Score, col = Method)) +
  geom_line(aes(lty = Method)) +
  geom_point(aes(shape = Method)) +
  theme_bw() +
  scale_x_continuous(breaks = seq(from = 1, to = 49, by = 3),
                     labels = paste0("R", seq(from = 1, to = 49, by = 3)))

# Save the comparison plot
ggsave("uwT_vs_T.pdf", width = 8, height = 5)

# --- Verification of minimized rules ---
for (r in seq_along(reglas_def)){

  X1 <- reglas_def[[r]]
  long <- length(reglas_def[[r]][[1]])
  X1names <- lapply(X1, names)
  X1names_unicos <- unique(X1names)
  reglas_ok <- logical(length(X1names_unicos))

  for (i in seq_along(X1names_unicos)){
    regs <- X1names_unicos[[i]]
    idx_i <- which(sapply(X1names, function(x) all(x == regs)))
    X1_i_mat <- do.call(rbind, X1[idx_i])
    dists <- dist(X1_i_mat, method = "manhattan")
    reglas_ok[i] <- all(dists >  1)  # Check that all distances are greater than 1
  }

  reglas_length_ok <- all(reglas_ok)
```

```
676   cat("Rules of length ", long, " OK: ", reglas_length_ok ,"\n")
677 }
```