UNIVERSIDAD DE SALAMANCA
DEPARTAMENTO DE ESTADÍSTICA
DOCTORADO EN ESTADÍSTICA MULTIVARIANTE APLICADA

GRÁFICO DE CONTROL ESTADÍSTICO DE PROCESOS MULTIVARIANTES PARA VARIABLES CUALITATIVAS

TESIS DOCTORAL

WILSON JAVIER ROJAS PRECIADO

DIRECTORES:

OMAR HONORIO RUIZ BARZOLA
PURIFICACIÓN GALINDO VILLARDÓN

SALAMANCA, ESPAÑA
2023

GRÁFICO DE CONTROL ESTADÍSTICO DE PROCESOS MULTIVARIANTES PARA VARIABLES CUALITATIVAS

DEPARTAMENTO DE ESTADÍSTICA

Memoria para optar al Grado de Doctor en Estadística Multivariante Aplicada por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Wilson Javier Rojas Preciado

Salamanca 2023

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

DR. OMAR HONORIO RUIZ BARZOLA

PROFESOR TITULAR DE LA ESCUELA POLITÉCNICA DEL LITORAL, ECUADOR.

DRA. PURIFICACIÓN GALINDO VILLARDÓN

CATEDRÁTICA DEL DEPARTAMENTO DE ESTADÍSTICA DE LA UNIVERSIDAD DE SALAMANCA, ESPAÑA.

CERTIFICAN:

Que D. **WILSON JAVIER ROJAS PRECIADO**, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo para optar al Grado de Doctor en Estadística Multivariante Aplicada, que presenta con el título **GRÁFICO DE CONTROL ESTADÍSTICO DE PROCESOS MULTIVARIANTES PARA VARIABLES CUALITATIVAS**, autorizando expresamente su lectura y defensa, y para que conste, firman el presente certificado en Salamanca a _____.

Firmado electrónicamente por:
OMAR HONORIO RUIZ BARZOLA

Dr. Omar Honorio Ruiz Barzola

Firmado electrónicamente por:
PURIFICACION GALINDO VILLARDON

Dra. Ma. Purificación Galindo Villardón

"La calidad comienza con la educación y termina con la educación"

Kaoru Ishikawa

"En Dios confiamos, los demás, traigan datos"

W. Edwards Deming

Dedicatoria

A mis hijos, Paúl, Javier, Luis, María Gabriela, a mi esposa Pilar y a mis padres, Wilson y Florcita.

Agradecimientos

Empiezo por agradecer a Dios, mi Señor, sin Él nada es posible.

Mi gratitud para la Dra. Purificación Galindo Villardón, por todo el apoyo que recibí de ella desde el inicio de este sueño que se convierte en realidad, por sus conocimientos impartidos, por su predisposición a ayudar en las diversas situaciones propias de este proceso, ya en Salamanca, ya en Ecuador. Dra. Puri, no he conocido a una persona más apasionada por su trabajo que usted. Muchas gracias.

Al Dr. Omar Ruiz Barzola, mi sincero agradecimiento por su invaluable conocimiento y su incansable dedicación, que han sido bases fundamentales en este trabajo de investigación. La combinación magistral de rigor científico y amabilidad en su trato, hicieron de nuestras largas jornadas de trabajo una experiencia memorable. Estoy eternamente agradecido.

Cada objetivo alcanzado lleva impreso el amor, sacrificio y dedicación de mi familia. Mención especial merecen mis hijos, Paúl, Javier, Luis, María Gabriela y mi amada esposa, Pilar, la fe inquebrantable de ellos y su amor constante me han empujado a alcanzar alturas que jamás imaginé; mis padres, Wilson y Florcita, un ejemplo de vida; mi querida familia, que es fuente inagotable de amor y apoyo. Sin ustedes, cada logro y meta alcanzada no tendría el mismo significado. Gracias por ser el motor detrás de mis sueños.

Los logros más significativos no se evalúan únicamente por la consecución de metas establecidas, sino también por aquellas manos que proporcionaron apoyo inquebrantable y

orientación en el camino. Gracias a mis profesores, compañeros y amigos de la Universidad Técnica de Machala y de la Universidad de Salamanca, por ser esas manos para mí.

A todos quienes han aportado al logro de este tan alto propósito, mi eterno agradecimiento.

Resumen

Los gráficos de control, en el control estadístico de procesos, son esenciales para definir parámetros y límites óptimos en procesos de producción, y monitorizar la calidad de los productos al reducir la variabilidad. Si bien originalmente estos gráficos se centraban en la monitorización univariante, la complejidad organizacional ha impulsado el desarrollo de herramientas multivariantes, siendo el gráfico T^2 de Hotelling el más utilizado, aunque tiene sus limitaciones.

Esta investigación propone un enfoque innovador al integrar variables cualitativas en gráficos de control multivariantes, considerando que dichas variables desempeñan un papel fundamental en áreas como economía, psicología, educación, en procesos productivos, industriales. El objetivo central es el desarrollo de una metodología que permita el control de estas variables usando técnicas estadísticas multivariantes en la fase I del control estadístico de procesos.

El gráfico de control propuesto, se denomina **T2Qv**, como un acrónimo de $\mathbf{T^2}$ (gráfico de Hotelling), ***Q****ualitative* y ***V****ariables*. Surge como una herramienta adaptada a bases de datos cualitativas que, partiendo del estadístico T^2 de Hotelling, introduce el estadístico T^2_{med}, aprovechando el vector de medianas para mejorar la robustez. Esta metodología detecta anomalías y utilizando técnicas estadísticas multivariantes, como el Análisis de Correspondencias Múltiples y los Métodos biplot, facilita la interpretación de comportamientos variables y su relación con estados fuera de control.

Adicionalmente, se ha desarrollado un paquete estadístico computacional, T2Qv, en el lenguaje R, accesible a través del repositorio oficial de R, para ampliar la facilidad y difusión del método.

No obstante, el T2Qv presenta algunas limitaciones, como la necesidad de bases de datos con un mínimo de cuatro variables y la pérdida de estabilidad en dimensiones bajas. Como oportunidades futuras, se sugiere la optimización para la fase II y la inclusión de técnicas multivariantes avanzadas.

En conclusión, la investigación presenta un avance significativo en la incorporación de variables cualitativas en gráficos de control multivariantes, llenando un vacío en el ámbito de la estadística aplicada, especialmente beneficioso para procesos sociales y campos interdisciplinarios.

Índice de contenidos

Índice de tablas

Índice de figuras

Notación

k	Índice de tabla (k = 1, 2, …, K)
K	Número total de tablas (profundidad del cubo de datos)
$\boldsymbol{'}$	Índice de matriz transpuesta (**X**')
n	Tamaño muestral de las k tablas
$\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$	Matriz de datos conformada por n vectores $\boldsymbol{x_i}$, i = 1, ..., n
mf	Vector de masas de filas
mc	Vector de masas de columnas
$\boldsymbol{Z}$	Matriz disyuntiva de valores binarios
$\boldsymbol{B}$	Matriz de frecuencias absolutas de la Tabla de Burt
$\boldsymbol{P}$	Matriz de frecuencias relativas
$\boldsymbol{S}$	Matriz de residuos estandarizados
$\boldsymbol{D_{fila}}$	Matriz diagonal que contiene las masas de las filas
$\boldsymbol{D_{columna}}$	Matriz diagonal que contiene las masas de las columnas
$\boldsymbol{U}$	Matriz formada por los vectores singulares izquierdos de **X**
$\boldsymbol{V}$	Matriz formada por los vectores singulares derechos de **X**
$\boldsymbol{D}$	Matriz diagonal p-dimensional de los valores singulares de **X**
$\boldsymbol{C}$	Matriz de coordenadas
$\boldsymbol{C'}$	Matriz de coordenadas normalizadas
$\mathbb{C}'$	Matriz concatenada
$\tilde{\mathbf{x}}_{\mathbf{C}'_k}$	Vector de mediana de la k matriz normalizada

$\tilde{\mathbf{x}}_{\mathbb{C}'}$	Vector de mediana de la matriz concatenada
$\boldsymbol{\Sigma}$	Matriz de covarianzas
$\boldsymbol{\alpha}$	Nivel de significación
p	Número de dimensiones (p variables)

Abreviaturas

ARL	Promedio de la longitud de las rachas
MRL	Mediana de la Longitud de la Secuencia
SPC	Control estadístico de procesos
MFA	Análisis Factorial Múltiple
PCA	Análisis de Componentes Principales
MCA	Análisis de Correspondencias Múltiples
SVD	Descomposición de valores singulares
KDE	Estimación de Densidad de Kernel
MEWMA	Gráfico multivariante de control de medias ponderadas
MCUSUM	Gráfico multivariante de control de sumas acumuladas
UCL	Límite de control superior
LCL	Límite de control inferior
CL	Valor central

Introducción

Los gráficos de control constituyen una de las herramientas más importantes para definir límites y parámetros óptimos de proceso de producción, así como para controlar la calidad de los productos mediante la reducción de la variabilidad. El uso de gráficos de control facilita la evaluación del comportamiento de las variables del proceso y contribuye al logro de los objetivos planificados. En el control estadístico de procesos, el uso de gráficos de control es muy importante para reducir la variabilidad del proceso, establecer límites y parámetros óptimos y evaluar el comportamiento de las variables.

Originalmente, el Control estadístico de procesos se centraba en la monitorización de una única variable, pero la creciente complejidad de las organizaciones dio lugar al control estadístico de procesos multivariantes. En esta investigación se hace una revisión de los principales gráficos de control y sus aportes al control estadístico de procesos multivariantes.

El gráfico de control multivariante que ha sido más utilizado es el T^2 de Hotelling, que, aunque eficaz para detectar cambios grandes en la media del proceso, tiene desventajas, como la necesidad de un tamaño grande de muestras y su dependencia de distribuciones normales (Montgomery, 2012; Das, 2009). En consecuencia, en esta investigación se revisa el desarrollo de mejoras a su rendimiento, así como propuestas para abordar datos complejos y alta correlación entre variables, en contextos no paramétricos.

En procesos de alta dimensionalidad, la presencia de múltiples variables correlacionadas puede llevar a redundancia en la información y comprometer la viabilidad del análisis. En tales escenarios, los métodos de reducción de dimensiones, basados en técnicas estadísticas multivariantes como Análisis de Componentes Principales, Análisis Factorial, Análisis de Correspondencias Múltiples, son útiles para extraer variables latentes, combinaciones lineales de las variables originales. Estas variables latentes capturan la mayor parte de la variabilidad en el conjunto de datos, permitiendo una representación más eficiente del proceso analizado.

La literatura científica es numerosa respecto de los gráficos de control para datos numéricos y mixtos en contextos multivariantes, sin embargo, no ha sido muy generosa en el estudio de gráficos de control para datos cualitativos. Estas variables cualitativas, ya sean nominales u ordinales, son esenciales en campos como la economía, la psicología, la sociología, la educación, entre otros, sin descartar procesos industriales. Dado que estas variables pueden codificar información compleja y valiosa, los métodos para incorporarlas en gráficos de control multivariantes representan una frontera emergente en la estadística aplicada; su análisis podría aportar perspectivas significativas para la toma de decisiones. El desarrollo de técnicas que integren estos tipos de datos en el marco del control de la calidad podría llenar una brecha importante y enriquecer el análisis en importantes sectores de la sociedad.

Este estudio ofrece al investigador un análisis de los principales aportes respecto del desarrollo de los gráficos de control estadístico de procesos multivariantes, a través de la aplicación de métodos teóricos. Se presenta un recorrido por los conceptos más relevantes relacionados con los gráficos de control, también se hace una caracterización de

los principales tipos de gráficos de control, así como su evolución histórica, desde los tradicionales de Shewhart, en su contexto univariante, hasta los gráficos multivariantes que aplican técnicas de reducción de dimensiones.

Al analizar los procedimientos publicados por diversos autores respecto de este tema, se detectan limitaciones que podrían restringir su aplicación, por ejemplo, el análisis de pocas características de la calidad, el uso de muestras constituidas por elementos individuales en vez de grupos, la dificultad de trabajar con muchas categorías de forma simultánea. Surge, entonces, la necesidad de un gráfico de control para la representación de p variables cualitativas, que pueda trabajar con múltiples categorías nominales y ordinales y que facilite la identificación de las causas que pueden llevar al proceso a un estado fuera de control y que pueda ser aplicado en procesos sociales.

Esta investigación atiende las limitaciones antes mencionadas en cuanto a gráficos de control para variables cualitativas y su aplicación en diversos entornos. Por tal motivo, su *objetivo general* es Desarrollar una metodología para el control de variables cualitativas mediante el uso de técnicas estadísticas multivariantes, que contribuya a la diversificación de técnicas en la fase I del control estadístico de procesos.

Los objetivos específicos de esta investigación son:

1. Identificar los principales aportes al desarrollo de los gráficos de control estadístico de procesos multivariantes, que han contribuido a la fundamentación teórica de propuestas aplicables al control de la calidad.
2. Establecer un estadístico que contribuya el control de procesos para variables cualitativas en el contexto estadístico multivariante.

3. Diseñar un gráfico que facilite la interpretación del comportamiento de procesos multivariantes para variables cualitativas como un aporte a la diversificación de técnicas en la fase I del control estadístico de procesos.
4. Demostrar la eficacia del nuevo gráfico de control, a través de la aplicación del gráfico propuesto al control de procesos estadísticos multivariantes, con bases de datos simulados y reales.
5. Determinar los niveles aceptables de confiabilidad de resultados de la aplicación del gráfico propuesto, en procesos que impliquen reducción de dimensiones y variación del número de variables contaminadas, mediante un análisis de sensibilidad.
6. Identificar los principales desafíos u oportunidades de mejora con relación al desarrollo de los gráficos de control para el control estadístico de procesos multivariantes.

Para el logro de estos objetivos, es necesaria la aplicación del método Histórico-Lógico, empleado especialmente en el establecimiento de los antecedentes del problema científico y de los fundamentos teóricos y metodológicos de la propuesta de solución al problema. Lo histórico se dirige al estudio del objeto en su trayectoria a través de su historia, con sus condicionamientos sociales, económicos y políticos en los diferentes periodos, mientras que, lo lógico interpreta lo histórico e infiere conclusiones. Lo histórico y lo lógico están estrechamente vinculados, lo lógico para descubrir la esencia del objeto requiere los datos que le proporciona lo histórico (Rodríguez y Pérez, 2017).

En esta investigación se hace una revisión de la evolución de los gráficos de control estadístico multivariantes, estableciendo etapas en las que se identifica el marco cronológico, los principales autores que contribuyeron a su desarrollo, las características

de las propuestas más destacadas y otras condiciones que influyeron en los cambios durante los períodos estudiados, lo que permite conocer la lógica de su evolución.

Otro de los métodos teóricos que se utiliza en esta investigación es el Analítico – Sintético, que posibilita descomponer mentalmente un todo en sus partes y cualidades, en sus múltiples relaciones, propiedades y componentes, así como establecer la unión o combinación de las partes previamente analizadas para el descubrimiento de relaciones y características generales entre los elementos de la realidad. El método Analítico – Sintético se manifiesta en la identificación de los elementos esenciales de cada una de las etapas de la evolución de los gráficos de control, los autores, los años, los aportes más relevantes y la forma en que estos elementos se relacionan, así como en la identificación de regularidades que se sintetizan para el establecimiento de conclusiones.

Por otra parte, se aplica el método sistémico-estructural-funcional, el cual considera al objeto de estudio como una realidad única y compuesta, basándose en la interrelación e interdependencia entre las partes del todo. El enfoque estructural-funcional distingue los elementos esenciales de los secundarios y se dirige a modelar el objeto como sistema, determinando sus componentes, estructura, jerarquía y relaciones funcionales (Rodríguez y Pérez, 2017). En esta investigación, se aplica este método en el análisis de procesos caracterizados por diferentes variables interrelacionadas que, a su vez, están conformadas por categorías que se asocian. El grado de asociación que existe entre estas categorías puede incidir en el comportamiento final del proceso.

Como técnica se utiliza el análisis documental, un procedimiento científico sistemático que indaga, recolecta, organiza, analiza e interpreta información alrededor de un tema afín al objetivo de la investigación, para proporcionar soporte teórico al

desarrollo de estudios científicos (Martínez, Palacios y Oliva, 2023). En esta investigación se hizo el análisis de documentos como Tesis de Doctorado, Tesis de Maestrías, libros, artículos científicos indexados en bases de datos como WoS, Scopus, Scielo, Taylor & Francis, ScienceDirect, Redalyc, Latindex, Google Académico, entre otros, relacionados con el desarrollo de los gráficos de control para procesos estadísticos multivariantes, en un marco cronológico que conduce al lector desde los autores clásicos hasta los aportes más recientes.

La estructura de este artículo incluye al Capítulo I que se refiere a los antecedentes teóricos de los gráficos de control estadísticos de procesos multivariantes para variables cualitativas. El Capítulo II se refiere al desarrollo del Gráfico de Control de Procesos Estadísticos Multivariantes para Variables Cualitativas (T2Qv), que constituye la propuesta metodológica de esta investigación. El segundo capítulo contiene también la aplicación de la propuesta metodológica al análisis de un caso con datos simulados, así como el Análisis de sensibilidad. El capítulo III caracteriza el Complemento computacional, T2Qv, versión 0.2.0, reproducible en R. El Capítulo IV describe la aplicación del paquete T2Qv en el análisis de un caso real, en el contexto de la educación superior. Seguidamente, se presentan la Discusión y las Conclusiones y, por último, las Referencias bibliográficas.

Capítulo 1

Antecedentes teóricos de los gráficos de control estadístico de procesos multivariantes para variables cualitativas

1.1. Introducción

La importancia del uso de los gráficos de control radica en que facilitan la reducción de la variabilidad de los procesos, permiten el establecimiento de límites y parámetros óptimos, así como la evaluación del comportamiento de las variables que intervienen, contribuyendo al logro de los objetivos planificados (Montgomery, 2012).

En el control estadístico de procesos es necesario identificar tipos y causas de variabilidad, asumiendo que un grupo de variables puede generar una diversidad de resultados. En consecuencia, es necesario registrar de manera sistemática las diferentes características de la calidad a lo largo de las fases del proceso observado: propiedades de los insumos, estado y calibración de los equipos, competencias del personal, calidad y pertinencia de procedimientos, cumplimiento de especificaciones, tiempos de entrega, la satisfacción de los usuarios, costos de operación y de reprocesos, entre otras.

Walter Shewhart desarrolló las bases para el control estadístico de procesos, quien reconoció que en toda producción industrial hay variación de procesos. Aseguró que no pueden producirse dos partes con las mismas especificaciones, lo cual se debe, entre otras cosas, a las diferencias que se dan en la materia prima, a las diferentes habilidades de los

operadores y las condiciones en que se encuentra el equipo. Más aún, hay variación en las piezas producidas por un mismo operador y con la misma maquinaria (Herrera, et al., 2018).

Shewhart estableció las diferencias entre la variabilidad natural o común y la provocada por causas asignables o especiales. La primera se manifiesta en todos los procesos, aun en los que se encuentran bajo control, la segunda puede llevar los procesos a un estado de fuera de control. De ahí que, un proceso está en control estadístico cuando en él sólo se presentan causas comunes de variación. Shewhart originalmente las denominó como "Causas originales de variación" y "Causas asignables", más tarde Edward Deming, las denominó "causas normales" y "causas especiales".

Comprender la fuente de la variación de los datos a través del monitoreo del proceso es esencial para garantizar su estabilidad (Zhao, 2023). Los gráficos de control diferencian causas especiales de variación de las comunes. Una vez detectadas las causas especiales, sus razones se identifican rápidamente. Estas herramientas visuales son fáciles de usar, siendo eficaces para monitorear datos del proceso a lo largo del tiempo (Song et al. 2023).

Los primeros gráficos (cartas) de control fueron propuestos por Shewhart para facilitar la representación y el control de la variabilidad en variables de tipo continuo y de atributos, como consecuencia se lograba la mejora del proceso (Ruiz-Barzola, 2013).

Shewhart estableció dos fases en el control de los procesos: la primera, denominada Fase de desarrollo, que describe el comportamiento estadístico de la variable analizada, determina los límites de control para el estimador del parámetro analizado y contribuye a la eliminación de causas asignables o especiales de variabilidad. La segunda, Fase de madurez,

establece la capacidad del proceso para cumplir con los requisitos, identifica la media del número de muestras antes de obtener falsas alarmas y favorece la disminución del número de muestras para la detección de cambios en el proceso (Ruiz-Barzola, 2013).

Al respecto, Montgomery (2012) indica que la Fase I tiene como objetivo obtener un conjunto de observaciones bajo control a partir del análisis de los m grupos preliminares, de manera que se pueda establecer los límites de control para la Fase II, que es el seguimiento de la producción futura.

Además de la estabilización de los procesos mediante la identificación y eliminación de causas especiales de variabilidad, otras ventajas de los gráficos o cartas de control se refieren al análisis del comportamiento del proceso a través del tiempo, especialmente en las variables de salida, pero también en las de entrada y aun en las de control interno del mismo proceso; por otra parte, el uso de cartas de control contribuye a que se detecten nuevas oportunidades de mejora y a que las ya implementadas se mantengan (Gutiérrez y de la Vara, 2013).

Marín (2016) identifica varios pasos que se deben tener en cuenta para el diseño y la construcción de gráficos de control, tales como, establecimiento de los objetivos del control de la calidad, identificación de la característica a controlar, selección del tipo de gráfico de control, elaboración del plan de muestreo, recogida de los datos, establecimiento de los límites de control, representación gráfica de la variación de los procesos y el análisis e interpretación de los resultados.

Además, esta autora sostiene que, para saber si el estadístico calculado en el control de procesos está o no dentro de los límites, se puede aplicar un contraste de hipótesis, en el

que la hipótesis nula (H_0) supone que el proceso está bajo control en cada una de las muestras seleccionadas, mientras que, la hipótesis alternativa (H_1) implica que hay uno o más puntos fuera de los límites de control. El incumplimiento de la H_0 y la consecuente aceptación de H_1, significa que el proceso no está bajo control.

En consecuencia, podría ocurrir un error tipo I, expresado mediante el nivel de significación α, esto es, la probabilidad de tener una observación fuera de los límites, cuando el proceso está bajo control. También podría haber un error tipo II, expresado mediante β, la probabilidad de no identificar una observación fuera de los límites de control cuando sí lo está.

Según Hawkins y Deng (2010), el control estadístico de procesos (SPC) busca determinar si hay cambios en la distribución de un proceso en relación con un objetivo. Para esto es esencial distinguir entre variaciones por cambios reales en la distribución, causas asignables, y aquellas por errores aleatorios, causas por azar. Los gráficos de control ayudan en esta tarea. Si no hay cambios y el proceso está bajo control, la probabilidad de una falsa señal del gráfico debe ser baja y conocida. No obstante, si hay desviaciones y el proceso no está controlado, el gráfico debe identificarlas rápidamente. La efectividad de estos gráficos se mide a través de la longitud media de corrida (ARL).

Ruiz-Barzola (2013) señala que la norma más utilizada para medir el rendimiento de un gráfico de control es “el promedio de la longitud de las rachas” (ARL, por sus siglas en inglés); es decir, el promedio del número de muestras sucedidas hasta obtener una señal de fuera de control. Gutiérrez y de la Vara (2013) aseguran que el ARL permite la medición de la velocidad con la que un gráfico de control detecta un cambio. En situación de control

estadístico, ARL = $1/p$, donde p es la probabilidad de que un punto caiga fuera de los límites de control, aunque no haya ocurrido un cambio en el proceso, a esto se denomina una falsa alarma.

Marín (2016) asocia el ARL a las hipótesis de la prueba estadística e identifica un ARL_0 y un ARL_1, en términos de

$$ARL_0 = \frac{1}{\alpha} \tag{1.1}$$

$$ARL_1 = \frac{1}{1-\beta} \tag{1.2}$$

En consecuencia, ARL_0 será el indicador de detección de una falsa alarma, mientras que, ARL_1 sería el inverso de la potencia de contraste de la prueba de hipótesis, que en el contexto de los gráficos de control se denomina curva característica de operación o curva OC (Operating Characteristics) por sus siglas en inglés y su representación gráfica permite visualizar el riesgo β. Para Ross y Adams (2012), el ARL_0 denota el número promedio de observaciones entre detecciones falsas positivas, suponiendo que no ha ocurrido ningún cambio, mientras que, ARL_1, el retraso promedio antes de detectar un cambio de tamaño. Das (2009) y Song, et. al (2023) sugieren utilizar también la Longitud de Corrida Mediana (MRL) como una medida más robusta.

Además de identificar de los puntos fuera de control, es importante prestar atención a señales de alarma para detectar patrones o comportamientos anormales de la distribución, aunque las observaciones se encuentren dentro de los límites.

Chiñas, Vásquez y López (2014) definen, en un gráfico de control univariante, un patrón como una secuencia numérica de dimensión n; $X = [x_1, x_2, x_3, \dots, x_n]$, es decir, un patrón es el comportamiento que tiene la serie de datos producida por un proceso aleatorio y la generación de los elementos de X se representa por $X = \mu + n_t + d_t$, donde, X es el patrón de comportamiento de los datos; n_t es la variación especial; d_t es la variación natural; y, μ es la media del proceso.

Asumiendo que la distribución de las observaciones es normal y que se cumple el teorema del límite central, Evans y Lindsay (2015) distinguen señales de que el proceso está en control estadístico: los puntos en el gráfico fluctúan de forma aleatoria entre los límites de control sin patrones reconocibles; la cantidad de puntos por encima y por debajo de la línea central es aproximadamente la misma; los puntos caen aleatoriamente por encima y por debajo de la línea central; la mayoría de los puntos está cerca de la línea central y solo algunos están cerca de los límites de control.

Asimismo, estos autores advierten que pueden surgir varios tipos de patrones poco comunes, tales como puntos fuera de los límites de control; cambio súbito en la media del proceso; ciclos breves repetitivos que alternan picos elevados y valles bajos; tendencias o cambios graduales de forma ascendente o descendente en relación con la línea central; apiñamiento en la línea central; apiñamiento en los límites de control. Como complemento, Marín (2016) señala dos patrones poco comunes para tener en cuenta: la inestabilidad, grandes fluctuaciones que provocan que algún valor eventualmente caiga fuera de los límites de control, y la sobre estabilidad: variabilidad es menor de la esperada.

Las características de calidad a menudo se encuentran correlacionadas. Por lo tanto, se debe considerar la monitorización conjunta de varias características de calidad simultáneamente (Qiu, 2013). En principio el control estadístico de procesos realizaba la monitorizaciónde una variable a la vez, no obstante, por la creciente complejidad de las organizaciones fue necesario realizar un análisis de varias características de calidad de forma simultánea, así surgió el control estadístico de procesos desde una perspectiva multivariante (Ramos-Barberán, 2017; Li, Tsung, y Zou, 2012).

En dependencia del número de características de calidad analizadas de forma simultánea se identifica dos tipos de gráficos de control, univariantes y multivariantes. Los primeros, diseñados para una sola variable, están mucho más estudiados en la bibliografía especializada y constituyen el punto de partida para los gráficos multivariantes (Alfaro, Mondéjar y Vargas, 2010).

Una propiedad necesaria para que un esquema de monitoreo multivariante sea útil es su eficacia para localizar con precisión la fuente de las causas asignables cuando se detecta un cambio. Sin embargo, en escenarios multidimensionales, donde características de calidad están correlacionadas, determinar las variables causantes es complejo (Song et al., 2023).

Se han propuesto diversos métodos para identificar las p características de la calidad responsables de la señal de fuera de control en el proceso, entre otros: la clasificación de componentes de un vector de observación según su contribución relativa a una señal (Doganaksoy, Faltin y Tucker, 1991); uso de Análisis de Componentes Principales (PCA) (Jackson, 1991; Tracy, Young y Mason, 1992; Kourti y McGregor, 1996); ajustes de regresión para variables individuales (Hawkins, 1993); la descomposición del estadístico T^2

en p componentes independientes (Mason, Tracy y Young, 1995); análisis basados en el uso del Biplot Dinámico (c); el uso de redes neuronales artificiales (Ruelas, et al., 2020); aplicación de Análisis de Correspondencias Múltiples (MCA) y determinación de distancias χ^2 entre las masas de columnas de la tabla fuera de control y de la tabla tomada como referente (Rojas-Preciado, et al., 2023).

A continuación, se hace una revisión de la evolución de los gráficos de control estadístico de procesos multivariantes, estableciendo etapas en las que se identifica un referente cronológico, los principales autores que contribuyeron y las características esenciales de sus propuestas.

1.2. Evolución histórica de los gráficos de control estadístico de procesos multivariantes

A partir de la aplicación del análisis documental y los métodos teóricos se identifican cinco etapas en la evolución histórica de los gráficos de control estadístico de procesos multivariantes, las dos primeras aparecen de forma secuencial, como respuesta a las necesidades de representación de múltiples variables que no podían ser abarcadas por los gráficos originales de Shewhart, mientras que, las últimas tres recorren una trayectoria casi simultánea, cuyas fronteras en algunos casos se traslapan, pero, han dado paso al desarrollo paralelo de propuestas bien diferenciadas, cada una con diferentes contribuciones y aportes significativos. Estas etapas se analizan a continuación:

a. El gráfico T^2 de Hotelling.
b. Mejoras al rendimiento del gráfico de control T^2 de Hotelling.
c. Gráficos de control multivariante en entornos no paramétricos.

d. Gráficos de control multivariante para variables cualitativas.
e. Gráficos para el control estadístico de procesos con técnicas estadísticas multivariantes.

1.2.1. Gráfico T^2 de Hotelling

El procedimiento de control y supervisión de procesos, en entorno multivariante, más conocido es el gráfico de control T^2 de Hotelling. Es el equivalente del gráfico univariante de Shewhart. Según Montgomery (2012), se trata del seguimiento del vector medio del proceso en el que se supone que la distribución de probabilidad conjunta de las *p*-características de calidad (variables) es la Distribución Normal Multivariante.

Harold Hotelling (1933) introdujo un gráfico de control multivariante fundamentado en el estadístico T^2. Este estadístico representa una generalización multivariante del estadístico t de Student y está basado en la distancia de Mahalanobis (1936). Para su formulación, emplea el vector de medias y la matriz de covarianzas de una distribución normal multivariante.

La distancia de Mahalanobis es una medida que se utiliza para cuantificar la distancia entre dos variables aleatorias (***X***, ***Y***) *p*-dimensionales, con igual función de distribución de probabilidades y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, teniendo en cuenta la correlación entre las variables y las desviaciones estándar de esas variables. Se parte del supuesto de que $X \sim N(\mu, \sigma)$.

Según Mukhopadhyay (2008), cuando tenemos una matriz de datos donde las columnas representan variables y las filas representan objetos, una forma común de comparar dos filas, $\boldsymbol{X}_r$ y $\boldsymbol{X}_r$, es a través de la distancia euclidiana:

$$\left|\left|\boldsymbol{X}_r - \boldsymbol{X}_s\right|\right|^2 = (\boldsymbol{X}_r - \boldsymbol{X}_s)'(\boldsymbol{X}_r - \boldsymbol{X}_s). \tag{1.3}$$

Sin embargo, cuando la variación en $\boldsymbol{X}$ es estocástica, es más conveniente considerar una transformación: $\boldsymbol{Z}_r = \boldsymbol{S}^{-\frac{1}{2}}(\boldsymbol{X}_r - \overline{\boldsymbol{X}}), r = 1, 2, \dots, n$. Esta transformación elimina la correlación entre las variables y estandariza sus varianzas.

$$\boldsymbol{S} = \frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{X}_r - \overline{\boldsymbol{X}})\,(\boldsymbol{X}_r - \overline{\boldsymbol{X}})' \tag{1.4}$$

donde, $\boldsymbol{S}$ es la matriz de covarianza de los datos, y $\overline{\boldsymbol{X}}$ es el vector de medias.

Una vez realizada la transformación, se puede calcular la distancia euclidiana entre las filas transformadas. Una de las distancias más significativas tras esta transformación es la distancia de Mahalanobis:

$$D^2 = \left|\left|\boldsymbol{Z}_r - \boldsymbol{Z}_s\right|\right|^2 = (\boldsymbol{X}_r - \boldsymbol{X}_s)'\boldsymbol{S}^{-\frac{1}{2}}(\boldsymbol{X}_r - \boldsymbol{X}_s) \tag{1.5}$$

$$D = \sqrt{(\boldsymbol{X}_r - \boldsymbol{X}_s)'\boldsymbol{S}^{-\frac{1}{2}}(\boldsymbol{X}_r - \boldsymbol{X}_s)} \tag{1.6}$$

Dependiendo del contexto, la distancia de Mahalanobis puede definirse de diferentes maneras:

- **Distancia entre parámetros:** Si se tienen dos distribuciones, $\boldsymbol{X}$ con media $\boldsymbol{\mu_1}$ y varianza $\boldsymbol{\Sigma}$ e $\boldsymbol{Y}$ con media $\boldsymbol{\mu_2}$ y la misma varianza $\boldsymbol{\Sigma}$, entonces la distancia de Mahalanobis entre los parámetros $\boldsymbol{\mu_1}$ y $\boldsymbol{\mu_2}$ se denota como $\boldsymbol{D_{\mu_1\mu_2}}$. Esta distancia refleja cuán diferentes son las dos distribuciones en términos de sus parámetros.
- **Distancia entre una variable y su media:** Si se tiene una variable aleatoria $\boldsymbol{X}$ con media μ y varianza $\boldsymbol{\Sigma}$, la distancia de Mahalanobis entre $\boldsymbol{X}$ y su media μ se define como D_{X_μ}. En este caso, la distancia de Mahalanobis es una variable aleatoria que mide cuán lejos está un punto observado de la media de su distribución.
- **Distancia entre dos variables aleatorias:** Si se tienen dos variables aleatorias, $\boldsymbol{X}$ con media $\boldsymbol{\mu_1}$ y varianza $\boldsymbol{\Sigma}$ e $\boldsymbol{Y}$ con media $\boldsymbol{\mu_2}$ y la misma varianza $\boldsymbol{\Sigma}$, entonces la distancia de Mahalanobis entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ se denota como $\boldsymbol{D_{XY}}$. Esta distancia refleja la diferencia entre las dos variables aleatorias en el espacio multivariante, teniendo en cuenta la estructura de correlación de las variables.

Ruiz-Barzola (2013) afirma que si $\boldsymbol{X'} = [X_1, X_2, \dots, X_P]$ con vector de medias $\mu'_x = [\mu_1, \mu_2, \dots, \mu_p]$ y matriz de covarianzas Σ_{pxp} (simétrica y definida positiva), el cuadrado de la distancia estandarizada de $\boldsymbol{X}$ a $\boldsymbol{\mu}$, es decir, el cuadrado de la distancia de Mahalanobis, es

$$d_m^2 = (\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{1.7}$$

Por otra parte, si $X \sim N(\mu, \sigma)$, $X^2 = \left(\frac{x-\mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu) \sim \chi^2_{gl=1}$

En consecuencia, si se toma una muestra aleatoria de la población $\boldsymbol{X} \sim N_p(\mu, \Sigma)$, el estadístico de prueba será

$$\chi_0^2 = n(\boldsymbol{X} - \mu)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \mu) = nd_m^2. \quad (1.8)$$

El estadístico T^2 de Hotelling tiene la siguiente relación con la distancia de Mahalanobis:

$$T^2 = nd_m^2(\boldsymbol{X}, \boldsymbol{Y}) \quad (1.9)$$

donde n es el tamaño de muestra y $d_m^2(\boldsymbol{X}, \boldsymbol{Y})$, el cuadrado de la distancia de Mahalanobis.

Cada componente de la distancia de Mahalanobis $(\boldsymbol{X} - \boldsymbol{Y})$ es análogo a una $Z \sim N(0, 1)$. Hazewinkel (2001) expresa esta relación explicando que si $Z_1, Z_2, \ldots, Z_k$ son variables aleatorias independientes, tales que $Z \sim N(0, 1)$ para $i = 1, 2, \ldots, k$, entonces, la variable aleatoria X definida por $X = Z_1^2 + Z_2^2 + \cdots + Z_k^2$

$$X = \sum_{i=1}^{k} Z_i^2 \quad (1.10)$$

tiene una distribución chi cuadrado con k grados de libertad.

Según Ruiz-Barzola (2013), el estadístico T^2 de Hotelling está dado por:

$$T^2 = n(\mu_k - \mu_0)'\Sigma_0^{-1}(\mu_k - \mu_0) \quad (1.11)$$

donde, n es el número de filas (individuos); μ_k es el vector de medias de la k-ésima muestra; μ_0, el vector de media de medias; y, Σ_0^{-1},la matriz de varianzas y covarianzas. Estos son parámetros de un proceso bajo control. Si no están disponibles los parámetros originales se deben estimar: $\hat{\Sigma} = \boldsymbol{S}$ y $\hat{\mu} = \bar{\bar{X}}$, por lo que, el estadístico cambia a

$$T^2 = n\left(\bar{\boldsymbol{X}} - \bar{\bar{\boldsymbol{X}}}\right)' \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{X}} - \bar{\bar{\boldsymbol{X}}}\right) \tag{1.12}$$

Cuando el proceso está bajo control, $\mu_k = \mu_0$, hay una probabilidad α que el estadístico exceda el UCL = $\chi^2_{p,\alpha}$, por tanto, la probabilidad de error (tipo I) se puede fijar al nivel α. Si el estadístico T^2 de alguna muestra supera este límite significa que el proceso está fuera de control. Cada muestra está representada por un punto en el gráfico

El gráfico T^2 de Hotelling (Figura 1.1) tiene un límite de control superior (UCL) = $\chi^2_{\alpha,p}$ y límite de control inferior (LCL) = 0; en donde, α es el nivel de significancia y p, el número de variables monitorizadas. Este gráfico mide la discrepancia entre vectores de promedios esperados y observados, considerando la matriz de covarianzas y utilizando el límite de control (UCL) para determinar si la distancia entre los dos vectores es lo suficientemente grande como para declarar el proceso fuera de control.

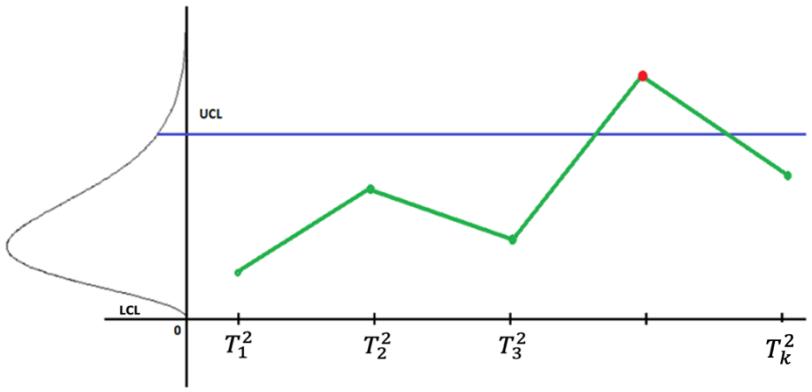


Figura 1.1 Gráfico T^2 de Hotelling

En el gráfico T^2 de Hotelling, cada muestra expresa el valor del estadístico T_k^2 y está representada por un punto; LCL se representa mediante una línea horizontal en el origen y UCL, mediante una línea horizontal por encima de LCL y depende de $\chi_{\alpha,p}^2$. Si alguna muestra $T_k^2 > UCL$, se representa como un punto sobre la línea UCL, lo que significa que el proceso está fuera de control, mientras que, si las muestras $T_k^2 < UCL$, sus puntos se grafican por debajo de la línea UCL, lo que significa el proceso está en control.

Cuando el proceso ha sufrido una desviación en al menos uno de los promedios de sus variables, el vector $\boldsymbol{\mu}_k$ se aleja del vector de promedios $\boldsymbol{\mu}_0$ del estado bajo control en una distancia d de Mahalanobis, entonces T^2 tiene una distribución chi-cuadrado no centralizada con p grados de libertad y con descentrado

$$\lambda = nd^2 = n(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0) \tag{1.13}$$

donde, n es el número de observaciones o tamaño de la muestra y **Σ,** la matriz de covarianzas de X durante el estado bajo control del proceso.

Entre las ventajas más relevantes del gráfico de control T^2 de Hotelling están la simplicidad de su procedimiento (Aparisi, 1996), que ofrece una mejor representación de los cambios en el proceso cuando se lo compara con el gráfico denominado elipse de control y que, para su aplicación, los datos pueden estar organizados en subgrupos o pueden ser observaciones individuales (Montgomery, 2012). Este gráfico puede monitorizar, de forma simultánea, múltiples características del proceso correlacionadas.

Como desventaja se puede señalar que, aunque el gráfico T^2 de Hotelling es eficaz para la detección de cambios grandes en la centralidad del proceso, pierde eficacia para

detectar cambios pequeños (Aparisi y García-Díaz, 2001; Alfaro, Mondéjar y Vargas, 2010). Otra desventaja consiste en la necesidad de disponer de un tamaño grande de muestras para el establecimiento de los límites de control, en esto coinciden Lowry y Montgomery (1995) y Jensen et al. (2006); además, el análisis de un número grande de muestras puede incurrir en elevados costos (Ruiz-Barzola, 2013).

Por otra parte, el estadístico T^2 de Hotelling se ve fácilmente afectado por la existencia de valores atípicos (Shabbak y Midi, 2012). Finalmente, el gráfico de control T^2 de Hotelling tiene una marcada vocación por distribuciones normales en las variables analizadas, requisito que en muchos casos no se cumple.

1.2.2. Mejoras al rendimiento del gráfico de control T^2 de Hotelling

Se ha escrito abundante literatura respecto de la naturaleza y los usos del gráfico T^2 de Hotelling. También se hicieron aportes dirigidos a la superación de sus limitaciones y a la mejora de su rendimiento.

1.2.2.1. Gráficos con memoria.- Propuestas como el gráfico de control de Suma Acumulada Multivariante (MCUSUM) (Crosier, 1988; Pignatiello y Runger, 1990) y el gráfico de control de Media Móvil Ponderada Exponencialmente Multivariante (MEWMA, por sus siglas en inglés) (Lowry et al., 1992), versiones multivariantes del gráfico de Sumas Acumuladas (CUSUM) (Page, 1954) y del gráfico de Promedios Móviles Exponencialmente Ponderados (EWMA) (Roberts, 1959), respectivamente, se diseñaron para el análisis de información histórica más allá de la última observación, lo que aumenta la sensibilidad en la detección de pequeños cambios en la media del proceso, por esto se les

denomina “gráficos con memoria” (García, 2014), superando con ello la falta de memoria del gráfico T^2 de Hotelling.

1.2.2.2. Gráfico T^2 con tamaño de muestra adaptable.- En ocasiones, tomar todos los elementos de la muestra para el análisis multivariante de un proceso se vuelve inconveniente, porque dilata los tiempos e incrementa costos de muestreo. Ruiz-Barzola (2013) señala que algunas actividades de muestreo resultan costosas o difíciles de medir y que es posible que la medición sea destructiva. Surge entonces la necesidad de mejorar la eficiencia del análisis.

En este sentido, Aparisi (1996) propuso optimizar el rendimiento del gráfico de control multivariante T^2 de Hotelling a través de su gráfico T^2 con tamaño de muestra adaptable. En su propuesta se utiliza un tamaño de muestra pequeño (n_i) para calcular T_k^2, mientras el estadístico está entre 0 y el límite de alerta (UWL), pero, si su valor está entre UWL y UCL, se utiliza una muestra más grande, de tamaño n_2 (Figura 1.2).

Si $T_k^2 > UCL$ se entenderá que el proceso está fuera de control. El gráfico de control T^2 con tamaños de muestra adaptables se grafica con límite de control $CL = \chi_{p,\alpha}^2$, utiliza un valor de α = 0,005 y tamaños de muestras: n_1 y n_2 ($n_2 > n_1$).

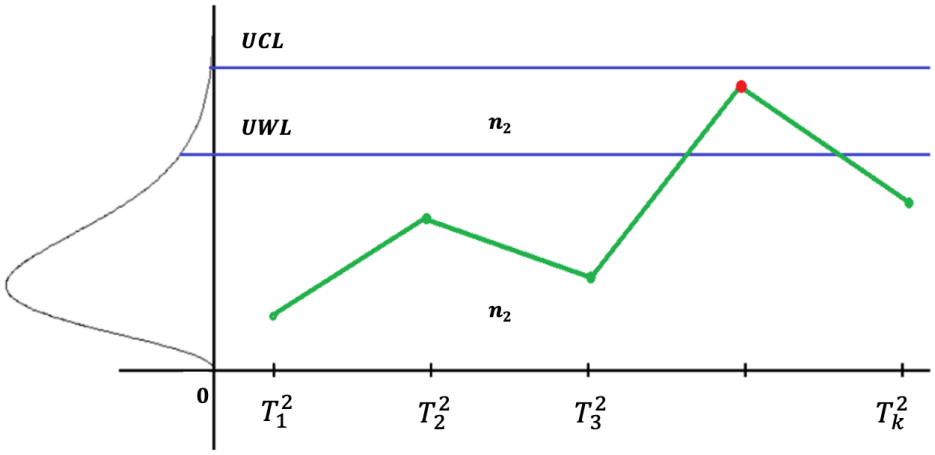


Figura 1.2. Gráfico T^2 de Hotelling con tamaños de muestra adaptativos

1.2.2.3. Gráfico de control de doble dimensión.- Ruiz-Barzola (2013) aborda el problema desde un nuevo esquema de muestreo para ser utilizado con el gráfico de control T^2, el gráfico de control de doble dimensión (DDT^2) y el Gráfico de control con dimensión variable (VDT^2).

En cada evento de muestreo, las p_1 variables que son baratas o rápidas de medir se evalúan y se determina el valor del estadístico T_{p1}^2. Si su valor es inferior a un límite de advertencia (*w*) el proceso se considera bajo control. Si su valor se ubica por encima del límite de control, el proceso (*CLp*) se considera fuera de control. Cuando el valor del estadístico está entre *w* y *CLp*, se miden las p_2 variables restantes y el estadístico global se compara con el límite de control apropiado (Figura 1.3), esto porque son diferentes los límites de control para el conjunto de variables p_1 de bajo costo y para el p conjunto total de variables ($p = p_1 + p_2$).

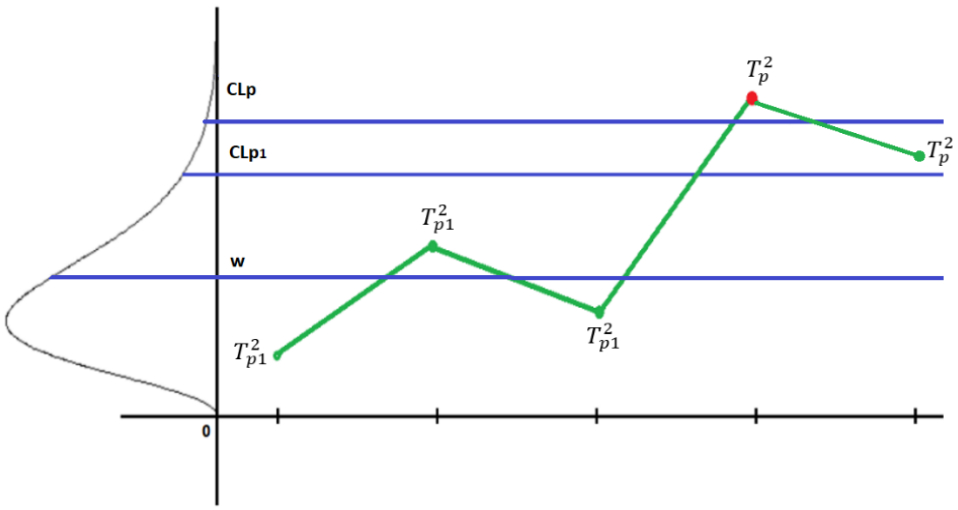


Figura 1.3. Gráfico de control DDT^2

Este procedimiento es similar al que se aplica en el gráfico de control de doble muestreo (DS), pero, se diferencia en que las mediciones de las p_2 características restantes son de otras variables en vez de constituir una muestra complementaria de una única variable. Las variables p_2, que son difíciles o costosas de medir, se miden solo cuando se necesita tomar más información.

Por consiguiente, el costo de muestreo con el gráfico de control de doble dimensión (DDT^2) disminuye en comparación con el costo de muestrear siempre todas las p variables. No obstante, dado que la información proporcionada por las p_2 características de la calidad aumenta significativamente la capacidad de detectar cambios en el proceso, la potencia del gráfico que utiliza solo p_1 variables es menor que la de un gráfico T^2 que utiliza todas las p variables.

1.2.2.4. Gráfico de control de dimensión variable.- Ruiz-Barzola (2013) propuso el gráfico de control T^2 para utilizar cuando el número de variables aleatorias que se debe evaluar es variable. La dimensionalidad de este gráfico depende de la complejidad que representa medir un grupo adicional de variables (p_2) de un proceso que tiene $p = p_1 + p_2$ variables.

De manera similar a lo que ocurre en el gráfico de control DDT^2, las p_1 variables corresponden a las menos costosas o menos complejas de medir, mientras que, las p_2 variables son más complejas o más costosas de monitorizar y solo se utilizarán cuando el proceso señale riesgo de estado fuera de control. Este método contribuye a optimizar el uso de recursos mejorando la eficiencia en el muestreo y es especialmente útil cuando el muestreo requerido para las p_2 variables es destructivo.

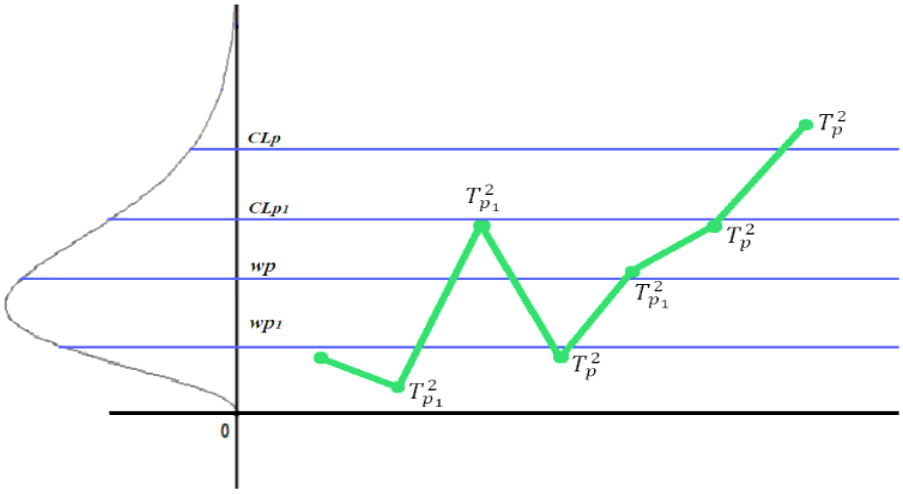


Figura 1.4. Gráfico de control con dimensión variable (VDT^2)

Este gráfico de control cuenta con cuatro parámetros, un límite de alerta w_{p_1}y un límite de control CL_{p_1} para el caso del estadístico T^2 calculado solo con variables p_1. Además, un límite de alerta w_p y otro de control CL_p que se aplican cuando se requiere el uso de todas las $p = p_1 + p_2$ variables (Figura 1.4). Hay otra versión más simple de este gráfico, que tiene un solo límite de alerta ($w_1 = w_2 = w$), por lo que solo cuenta con tres parámetros.

Para la evaluación de sus gráficos de control, este autor sugiere dos medidas de rendimiento: el costo medio por muestra cuando el proceso está bajo control y el número promedio de muestras necesarias para detectar un cambio en el proceso (ARL).

1.2.2.5. Gráfico T^2 de Hotelling que incorpora técnicas multivariantes y estimación de densidad de Kernel.- Mashuri, et al. (2021), proponen un enfoque integrado que combina el Análisis de Componentes Principales (PCA) y el gráfico T^2 de Hotelling para mejorar la detección de intrusiones en redes.

El primer paso en PCA es estimar la matriz de covarianza de los datos denotados como *C* con dimensión $p\ x\ p$ usando la ecuación en forma de matriz:

$$\boldsymbol{C} = \frac{1}{n} X'X. \tag{1.14}$$

Seguidamente, se descompone la matriz **C** para obtener el valor propio y el vector propio: $\boldsymbol{C} = \mathbf{V}\Lambda\mathbf{V}'$, donde $\mathbf{V} = \left(v_1, \dots, v_p\right)$ es la matriz de vectores propios con $p\ x\ p$ dimensiones y $\boldsymbol{\Lambda} = diag\left(\lambda_1, \dots, \lambda_p\right)$ es la matriz diagonal de valores propios. El estadístico

T^2 basado en PCA emplea los primeros k componentes principales para construir un gráfico de control, cuyo estadístico se escribe como:

$$T_{i,PCA}^2 = (\boldsymbol{y}_i - \overline{\boldsymbol{y}})' \boldsymbol{\Lambda}^{-1} (\boldsymbol{y}_i - \overline{\boldsymbol{y}}). \tag{1.15}$$

El algoritmo de Determinante de Covarianza Mínima Rápida (FMCD), que usa iteración de la distancia de Mahalanobis para calcular el vector medio robusto $\bar{y}_{n_t}$ y la matriz de covarianza S_{n_t}, se utiliza para estimar de manera robusta el vector medio y la matriz de covarianza, resultando el estadístico

$$\tilde{T}_{i,FMCD}^2 = \left(y_i - \bar{y}_{n_t}\right)' {S_{n_t}}^{-1} \left(y_i - \bar{y}_{n_t}\right). \tag{1.16}$$

Dada la dificultad de que los datos de tráfico de la red sigan una distribución normal multivariante, se emplea el procedimiento de Estimación de Densidad del Núcleo (KDE) para establecer un límite de control óptimo. Este límite de control considera la función de distribución de $\widehat{f_h}(t)$, y se denota por $CL_{KDE} = \hat{F}_h^{-1}(t)(1-\alpha)$.

Los autores aseguran que el enfoque propuesto supera a los métodos convencionales y otros clasificadores en términos de precisión y tasa de falsos negativos. Este enfoque también se compara favorablemente con alternativas anteriores como PCA-Bootstrap y PCA-KDE, mostrando aplicabilidad tanto en contextos industriales como en ciberseguridad.

Otras propuestas de mejora de la eficacia y la eficiencia del gráfico T^2 de Hotelling se han desarrollado, pero, por su propia naturaleza, se pueden clasificar también dentro de otras categorías en la evolución histórica de los gráficos de control estadístico de procesos multivariantes, en consecuencia, se describen más adelante en este documento.

Entre estas propuestas está el Gráfico de control multivariante T^2 basado en bootstrap (Phaladiganon et al., 2011), el Gráfico χ^2 para distribuciones multivariantes binomiales y de Poisson (Patel, 1973), Gráfico D^2 de control de atributos multivariante (Mukhopadhyay, 2008), Gráfico de control de calidad difuso multivariante (Taleb, Limam y Hirota, 2006), Gráfico de control multivariante basado en la combinación de PCA para características de calidad de atributos y variables (Ahsan, et al., 2018), gráfico T^2 basado en PCA Mix con límite de control de densidad de Kernel para mezclas de datos categóricos y continuos (Ahsan, et al., 2022), PCA Mix Chart para monitorear las características de calidad mixta en presencia de datos atípicos (Ahsan, et al., 2021), Gráfico de control multivariante no paramétrico para variables numéricas y categóricas (Jin y Loosveldt, 2022), el Gráfico de control estadístico de procesos multivariantes para datos cualitativos (Rojas-Preciado, et al., 2023).

1.2.3. Gráficos de control estadístico multivariante en entornos no paramétricos

En la evolución del control estadístico de procesos multivariantes, se ha observado que los gráficos del tipo T^2 son efectivos en los procesos tradicionales que generan datos numéricos independientes con distribución normal, pero no pueden procesar los datos complejos de alta dimensión que se encuentran frecuentemente en los sistemas modernos de producción (Liu, Liu y Jung, 2020). En esto coinciden autores como Chiñas, Vásquez y López (2014); Chakraborti, Van Der Laan y ST Bakir (2001); Sun y Tsung (2003); Pacella, Semeraro y Anglania (2004). Por ello se han realizado numerosos estudios para el desarrollo de gráficos de control multivariantes en contextos no paramétricos.

En relación con estos gráficos, se pueden distinguir dos grupos, según su utilización en procesos de dimensión moderada o de alta dimensión, los de alta dimensión han ganado terreno en periodos más recientes. A continuación, se describen estos gráficos.

1.2.3.1. Gráficos de control no paramétricos para procesos de dimensión moderada.- Qiu (2008) señala que, comúnmente en aplicaciones prácticas, la distribución los datos es desconocida y las mediciones multivariantes no son gaussianas, en particular con variables discretas. Las herramientas para datos no gaussianos multivariantes son escasas. Qiu aporta dos contribuciones a la literatura de Control Estadístico de Procesos (SPC). La primera es un método para estimar la distribución de mediciones en control, no Gaussianas y *p*-variantes, transformando cada componente de medición en una variable binaria según su mediana en control. Luego usa un modelo log-lineal para describir asociaciones entre estas variables binarias en la Fase I del control estadístico. Esta idea se puede generalizar a variables categóricas con $q \geq 2$ categorías. La segunda, sugiere un procedimiento MCUSUM para detectar cambios en el vector de parámetros en la Fase II del SPC, que es libre de distribución y apto para la mayoría de los problemas multivariantes de SPC.

En esta misma intención, Das (2009) propone un nuevo gráfico de control multivariante no paramétrico basado en la prueba de signo bivariante para superar las limitaciones del gráfico Hotelling T^2 cuando los datos no siguen una distribución normal multivariante. El nuevo gráfico de control es no paramétrico, lo que elimina la necesidad de suposiciones distribucionales. Se evalúa el rendimiento del nuevo método mediante el ARL y la "fracción de clasificación correcta", en estados de control y fuera de control.

El método propuesto muestra una mejor ARL en estado de control en comparación con Hotelling T^2, aunque su rendimiento disminuye al aumentar la tasa de error de Tipo-I, sin embargo, es levemente inferior en estados fuera de control. Se critica el uso exclusivo de ARL debido a su sesgo y se sugiere usar también la Mediana de la Longitud de la Secuencia (MRL) y la desviación estándar de la Longitud de la Secuencia (sdRL). El método es eficaz especialmente para muestras de tamaño mayor a 10 y solo en casos bivariantes.

Hawkins y Deng (2010) manifiestan que la asunción de distribuciones conocidas en controles es una idealización, en el mejor de los casos aproximadamente verdadera. El desarrollo reciente de métodos de punto de cambio basados en la normalidad ha permitido relajar la suposición de la media y la varianza en control exactamente conocidas, pero ha mantenido la suposición de normalidad. Estos autores presentan el gráfico de control de punto de cambio no paramétrico, y aseguran que ha demostrado ser efectivo, superando incluso métodos paramétricos en ciertas condiciones. Esta técnica, con mínimos supuestos y requerimientos, ofrece una alta flexibilidad frente a incertidumbres distribucionales y se constituye en una buena alternativa frente a métodos anteriores en el control de procesos de la Fase II.

Phaladiganon (2011) propuso un gráfico de control multivariante T^2 basado en bootstrap, método de remuestreo propuesto por Efron (1979), que facilita un monitoreo eficiente de un proceso cuando la distribución de los datos observados no es normal o es desconocida. Al hacer un análisis comparativo con datos simulados, los resultados mostraron que este gráfico tenía un mejor rendimiento que el T^2 de Hotelling tradicional y similar al gráfico de control T^2 basado en la estimación de la densidad del Kernel (KDE).

Según Shabbak y Midi (2012), la detección de valores atípicos correctos en la Fase I es importante para una especificación correcta del modelo. El estadístico T^2 de Hotelling no detecta múltiples valores atípicos de dispersión para observaciones individuales y, aunque sí lo hace cuando están en número pequeño, puede sufrir un efecto de enmascaramiento, en consecuencia, se vuelve menos eficaz para detectar valores atípicos, tanto de dispersión como en los cambios de paso sostenidos en el vector medio.

Como alternativa, se propone un gráfico de control robusto basado en el procedimiento de potencial generalizado robusto para el diagnóstico (DRGP) y Límites de Control Superior (UCL) calculados a partir de la mediana y la Desviación Absoluta Mediana (MAD). Se enfoca en la Fase I del esquema de monitoreo, donde la detección adecuada de atípicos es crucial para evitar la incorrecta especificación del modelo en la Fase II. Se introduce una métrica que no solo cuenta el número de atípicos detectados, sino también su correcta ubicación en el conjunto de datos. Los hallazgos son aplicables principalmente a conjuntos de datos de dimensiones moderadas con distribuciones no paramétricas.

Para Ross y Adams (2012), las tablas de control tradicionales a menudo carecen de datos suficientes para determinar la distribución previa al cambio en un proceso. En situaciones con muestras pequeñas o inexistentes, estimar parámetros es complicado. Por ello, se requieren gráficos no paramétricos que no asuman una distribución específica, pero mantengan un rendimiento constante. Ross y Adams desarrollaron dos gráficos que detectan variaciones arbitrarias en la distribución durante la Fase II, el CvM CPM y KS CPM, respectivamente. Basados en el modelo de punto de cambio (CPM), que tradicionalmente se ha utilizado para detectar cambios durante la Fase I, estos gráficos se adaptan para detectar

más tipos de cambios, incorporando pruebas "ómnibus", como Cramer-von Mises y Kolmogorov-Smirnov.

Xue y Qiu (2020) notaron que el control de procesos estadísticos multivariantes asume que las observaciones son independientes y tienen una distribución paramétrica, como la gaussiana, cuando el proceso está controlado. Los gráficos de control no paramétricos actuales asumen independencia de observaciones, supuesto a menudo no se cumple, dado que son comunes las correlaciones de datos en series temporales. En consecuencia, propusieron un gráfico de control de procesos multivariantes no paramétrico flexible que puede acomodar correctamente la correlación de datos en serie estacionaria.

1.2.3.2. Gráficos de control no paramétricos para procesos de alta dimensión.- Tuerhong y Kim (2014) propusieron un gráfico de control multivariante no paramétrico basado en la distancia de Gower, que puede manejar eficientemente los procesos caracterizados por una mezcla de datos continuos y categóricos en dimensiones elevadas. El coeficiente de disimilaridad de Gower es el promedio ponderado de las distancias calculadas lejos de cada variable después de escalar cada variable, se expresa en una escala que va desde 0 a 1. Los límites de control se calculan mediante *bootstrapping percentile*, método de remuestreo utilizado cuando se desconoce la distribución de la población subyacente. Los resultados revelaron que, cuando se incrementa el número de variables categóricas, este gráfico superó el rendimiento de otros.

Yue y Liu (2017) desarrollaron un gráfico de control de media móvil adaptable, multivariante, no paramétrico y ponderado exponencialmente con un intervalo de muestreo variable, que utiliza el concepto de profundidad de Mahalanobis para abordar un proceso

multivariante y reducir cada medición a un índice univariante. Este es un gráfico adaptativo y tiene un intervalo de muestreo variable que facilita la detección de varias magnitudes de cambios. La propuesta de Yue y Liu empieza por una tabla de control EWMA adaptativa no paramétrica para procesos multivariantes (MANE). Luego, incorpora un intervalo de muestreo variable (VSI) al gráfico MANE, por lo que el gráfico se denomina VSI-MANE.

Otra propuesta interesante es la de Liu, Liu y Jung (2020), quienes señalaron que la mayoría de los gráficos de control existentes no pueden manejar de manera eficiente situaciones con patrones de observaciones no lineales o multimodales. Estos investigadores desarrollaron un gráfico de control de peso de novedad sensible a la densidad (DNW) que mide el grado relativo de novedad sensible a la densidad como estadístico de seguimiento, utilizando el algoritmo de vecinos más cercanos (kNN), que puede monitorear de manera eficiente un proceso cuando se desconoce la distribución de las observaciones. Los límites de control se calculan a partir del percentil del estadístico DNW derivado de las muestras Bootstrap.

Según Li, Pei y Wu (2020), las gráficas de control multivariante originalmente se diseñaron para datos con distribuciones normales. Aplicar estos gráficos a datos no normales no es óptimo y aunque transformar estos datos podría permitir su uso con gráficos para distribuciones normales, encontrar una transformación adecuada es complicado. Por ello, la tendencia ha sido usar pruebas no paramétricas multivariantes que utilizan clasificación, análisis de rangos, estadísticas de orden, entre otras técnicas, como la prueba de corridas, para detectar tendencias o diferencias entre grupos de datos. Li, Pei y Wu presentaron el HAMEWMA, un gráfico de control multivariante no paramétrico, que se fundamenta en la

prueba de corridas y el camino hamiltoniano más corto, siendo apto para datos de alta dimensión con distribuciones desconocidas.

Adegoke et al. (2022) manifiestan que, en el control estadístico de procesos, las necesidades actuales implican variables aleatorias discretas o combinaciones de variables discretas y continuas, para las cuales los gráficos tradicionales son inadecuados. Estos autores propusieron el Gráfico de covarianza multivariante no paramétrico para la monitorización de observaciones individuales, tipo Shewhart, que proyecta los datos en el espacio euclidiano y emplea una relación de probabilidad específica para garantizar una estimación robusta de la matriz de covarianza. No asume modelos paramétricos y se adapta a cualquier medida de distancia. Mediante un procedimiento de arranque, establece límites de control, demostrando su eficacia en la detección de cambios en la matriz de covarianza según simulaciones realizadas.

Merlo et al. (2022) destacan que la normalidad multivariante rara vez se observa en situaciones prácticas, lo que eleva la relevancia de los esquemas no paramétricos. Proponen un nuevo gráfico multivariante que combina la distancia de Mahalanobis con el estadístico de Mann-Whitney, diseñado para monitorear flujos de datos multivariantes, especialmente en procesos de alta dimensión. Este gráfico utiliza estadísticas de prueba de dos muestras para comparar medidas de distancia. Considera la variación entre profesionales, abordando la incertidumbre asociada con la longitud de ejecución. Utiliza la probabilidad de falsa alarma condicional acumulativa para gestionar la complejidad computacional, facilitando una rápida determinación de límites de control, algo poco explorado en la monitorización no paramétrica.

Song et al. (2023) presentan métodos avanzados de monitoreo de procesos estadísticos multivariantes, focalizándose en cambios de estructura de dependencia y en atributos de calidad del producto. Emplean pruebas de Lepage y Cucconi basadas en rangos, integradas en un esquema MEWMA, que resultan robustas y casi libres de suposiciones distribucionales. La propuesta incluye dos variantes: MEWMA-LC y MEWMA-CC, que monitorean tanto las distribuciones marginales como la cópula. Se adopta la Mediana de Longitud de Ejecución (MRL) como métrica de rendimiento robusta, abordando limitaciones inherentes al ARL.

Los esquemas propuestos, según estos investigadores, superan a alternativas no paramétricas en la detección de cambios de escala y ofrecen mecanismos de seguimiento posterior a la señal eficaces. Como limitaciones y oportunidades de mejora de esta propuesta se señala la falta de libertad completa de distribución y la necesidad de monitorear conjuntamente ubicación, escala y forma del proceso.

Tang, Mukherjeey Wang (2023) presentan novedosos esquemas MEWMA para la monitorización estadística de procesos multivariantes y de alta dimensión (HD). Incorporan métricas de distancia euclidiana de origen e inter-punto para abordar la complejidad inherente al monitoreo HD. Los esquemas propuestos son condicionalmente libres de distribución y particularmente robustos en escenarios de baja muestra de referencia en la Fase I. La muestra de referencia se particiona y el gráfico está diseñado condicionalmente en un pequeño grupo de muestras de sub-referencia seleccionadas al azar.

Estos investigadores aseguran que, en la Fase II del esquema basado en distancia inter-punto (IP), se calculan las distancias medias de cada muestra de prueba a puntos de sub-

referencia seleccionados al azar en la Fase I. Estas distancias medias sirven como valores de distancia de prueba para monitoreo en tiempo real. El diseño, condicional a muestras de sub-referencia, asegura propiedades libres de distribución, lo que simplifica la implementación y aumenta la robustez estadística. El estudio adopta la longitud media de ejecución (MRL) como métrica de rendimiento. A través de simulaciones y estudios comparativos, los esquemas MEWMA propuestos muestran superioridad en la detección de cambios en escala y ubicación sobre métodos existentes.

1.2.4. Gráficos de control multivariante para variables cualitativas

En el contexto de los gráficos de control multivariante para variables cualitativas, algunas propuestas hacen aportaciones al gráfico T^2 de Hotelling y otras se desarrollan con enfoques diferentes. La literatura científica ha cubierto ampliamente los gráficos de control en entornos multivariantes para datos cuantitativos y mixtos, pero, son pocas las aportaciones en el desarrollo de gráficos de control multivariante para variables cualitativas.

Hay procesos en las que las características de calidad del producto son de tipo atributo y siguen alguna distribución discreta, como la binomial multivariante o la distribución de Poisson multivariante. En el caso univariante, los gráficos de control *p* o *np* para datos distribuidos binomialmente y los gráficos *c* o *u* para datos distribuidos Poisson se utilizan para monitorear las características de calidad de tipo atributo (Montgomery, 2009). Cuando el enfoque está en el número de defectos en cada artículo y estos defectos pueden clasificarse en varias categorías, el proceso podría modelarse mediante una distribución de Poisson multivariante (Raza y Aslam, 2019).

Para facilitar la caracterización de los gráficos de control estadístico multivariantes para variables cualitativas, se presenta una clasificación de las propuestas centrada en el análisis de variables que siguen distribuciones Binomial, distribución Poisson y el análisis de variables multinomiales.

1.2.4.1. Gráficos de control para variables cualitativas con distribución Binomial.- Patel (1973) estudió casos en que la inspección de varios componentes de un ensamblaje se realiza por atributos. Si se inspeccionan p atributos en de los n ensamblajes fabricados en puntos sucesivos en el tiempo, se obtiene un patrón de ceros y unos, consistente en n filas y p columnas. Si Y_i denota el i-ésimo atributo (i = 1, 2, ... , p), dado que todos estos atributos son inspeccionados en el mismo ensamblaje, el vector $\boldsymbol{Y}(p\ x\ 1)$ tiene una distribución binomial puntual multivariante y los vectores fila sucesivos así obtenidos pueden depender del tiempo.

Patel propuso el estadístico $G = (\boldsymbol{X} - \bar{x})'\boldsymbol{S}^{-1}(\boldsymbol{X} - \bar{x})$ con distribución aproximada χ^2 y p grados de libertad, donde, $\boldsymbol{X}$ es un vector de observación; $\bar{x}$ es la media de la muestra obtenida del patrón base, tomada como estándar deseable del proceso bajo control; y, $\boldsymbol{S}$**,** la matriz de covarianza de la muestra del patrón base. Para determinar si la calidad de un conjunto futuro de vectores de observación, $X_1, X_2, .., X_k$, cumple con los estándares deseables, propuso el gráfico χ^2, con un solo UCL**,** similar al gráfico T^2 de Hotelling.

En la evolución de los gráficos de control para variables cualitativas, Lu (1998) se aleja de los gráficos tipo T^2 de Hotelling, como el χ^2 de Patel, porque considera que sólo tienen un límite de control superior, lo que les impide ser usados para la detección tanto del deterioro como de la mejora del proceso. En su lugar, presentó un gráfico de control MNP,

tipo Shewhart, para procesos de atributos multivariantes cuyas características de calidad se clasifican como conforme o no conforme para una unidad de producto, un enfoque que maneja características de calidad binomiales.

El gráfico de control MNP de Lu utiliza el estadístico X, que es la suma ponderada de los recuentos de unidades no conformes respecto de todas las características de calidad como estadístico de análisis, y utiliza límites de control *UCL* y *LCL*, tomando como referencia un valor central (*CL*). Además, propone un enfoque fácil y eficiente para identificar el contribuyente crítico a una señal fuera de control, así como la interpretación de la señal.

1.2.4.2. Gráficos de control para variables cualitativas con distribución Poisson.- La primera propuesta en este campo fue presentada por Holgate (1964). Su modelo admite que hay un factor común para todas las variables, más un factor único para cada una de las características de la calidad observadas. Este modelo está dado por $X_1 = Y_0 + Y_1$, donde, Y_0 representa a una variable común y explica la covarianza de todas las variables; mientras que, Y_1 es específica de cada variable y explica la variabilidad única de dicha variable.

Holgate desarrolló un trabajo sobre la distribución Poisson bivariante para variables correlacionadas, que fue tomado como referente para otras investigaciones de autores como Chiu y Kuo (2008); Lee y Branco (2009); Laungrungrong, Borror, y Montgomery (2011); Epprecht, Aparisi y García-Bustos (2013).

Luego, Patel (1973) desarrolló un método de control de calidad, análogo al tipo Hotelling, para poblaciones binomiales multivariantes o con distribución Poisson multivariante. Además, proporcionó técnicas para abordar situaciones problémicas, como matrices de covarianza singulares o dependencia temporal entre las observaciones, autocorrelación.

Otros autores trataron métodos que pretendían establecer límites óptimos para conteos de Poisson multivariantes. Entre ellos, Jones, Woodall y Conerly (1999), quienes propusieron un sistema de clasificación por deméritos para productos con múltiples defectos, desarrollando el estadístico D, basado en una combinación lineal de conteos de defectos y estableciendo límites de control (UCL y LCL) a partir de combinaciones lineales de variables aleatorias de Poisson. Por otra parte, Jiang et al. (2002) diseñaron un gráfico c simétrico, estableciendo límites de control que minimizaban la desviación de la longitud de corrida promedio en control (ARL_0) y abordaron las limitaciones de las aproximaciones normales, especialmente con bajos valores de λ en Poisson. Sin embargo, estos métodos requieren gráficos individuales para cada característica de calidad, lo que resulta ineficiente en el contexto de procesos con múltiples características.

Chiu y Kuo (2008) señalaron como desventajas del gráfico de control MNP de Lu, la suposición de normalidad y la falta de discusión sobre el rendimiento. Chiu y Kuo se enfocaron en llenar los vacíos de conocimiento relacionados con análisis de datos de conteos correlacionados múltiples. Señalaron que cuando el foco está en el número de defectos en cada unidad y el número de defectos se clasifica en más de dos categorías, los datos podrían modelarse utilizando la distribución de Poisson multivariantes. Chiu y Kuo desarrollaron un

gráfico de control para datos correlacionados del recuento de Poisson multivariante, conocido como gráfico MP. Los límites de control superior (*UCL*) e inferior (*LCL*) del gráfico se establecen mediante un método de probabilidad exacta. El modelo multivariante de Poisson adoptado en el gráfico MP, extensión de un modelo bivariante de Holgate (1964), es relativamente simple y puede monitorear muchas características de calidad en un solo gráfico de control.

Lee y Branco (2009) propusieron tres gráficos de control único para monitorear observaciones individuales en un proceso de Poisson bivariante. Se evaluó el desempeño de estos gráficos en comparación con dos gráficos de control univariantes, uno para cada característica de calidad, y se determinaron los límites de control y los riesgos de falsa alarma. Estos gráficos señalan un riesgo de falsa alarma con un nivel de significancia α menor o igual a 0.0027. Los resultados indican que los gráficos de control único presentan un mejor desempeño en la mayoría de los casos. La principal ventaja de estos gráficos es la sencillez de juzgar si el proceso está bajo o fuera de control por un gráfico único en lugar de dos gráficos de control separados.

Los gráficos dependen, según Lee y Branco, de las variaciones en la distribución Poisson bivariante, por ejemplo, la distribución de $DF = X_1 - X_2$ (Diferencia) es la base del gráfico DF; la distribución de $SM = X_1 + X_2$ (Suma) es la base del gráfico SM; y, la distribución de $MX = max(X1, X2)$ (Máximo), corresponde al gráfico MX; y dos gráficos separados, uno para X_1 y otro para X_2, denominado gráfico 2C. Las pautas para aplicar los gráficos de control propuestos dependen de la correlación entre los tipos de defectos observados. Si la correlación es alta, el gráfico SM es la mejor opción; si se mantiene estable

en ambos tipos de defectos, los gráficos MX o 2C son mejores; pero, si solo un tipo de defectos aumenta, el gráfico DF es la mejor opción.

Laungrungrong, Borror, y Montgomery (2011) aseguran que la monitorización simultánea de distintas características de calidad correlacionadas que siguen una distribución Poisson multivariante ha sido tradicionalmente implementado utilizando una aproximación normal a la distribución de Poisson para determinar los límites de control apropiados, mas, en su trabajo destacan la importancia de la distribución de Poisson en el monitoreo de datos que siguen esta distribución frente a la suposición de una distribución normal.

Estos investigadores proponen un esquema MPEWMA para tasas de conteo utilizando directamente el marco de distribución multivariante de Poisson. Afirman que el gráfico EWMA multivariante basado directamente en esta distribución es superior a uno basado en la distribución normal, en condiciones de ARL bajo control. Además, discuten la aplicación de su gráfico para la monitorización simultánea de varias características de calidad correlacionadas.

Epprecht, Aparisi y García-Bustos (2013) presentaron un nuevo gráfico de control multivariante llamado Gráfico de Combinación Lineal de Recuentos de Poisson (LCP) para la monitorizaciónde variables de Poisson correlacionadas. Este gráfico optimiza el ARL en control mediante una combinación lineal de p variables de Poisson, solventando carencias previas en métodos de control para dichas variables. La propuesta se presentó mediante un software especializado que facilita la calibración, determinando tanto los coeficientes óptimos para la combinación lineal como los límites de control que minimizan el ARL fuera de control, todo ello condicionado por un ARL en control predefinido (ARL_0).

Un rasgo distintivo de la propuesta de Epprecht, Aparisi y García-Bustos es la adaptabilidad a cualquier ARL_0 especificado, permitiendo límites de control en números reales. Este enfoqtenías dos palabras juntasue ofrece una herramienta robusta y flexible para la monitorización multivariante y tiene ventajas, como no estar restringido a valores enteros y acercarse más a la tasa de falsas alarmas deseada, lo que permite una detección más rápida del cambio en el proceso.

Raza y Aslam (2019) presentan dos nuevos gráficos de control para monitorización de datos de conteo de Poisson multivariante, que responden a dos esquemas: MDS y GMDS. Estos gráficos utilizan información de muestras anteriores para mejorar la sensibilidad en la detección de desplazamientos del proceso y aplican doble límite superior (UCL_1 y UCL_2) y doble límite inferior (LCL_1 y LCL_2), mejorando así la detección de anomalías. Se emplea el ARL para evaluar el rendimiento.

El estadístico $D = \sum_{j=1}^{p} Y_j$ se utiliza en esta propuesta, donde, $Y_j, (j = 1, 2, \dots, p)$ es el número de no conformidades o defectos de las características de la calidad que siguen una distribución de Poisson p-variante con media $\lambda_j, (j = 1, 2, \dots, p)$ y covarianza entre dos variables definidas como $(Y_1, Y_j) = \rho_{ij}$, para $i \neq j$. Se asume que $\rho_{ij} = \rho_0$.

En el esquema de muestreo de estado múltiple dependiente (MDS), el proceso se declara bajo control si el estadístico D cae dentro de $LCL_2 \leq D \leq UCL_2$; de lo contrario, se lo reporta fuera de control. El valor de i puede ser decidido por el experto en control de calidad. En el esquema GMDS, una versión generalizada de MDS, se declara que el proceso está bajo control si al menos k de los m estadísticos D anteriores se encuentran dentro de $LCL_2 \leq D \leq UCL_2$; de lo contrario, se activa una señal de fuera de control. Los valores de

k y m los especifican los expertos en calidad de acuerdo con la sensibilidad del proceso que se está monitoreando.

Según Raza y Aslam, los resultados muestran que los nuevos gráficos superan al esquema MDS tradicional (Aslam, Azam y Jun, 2014) y a otros gráficos existentes en eficacia y rapidez para detectar cambios en el proceso y sugieren que la generalización a otros tipos de gráficos de control es posible, así como la ponderación diferenciada de los tipos de defectos.

1.2.4.3. Gráficos de control para variables cualitativas con datos multinomiales.- Mukhopadhyay (2008), desarrolló un gráfico de control, D^2, con la distribución subyacente como normal multivariante. Retomó el análisis de la distancia de Mahalanobis D^2, considerando que esta distancia es la base del estadístico T^2 de Hotelling. En su análisis, este autor comparó las filas de la matriz de datos, es decir, el vector de proporción de defectuosos, correspondiente a un punto en el tiempo particular ($\boldsymbol{p_i}$) con el vector de proporción de defectuosos promedio ($\overline{\boldsymbol{p}}$). A partir de una generalización de la Distancia de Mahalanobis, definida por

$$D_i^2 = (\boldsymbol{p_i} - \overline{\boldsymbol{p}})' \boldsymbol{\Sigma_i^-} (\boldsymbol{p_i} - \overline{\boldsymbol{p}}) \tag{1.17}$$

donde, $\boldsymbol{\Sigma_i}$ es la matriz de varianza-covarianza del vector pi, planteó la fórmula en términos de

$$D^2_{\boldsymbol{p}_i,\overline{\boldsymbol{p}}} = N_i \sum_{i=1}^{k} \frac{\left(p_{ij} - \bar{p}_j\right)^2}{\bar{p}_j} \tag{1.18}$$

donde, $\boldsymbol{p}_i' = [p_{i1}, p_{i2}, \ldots, p_{iK}]$; $\overline{\boldsymbol{p}}_j = [\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_K]$ y N_i es el tamaño de muestra en el i-ésimo punto en el tiempo, para una clasificación de individuos en K categorías. El gráfico de control D^2 tiene un *UCL*, mientras que su *LCL* = 0.

Mukhopadhyay presentó un análisis que se centra en la adecuación de usar una distribución chi-cuadrado para el caso multinomial de la distancia. Usar una distribución chi-cuadrado parece adecuado, pero, tiene restricciones, como la de garantizar que la frecuencia esperada en cada categoría sea al menos cinco, especialmente cuando el tamaño de la muestra es pequeño o hay pocos defectos. Trabajar con múltiples categorías añade complejidad debido a las variaciones en *K*. Para Mukhopadhyay, es preferible, en lugar de χ^2, utilizar la distribución T^2, que considera el tamaño de muestra y se alinea más con el gráfico p tradicional.

Otra propuesta interesante es la aplicación de gráficos de control multivariantes a los procesos de atributos multinomiales, que depende del método de muestreo utilizado. Dos escenarios son posibles:

En el escenario 1, cuando los artículos se clasifican sucesivamente y por separado con respecto a todos los controles de calidad, Taleb, Limam y Hirota (2006), proponen dos enfoques, el enfoque difuso y el enfoque de probabilidad, para desarrollar estadísticos de seguimiento y gráficos de control.

Enfoque difuso utiliza escalas intermedias para las características de calidad y las representa mediante datos lingüísticos. Cada conjunto de términos se asocia a subconjuntos

difusos y se transforma en valores escalares, conocidos como valores representativos. El gráfico de control de calidad difuso multivariante (MFQCC) utiliza un estadístico de prueba

$$T^2 = (R - \mu)'\Sigma^{-1}(R - \mu) \tag{1.19}$$

Aquí R indica el valor representativo del número difuso en una muestra, mientras que μ y Σ refieren al vector de medias y la matriz de covarianza de las características de calidad, derivados de muestras preliminares. Una vez que el proceso está bajo control, se utilizan S (matriz de covarianza muestral) y R (vector de media muestral) como estimaciones de Σ y μ.

Al recolectar datos de una fase estable, se determinan límites de control para futuras medias de muestras. Se usa la matriz de covarianza S y el vector R para estimar Σ y μ. Al hacer esto, el estadístico se modifica a

$$T_f^2 = (R - \bar{R})'\mathrm{S}^{-1}(R - \bar{R}) \tag{1.20}$$

El gráfico de control MFQCC es similar al gráfico de control T^2 tradicional. Sin embargo, hay diferencias. La distribución del estadístico de Hotelling T^2 es conocida y sus límites de control pueden determinarse fácilmente, mientras que, para establecer límites de control en el gráfico MFQCC se debe determinar la distribución de T_f^2. Sin embargo, la suposición de normalidad no se cumple para la distribución del valor representativo y, entonces, la determinación de la distribución de T_f^2 se vuelve un desafío.

En el enfoque de probabilidad, Taleb, Limam y Hirota utilizan el estadístico W_i^2, que es una combinación lineal de las p características de calidad correlacionadas y es difícil determinar directamente su distribución. Sin embargo, se puede aproximar mediante χ^2 con

v_i grados de libertad. El valor de v_i puede estimarse mediante la aproximación de Satterthwaite. El gráfico de control de calidad de atributos multivariantes (MAQCC) tiene un límite de control superior que se determina utilizando los percentiles de la distribución chi-cuadrado con *v* grados de libertad. Está dado por

$$W_i^2 = \sum_{j=1}^{p} Z_{ij}^2 \tag{1.21}$$

donde, *Z* es un estadístico que se utiliza en la monitorizaciónde cada periodo respecto de las características de calidad.

En el escenario 2 cada elemento se controla simultáneamente con respecto a todos los controles de calidad. Según Taleb (2009), en este caso los elementos de una muestra dada se clasifican en la tabla de contingencia $\boldsymbol{q}_1$ x $\boldsymbol{q}_2$ x ... x $\mathrm{q_p}$, donde $\boldsymbol{q_1}, \ldots, \boldsymbol{q}_\mathrm{p}$ son los números de categorías de las características de la calidad, $1, \ldots, p$. Este autor plantea dos enfoques para la elaboración de gráficos de control que monitoricen procesos de atributos multivariantes con datos lingüísticos multidimensionales, uno basado en la teoría difusa y otro, en la teoría de la probabilidad.

El Enfoque difuso aborda la representación de perfiles de calidad en procesos de atributos mediante el uso de lógica difusa. Se plantean dos casos: el intramétodo difuso (FI) y el intermétodo difuso (FN). En el Caso I (FI), cada perfil de calidad está compuesto por un vector de términos lingüísticos como “bueno” o “malo”, que puede ser modelado por un subconjunto difuso con su correspondiente función de membresía. La cantidad de perfiles depende del número de subconjuntos para cada variable lingüística y los subconjuntos

difusos derivan de las variables lingüísticas originales. En el Caso II (FN), un perfil de calidad está representado por un único término, una nueva variable lingüística que se construye mediante la combinación de múltiples términos difusos, también representados por subconjuntos difusos y funciones de membresía.

De este enfoque se desprenden dos Gráficos de Control Difuso Multivariante: método FN (los gráficos del enfoque FN superan a los basados en el enfoque FI). En el gráfico para el método Probabilístico Difuso (FNP) cada subconjunto difuso Fl se convierte en su valor representativo rv_l. El valor atribuido a una muestra dada i de tamaño n es la media de rv_l dada por

$$\overline{rv_l} = \frac{1}{n}\sum_l Y_l \, rv_l \tag{1.22}$$

Mientras que, en el método de Membresía Difusa (FNM), FSM es la media difusa de la muestra *i* y μFSM su función de membresía, dada por

$$FSM = \frac{1}{n}\sum_l Y_l \, F_l \tag{1.23}$$

Los dos gráficos de este enfoque tienen *UCL* y *LCL*.

En el Enfoque de probabilidad, Taleb asegura que, para la monitorizaciónde perfiles de calidad de producto, se consideran dos casos: cuando las probabilidades π_l son conocidas a priori (Caso I) y cuando se estiman a partir de datos (Caso II). En el Caso I, se utiliza la estadística de bondad de ajuste de Pearson, cuya distribución asintótica es una χ^2 con v grados de libertad. En el Caso II, se aplica una prueba de homogeneidad de proporciones

entre un período base y los subsiguientes períodos de monitoreo. Ambos casos son asintóticos y sujetos a restricciones sobre el tamaño de la muestra para garantizar que la distribución χ^2 sea aplicable.

De este enfoque se deriva el Gráfico de Control de Probabilidad Multivariable (MPCC), que utiliza un estadístico dado por

$$\chi_i^2 = n_i n_o \sum_{l=1}^{r} \frac{(p_{il} - p_{0l})^2}{Y_{il} - Y_{0l}} \tag{1.24}$$

donde, p_{il} y p_{ol} son las frecuencias esperadas correspondientes en las celdas de la tabla de contingencia en el período ***i*** y ***0;*** Y_{il} e Y_{0l} denotan el número de observaciones en ***el*** perfil ***l*** para el período ***i*** y 0, respectivamente; y, n_i y n_o son el tamaño de la muestra de los períodos ***i*** y 0. La distribución de χ_i^2 se aproxima a $\chi^2(v)$. El MPCC utiliza un UCL que es un percentil de la distribución χ^2.

Kumar y Mohapatra (2012) avanzan en el campo de los gráficos de control de calidad multivariantes, poniendo énfasis en la evaluación subjetiva de las características del producto por parte de un panel de expertos. Su método utiliza la teoría de conjuntos difusos para abordar la incertidumbre y vaguedad inherentes en evaluaciones cualitativas. En su metodología, los expertos asignan rankings a diversas características de calidad, las cuales pueden estar correlacionadas. Estos rankings se transforman en números difusos que luego se ponderan y suman interactivamente para obtener una métrica compuesta de la calidad del producto. Posteriormente, se diseñan gráficos de control utilizando medidas de posibilidad y

necesidad, entre 0 y 1, que van de imposible a posible e innecesario a necesario, respectivamente.

De este modo, Kumar y Mohapatra ofrecen un esquema robusto y flexible para monitorear productos con múltiples atributos correlacionados en escenarios donde la evaluación objetiva puede ser difícil o imposible de lograr y, contribuyendo a la toma de decisiones en entornos complejos y ambiguos, donde la subjetividad y la incertidumbre son factores predominantes.

Pastuizaca-Fernández, Carrión y Ruiz-Barzola (2015) propusieron un gráfico de control de procesos difuso para características de calidad de tipo multi-atributos correlacionados. Emplearon un enfoque que integra el gráfico de control multivariante T^2 de Hotelling con la teoría de conjuntos difusos. En este marco, las variables lingüísticas capturan la incertidumbre y la subjetividad en la evaluación de la calidad y se modelan como Números Triangulares Difusos (TFN) que cumplen con propiedades de normalidad y convexidad, haciendo que la contribución de cada variable individual fuera descompuesta para un análisis más preciso. Los términos lingüísticos se asumen como funciones de pertenencia normales y convexas en el intervalo [0,1], lo que facilita su integración en el gráfico T^2 de Hotelling.

La expresión mediante la cual la variable multinomial correspondiente a cada característica de calidad Q_j de una muestra *i* se convierte en un valor numérico está dada por

$$R_{ij} = \frac{1}{n}\sum_{k=1}^{s_j} n_{ij} vr_{jk} \tag{1.25}$$

donde, vr_{jk} es el valor representativo correspondiente a la función de pertenencia de la categoría k de la característica de la calidad j y S_j es una combinación lineal de los s_j valores n_{ijk} correspondientes al número de productos clasificados con las categorías q_{jk} de la característica de calidad Q_j en la muestra A_i. La muestra A_i de n observaciones está representada por el vector $R_i = \left(R_{i1}, \dots, R_{ij}, \dots, R_{ip}\right)'$, donde R_{ij} cumple una distribución normal con media μ_{ij}.

El conjunto de valores representativos de las p características de calidad está dado por el vector $\boldsymbol{R_i}$, el estadístico de prueba a ser representado en el gráfico de control para cada muestra es

$$T_i^2 = (\boldsymbol{R}_i - \boldsymbol{\mu}_R)'\boldsymbol{\Sigma}_R^{-1}(\boldsymbol{R}_i - \boldsymbol{\mu}_R) \tag{1.26}$$

donde $\boldsymbol{\mu}_R = \left(\mu_1, \dots \mu_p\right)$ es el vector de medias para cada característica de la calidad y $\boldsymbol{\Sigma}_R$ es la matriz de covarianzas de las características de calidad. Por otra parte, los parámetros $\boldsymbol{\mu_R}$ y $\boldsymbol{\Sigma_R}$ serán estimados por $\bar{\boldsymbol{R}}$ y $\boldsymbol{S}$, respectivamente. Luego, el valor del estadístico T^2 para $\boldsymbol{R}_i$, que puede representarse en un gráfico T^2 de Hotelling en Fase I, responde a

$$T_i^2 = (\boldsymbol{R}_i - \bar{\boldsymbol{R}})'\mathbf{S}^{-1}(\boldsymbol{R}_i - \bar{\boldsymbol{R}});\ i = 1, 2, \dots, m, \tag{1.27}$$

En Fase II, sea un conjunto de datos históricos (HDS) de m observaciones y una sola futura observación Υ, el valor del estadístico

$$T_i^2 = (\Upsilon - \bar{R})'\mathrm{S}^{-1}(\Upsilon - \bar{R}) \tag{1.28}$$

se puede representar en un gráfico T^2 de Hotelling. Los parámetros $\boldsymbol{\mu_R}$ y $\boldsymbol{\Sigma_R}$ deberán ser estimados por $\boldsymbol{\bar{R}}$ y $\boldsymbol{S}$ a partir del análisis del HDS formado por muestras de tamaño n, tomadas cuando el proceso está bajo control. El vector de medias $\boldsymbol{\mu_R}$ será estimado por:

$$\bar{R} = \left(\bar{R}_1, \dots, \bar{R}_j, \dots, \bar{R}_p\right). \tag{1.29}$$

Para la estimación de $\boldsymbol{\Sigma_R}$, Pastuizaca-Fernández, Carrión y Ruiz-Barzola, siguiendo a Sullivan y Woodall (1999), consideraron que un gráfico más eficaz estima la matriz de covarianzas de las diferencias vectoriales entre observaciones sucesivas, asumiendo que las observaciones sucesivas tendrán similar vector de medias, que es una estimación más robusta para observaciones individuales. Este estimador se denomina Diferencia Sucesiva Cuadrática Media (MSSD).

1.2.5. Gráficos para el control estadístico de procesos con técnicas estadísticas multivariantes

La complicada estructura de las variables asociada con características altamente correlacionadas ha promovido el uso, cada vez mayor, de sustitutos de los métodos convencionales. En este sentido, se han desarrollado métodos para abordar el problema de la alta correlación entre características, transformándolas en un conjunto de variables no correlacionadas (Farokhnia y Niaki, 2020).

El procedimiento estadístico para la ejecución de estos gráficos se basó, al inicio, en la aplicación de la distancia de Mahalanobis; no obstante, poco a poco se fueron incorporando otras técnicas estadísticas multivariantes tales como Análisis de Componentes Principales (Pearson, 1901), Análisis de conglomerados (Edwards y Cavalli-Sforza, 1965), Métodos

Biplot (Gabriel, 1971; Galindo-Villardón, 1986), Análisis de correspondencias (Benzécri, 1973), STATIS (L'Hermier des Plantes, 1976; Robert y Escoufier, 1976; Lavit, 1988), Coordenadas paralelas (Inselberg y Dimsdale, 1990), PCA Mix (Ahsan, et al., 2018).

1.2.5.1. Control estadístico de procesos multivariantes con STATIS.- Filho y Luna (2015) presentaron un gráfico de control multivariante que utiliza STATIS para la monitorizaciónde procesos por lotes en entornos no paramétricos. El enfoque de monitoreo del comportamiento del proceso centrado en el tiempo es una de las principales contribuciones del método de análisis de datos propuesto. El gráfico propuesto considera la correlación cruzada y la estructura de correlación de los datos, preservando la información sobre la variabilidad del proceso a lo largo del eje de tiempo y ofrece una detección precisa de estados de proceso fuera de control que permite la adopción oportuna de acciones correctivas.

El método STATIS, es una técnica de análisis de datos desarrollada por L'Hermier des Plantes (1976), Escoufier (1987) y Lavit, Escoufier, Sabatier y Traissac (1994), que maneja datos de tres vías como un conjunto de K matrices y calcula la distancia euclidiana entre configuraciones de las mismas observaciones obtenidas en T momentos diferentes (Lavit, Escoufier, Sabatier y Traissac, 1994). Este método analiza la estructura de varias tablas de datos que son conjuntos de medidas sobre los mismos individuos, aunque no sean las mismas variables (Ramos-Barberán, et al., 2018). El interés del método STATIS reside en la comparación de los individuos, privilegiando sus posiciones relativas. Por el contrario, si las T matrices representan las mismas variables, pero los individuos pueden variar en T ocasiones distintas, se aplica el método STATIS Dual, cuyo interés se centra en la relación entre variables (González, 2015).

1.2.5.2. Control estadístico de procesos multivariantes con STATIS Dual y Coordenadas paralelas.-

Ramos-Barberán, et al. (2018) utilizaron STATIS Dual y Coordenadas paralelas en el control de procesos multivariantes relacionados con la producción por lotes en entornos no paramétricos, para la monitorización fuera de línea de lotes y variables, así como de la agrupación visual de observaciones dentro de los diagramas de bolsa (*bagplots*) robustos lotes. Estos procesos utilizan *bagplots* robustos para la configuración de regiones de control que favorecen la representación de la dinámica entre las variables.

La propuesta de estos autores representa tablas como puntos en mapas de PCA bidimensionales del espacio entre Estructuras, así como el espacio Infraestructura o mapas de Compromiso. Estos gráficos, denominados Biplot por su autor, Gabriel (1971), representan variables e individuos en un espacio de dos dimensiones y se muestran en dos versiones: el GH-Biplot que prioriza la representación de variables (columnas) y el JK-Biplot, que representa mejor los individuos (filas). Galindo-Villardón (1986) aportó, con su HJ-Biplot, a mejora de la calidad de representación de individuos y variables de forma simultánea sobre un mismo plano de referencia.

La metodología de Coordenadas paralelas fue desarrollada por Inselberg y Dimsdale (1990), permite visualizar relaciones multivariantes y problemas multidimensionales, induce un mapeo no proyectivo entre conjuntos N-Dimensional y 2-Dimensional, como un sistema que representa datos multivariantes en un sistema bidimensional. Las coordenadas paralelas preservan la naturaleza multilineal del conjunto de datos y no precisan la creación de variables latentes para reducir la dimensionalidad. El análisis de cada variable, así como al

interior de los lotes, se muestra en un sistema todo en uno, de esta manera se reconocen patrones locales y su comportamiento dinámico global (Ramos-Barberán, et al., 2021).

1.2.5.3. Control estadístico de procesos multivariantes con algoritmos de clasificación.- Según Liu, Liu y Jung (2020), el procesamiento de datos complejos de alta dimensión, que se encuentran frecuentemente en los sistemas modernos, ha merecido que muchos investigadores dediquen esfuerzos para el desarrollo de gráficas de control multivariantes que, utilizando métodos no paramétricos, aborden las limitaciones relacionadas con el supuesto de distribución. En este sentido, se ha trabajado con algoritmos de clasificación (OCC) que identifican si los objetos pertenecen o no a una clase específica a partir del análisis de sus características más relevantes; también, el uso de puntuaciones de novedad de los algoritmos OCC como estadísticas de seguimiento de los gráficos de control, un tipo de algoritmos de minería de datos que asume que solo los datos bajo control se usan para determinar el grado de anormalidad de una nueva observación no clasificada.

Además, se han desarrollado gráficos de control basados en puntuaciones de novedad que utilizan el algoritmo k-vecino más cercano (kNN) para el análisis de la densidad de los datos bajo control. Estos autores desarrollaron un gráfico de control (DNW) que utiliza la puntuación de novedad sensible a la densidad como estadístico de seguimiento utilizando el algoritmo k-vecino más cercano (kNN).

1.2.5.4. Control estadístico de procesos multivariantes con Análisis de Componentes Principales.- Según Farokhnia y Niaki (2020), los gráficos de control basados en Análisis de Componentes Principales (PCA) se utilizan ampliamente para superar el problema de la correlación entre las variables medidas mediante la definición de

transformaciones lineales de las variables existentes en un nuevo espacio no correlacionado. Las variables transformadas explican diferentes cantidades de varianza, de esta forma, en el primer componente principal explican la cantidad más alta, en el segundo PC explican la segunda más alta, y así sucesivamente. El PCA también reduce la dimensión al perder una parte de la información de la variación observada en las variables originales.

Sin embargo, el supuesto subyacente de que las observaciones se distribuyen normalmente ha limitado la aplicabilidad de los esquemas basados en PCA, ya que el supuesto de normalidad no siempre se cumple en las prácticas reales. Luego, un método de distribución libre para establecer los límites de los gráficos de control basados en PCA puede ser una propuesta confiable cuando no se cumple el supuesto de normalidad.

El método presentado por Farokhnia y Niaki se basa en máquinas de vectores de soporte (SVM) como sustituto de los métodos convencionales para construir límites de control en gráficos de control basados en PCA. Como SVM utiliza observaciones del proceso en el mundo real, no se requiere ningún supuesto de distribución para construir límites de control.

1.2.5.5. Control estadístico de procesos multivariantes con PCA Mix.- El PCA Mix, Gráfico de control multivariante basado en la combinación de PCA para características de calidad de atributos y variables, fue introducido por Ahsan, et al. (2018). El método está basado en el gráfico de control T^2 que puede manejar conjuntamente datos continuos y categóricos, por medio de una combinación de Análisis de Componentes Principales (PCA) y Análisis de Correspondencias Múltiples (MCA).

Para estimar el límite de control utilizaron el método *Kernel Density Estimation* (KDE). El desempeño del gráfico propuesto se evalúa utilizando ARL. Los límites de control T^2 obtenidos del método KDE producen un ARL_0 estable alrededor de 370 para $\alpha = 0.00273$, mientras que, el ARL_0 del gráfico de control T^2 que utiliza un límite de control convencional no es estable. Por otra parte, el ARL_1 del gráfico propuesto se reduce rápidamente a medida que aumenta el cambio tanto de las características variables como de los atributos. Considerando el desplazamiento del proceso, el gráfico propuesto tiene un mejor rendimiento cuando utiliza una cantidad adecuada de componentes principales.

Ahsan, et al. (2022) explican que la presencia de valores atípicos puede llevar a una detección errónea en las observaciones fuera de control en la Fase II, por lo tanto, deben limpiarse en la Fase I. Proponen un gráfico T^2 basado en PCA Mix con límite de control de densidad de Kernel para mezclas de datos categóricos y continuos. Evaluaron el rendimiento del gráfico propuesto en la detección de valores atípicos a partir de datos limpios y contaminados.

Se encontró que el gráfico propuesto tiene un mejor rendimiento que el punto de referencia en el seguimiento de datos limpios, mientras que, para datos contaminados, tiene un rendimiento óptimo en situaciones en las que los datos categóricos se generan a partir de una distribución multinomial con parámetros equilibrados. En comparación con los gráficos convencionales y otros gráficos robustos, el gráfico propuesto demostró un gran rendimiento por el éxito en la detección correcta de valores atípicos.

El PCA Mix es eficaz para monitorear simultáneamente variables numéricas y categóricas en un solo gráfico, no obstante, cuando éstas tienen una proporción

desequilibrada se producen algunas dificultades. Para superarlas, Ahsan, et al. (2022) desarrollaron el gráfico de control Kernel PCA Mix. En este gráfico, los datos categóricos se transforman en variables ficticias (*dummy*), después se combinan con los datos numéricos para dar lugar a la función Kernel. La descomposición de valores propios se realiza en el espacio de características mediante la realización de un mapeo no lineal y se calculan las puntuaciones de los componentes principales. Finalmente, el estadístico T^2 se estima a partir de los componentes principales calculados y para el cálculo del límite de control se utiliza la KDE.

Jin y Loosveldt (2022) abordan el tema desde un gráfico de control no paramétrico que trabaja con una combinación de variables numéricas y categóricas simultáneamente. Esta propuesta incluye el análisis de componentes principales Mix (PCA Mix), el gráfico T^2 de Hotelling y dos enfoques no paramétricos: estimación KDE y *bootstrap*, debido a la naturaleza desconocida de la distribución subyacente.

1.2.5.6. Control estadístico de procesos multivariantes con Análisis de Correspondencias Múltiples.- Rojas-Preciado, et al. (2023), desarrollaron el gráfico de control T2Qv, una herramienta de control estadístico de procesos multivariantes para el análisis de datos cualitativos, nominales y ordinales, basado en técnicas de Análisis de Correspondencias Múltiples (MCA) y el gráfico T^2 de Hotelling. El T2Qv es una técnica estadística multivariante de tres vías que trabaja con bases de datos cualitativos para n individuos, con p variables en K momentos distintos. Utiliza el estadístico T^2 de Hotelling ajustado, que está dado por

$$T_{med}^2 = n(\tilde{x}_k - \tilde{x}_0)' \sum_0^{-1} (\tilde{x}_k - \tilde{x}_0) \quad (1.30)$$

donde, n es el número de individuos de la tabla k; $\tilde{x}_k$ es el vector de medianas de la tabla k; $\tilde{x}_k$ es el vector de medianas de la tabla concatenada, que ha sido tomada como referente; y Σ, la matriz de covarianzas. Dado que este gráfico de control está basado en distancias de Mahalanobis ponderadas, su límite de control viene dado por $UCL = \chi^2_{\alpha,p}$, donde α es la significancia predeterminada ($\alpha = 0.0027$) y p, el número de dimensiones. Cuando el estadístico T_{med}^2 de alguna muestra (tabla) supera este límite significa que el proceso está fuera de control.

La interpretación de los puntos fuera de control se realiza comparando la ubicación de las categorías de las variables en el gráfico MCA de la tabla concatenada y el de la tabla fuera de control. Las categorías que están incidiendo en el estado fuera de control son las que muestran mayores diferencias en su ubicación al comparar ambos gráficos. Para cuantificar la magnitud de dichas diferencias o del comportamiento anómalo, se calcula las distancias Chi-cuadrado entre las masas de columnas de la tabla concatenada y las de la tabla fuera de control. La implementación del T2Qv se facilita con la aplicación de un paquete de R (T2Qv), que permite visualizaciones interactivas a través de una interfaz Shiny y está disponible en CRAN.

En resumen, se puede decir que la evolución de los gráficos estadísticos para el control de procesos multivariantes ha transcurrido por diferentes fases, cada una con diferentes aportes significativos y enfoques específicos, lo que ha permitido el desarrollo de métodos estadísticos y herramientas informáticas cada vez más sofisticadas y efectivas para

la monitorizaciónde procesos relacionados con la gestión de la calidad en diferentes ámbitos y sectores.

La gestión de la calidad y la monitorizaciónde procesos son cada vez más complejos debido al aumento en la cantidad y variedad de datos generados por las empresas. La combinación de gráficos de control estadístico de procesos con técnicas modernas de minería de datos, inteligencia artificial, *big data, randon forest, statistical learning* y otras que pudieran ser incorporadas a la realización de los gráficos de control, constituyen un desafío que debe ser abordado por los investigadores de los gráficos de control estadístico de procesos multivariantes, porque contribuyen a la identificación temprana de problemas y la toma de decisiones informadas en tiempo real. Por lo tanto, es importante que los esfuerzos se dirijan al desarrollo de herramientas y recursos que les permitan implementar estas técnicas y asegurar la calidad de sus procesos para mantenerse competitivas en el contexto actual.

1.3. Contribuciones del Capítulo 1

Los gráficos de control son una herramienta esencial en el control estadístico de procesos para reducir la variabilidad del proceso, establecer límites y parámetros óptimos y evaluar el comportamiento de las variables. Han evolucionado desde la monitorización de una única variable hasta el control estadístico de procesos multivariantes, lo que ha permitido el análisis simultáneo de múltiples características de calidad.

Se identifican cinco etapas en el desarrollo de gráficos de control para el control estadístico de procesos multivariantes. Las dos primeras surgen para responder a las necesidades de representación de datos que no podían ser abarcados por los gráficos

originales de Shewhart, mientras que las últimas tres recorren una trayectoria casi simultánea y dan lugar al desarrollo de propuestas bien diferenciadas para la monitorizaciónde procesos. Estas etapas son el Gráfico T^2 de Hotelling, Mejoras al rendimiento de este gráfico, Gráficos de control en entornos no paramétricos, Gráficos de control para variables cualitativas y Gráficos para el control de procesos con técnicas estadísticas multivariantes.

El T^2 de Hotelling es uno de los gráficos más importantes en el control estadístico de procesos multivariantes, se utiliza para el seguimiento del vector medio del proceso en el que se supone que la distribución de probabilidad conjunta de las p-características de calidad es la normal multivariante.

Como ventajas tiene que el gráfico T^2 de Hotelling es eficaz para detectar cambios grandes en la media del proceso, ofrece una mejor representación en comparación con el gráfico denominado elipse de control y es capaz de monitorear simultáneamente múltiples características del proceso correlacionadas. Sus desventajas se relacionan con la necesidad de un tamaño grande de muestras, la poca eficacia para detectar pequeños cambios, vulnerabilidad ante valores atípicos y su dependencia de distribuciones normales, requisito que en muchos casos no se cumple.

Se han propuesto gráficos con memoria como MCUSUM y MEWMA para mejorar la eficacia del gráfico de control T^2 de Hotelling y detectar pequeños cambios en la media del proceso. Además, se han aplicado técnicas como la modificación de los límites de control y la modelización de series temporales para eliminar la autocorrelación en el uso de los gráficos. Los métodos de muestreo como el gráfico de control T^2 adaptable, DDT^2 y VDT^2

mejoran la eficiencia del análisis multivariante del proceso, disminuyen el costo y mejoran la capacidad de detectar cambios.

Los gráficos de control multivariantes tradicionales no pueden manejar datos complejos de alta dimensión comunes en los sistemas modernos de producción, cuyos datos generalmente no siguen una distribución gaussiana. Como alternativa se han propuesto gráficos de control multivariantes no paramétricos más flexibles que manejan mezclas de datos continuos y categóricos y distribuciones no paramétricas, utilizando métodos de remuestreo como el Bootstrap, que pueden acomodar correctamente la correlación de datos en serie estacionaria, detectar valores atípicos y cambios en el vector de parámetros de ubicación de la distribución.

La literatura ha abordado con generosidad los gráficos de control multivariantes para datos numéricos y mixtos, pero no hay muchas aportaciones para variables cualitativas. Han surgido propuestas para datos de conteo de Poisson y datos lingüísticos, como el gráfico MP, LCP y gráficos de control difuso. Éstos, particularmente, son útiles para manejar situaciones donde los datos estadísticos son inciertos, incompletos o hay subjetividad humana. Son una buena opción para procesos con características de calidad multinomiales y correlacionadas.

La alta correlación entre características de la calidad ha impulsado el desarrollo de métodos no convencionales en la monitorizaciónde procesos. El PCA, el MCA, el método STATIS, los Biplots, las Coordenadas paralelas, Análisis de conglomerados, técnicas bootstrap y los gráficos de control basados en KDE se cuentan entre las técnicas más utilizadas. Además, se han propuesto gráficos de control multivariantes que abordan la distribución no normal de los datos y el desequilibrio en las proporciones categóricas. Las

técnicas de análisis estadístico multivariante han enriquecido el control estadístico de procesos.

La aplicación de métodos teóricos ha facilitado el análisis de la evolución de los gráficos estadísticos para el control de procesos multivariantes, mediante el establecimiento de fases con enfoques específicos y aportes significativos que han permitido el desarrollo de métodos estadísticos y herramientas informáticas cada vez más sofisticadas y efectivas para la gestión de la calidad en diferentes ámbitos y sectores.

La implementación de gráficos de control estadístico de procesos en combinación con técnicas modernas de minería de datos, inteligencia artificial, *big data, randon forest, statistical learning* y otras, constituye una oportunidad clave para la mejora de la calidad y la eficiencia en la gestión de procesos en diversos ámbitos. La capacidad de analizar grandes cantidades de datos de manera automatizada y en tiempo real, brinda la posibilidad de detectar patrones y anomalías con mayor precisión y rapidez, lo que favorece la toma de decisiones más informada y oportuna.

La combinación de estas herramientas estadísticas con la tecnología moderna puede ser un gran aliado en la mejora continua en pro de la excelencia en la gestión de la calidad. Esto deja abierta la posibilidad de inéditas aportaciones a los gráficos de control, aplicando nuevas metodologías estadísticas desarrolladas para el manejo de grandes volúmenes de datos.

Como contribución de este capítulo, se ha escrito un artículo científico titulado "*Multivariate statistical process control charts for qualitative variables: A review from genesis to development perspective*", que fue sometido el 05 de octubre de 2023, con número

de submission 230632842, a la revista *Quality Engineering*, de Taylor and Francis. Las métricas de esta revista para 2022 son: CiteScore Best Quartile: Q2; CiteScore Scopus: 3.5; Impact Factor: 2.0; Impact Factor Best Quartile: Q2; SJR: 0.604. Al momento de escribir este capítulo de la Tesis, el artículo está en fase de revisión por pares ciegos.

Capítulo 2

Gráfico de control de procesos estadísticos multivariantes para variables cualitativas

2.1. Introducción

El control estadístico juega un rol muy importante en la mejora continua de los procesos y dentro de éste, los gráficos de control, los cuales ayudan a monitorizar los procesos, han sido extensamente utilizados desde su creación por Walter Shewhart hasta nuestros días (Gutiérrez y De la Vara, 2013).

Shewhart estableció dos fases en el control de los procesos: la primera, denominada Fase de desarrollo, que describe el comportamiento estadístico de la variable analizada, determina los límites de control para el estimador del parámetro analizado y contribuye a la eliminación de causas asignables o especiales de variabilidad. La segunda, Fase de madurez, establece la capacidad del proceso para cumplir con los requisitos, identifica la media del número de muestras antes de obtener falsas alarmas y favorece la disminución del número de muestras para la detección de cambios en el proceso (Ruiz-Barzola, 2013). A esto, Montgomery (2012) le denomina el seguimiento de la producción futura.

A partir de los gráficos univariantes se han desarrollado diversas propuestas, las cuales incorporaron la opción de monitorizar varias variables a la vez (Ramos, 2017; Li, Tsung, y Zou 2012), abriendo con ello el control estadístico de procesos multivariantes. Las opciones más conocidas son: El gráfico de control T^2 de Hotelling (Hotelling, 1947), el cual se podría considerar la versión multivariante del gráfico de medias de Shewhart; el MEWMA (Lowry et al., 1992), versión multivariante del gráfico de medias ponderadas EWMA

(Roberts, 1959); o el MCUSUM (Crosier, 1988), el cual es la versión multivariante del grafico de control de sumas acumuladas CUSUM (Page, 1954).

A estos gráficos de control multivariante se le han realizado varias mejoras tales como: su optimización, determinando de forma analítica (Aparisi, 1996; Aparisi y Haro, 2001; Faraz y Parsian, 2006) o heurística los valores óptimos de sus parámetros (Ruiz-Barzola, 2013); otra propuesta es la de trabajar sin distribuciones probabilísticas o versiones no paramétricas (Shabbak y Midi, 2012; Liu, Liu y Jung 2020; Xue y Qiu, 2020), para procesos continuos o por lotes (Ramos, 2017).

Todos estos gráficos de control multivariante tienen un enfoque cuantitativo; es decir, las variables monitorizadas son esencialmente cuantitativas, ya sean discretas o continuas. Para ello, los diferentes autores, inicialmente se valieron de la distancia de Mahalanobis (Mahalanobis, 1936). Posteriormente, para el análisis de una combinación de variables continuas y categóricas se desarrolló el gráfico basado en la distancia de Gower (Tuerhong y Kim, 2014).

Sin embargo, el abordaje de problemas como la alta correlación entre características y en presencia de datos mixtos, requirió la incorporación de técnicas estadísticas multivariantes clásicas, como Análisis de Componentes Principales (Pearson, 1901), Métodos Biplot (Gabriel, 1971; Galindo-Villardón 1986), Análisis de Correspondencias (Benzécri, 1973), STATIS (L'Hermier des Plantes, 1976; Robert y Escoufier, 1976; Lavit, 1988), Coordenadas paralelas (Inselberg y Dimsdale, 1990), Análisis de conglomerados (Edwards y Cavalli-Sforza, 1965).

Dentro de las aportaciones referentes a los gráficos de control que incorporan técnicas multivariantes destacan el gráfico basado en STATIS para la monitorizaciónde procesos por lotes en entornos no paramétricos (Filho y Luna, 2015); los diagramas de bolsa robustos que utilizan STATIS Dual y Coordenadas paralelas (Ramos-Barberán, et al., 2018); el Gráfico de control multivariante PCA para datos mixtos, que aplica una combinación de Análisis de Componentes Principales y Análisis de Correspondencias Múltiples (Ahsan, et al., 2018); el gráfico de control de peso de novedad sensible a la densidad (DNW) utiliza el algoritmo k-vecino más cercano (kNN) (Liu, Liu y Jung, 2020); el gráfico basado en Kernel PCA Mix (Ahsan et al., 2022); el Gráfico T^2 basado en la combinación de PCA para datos continuos y cualitativos con detección de datos atípicos (Ahsan et al., 2021); los gráficos de control basados en PCA para entornos no paramétricos (Farokhnia y Niaki, 2020; Jin y Loosveldt, 2022).

No obstante, las aportaciones al desarrollo de gráficos de control multivariante para variables cualitativas no han sido numerosas. En este campo las propuestas se han desarrollado alrededor del análisis de variables que siguen una distribución Poisson y el análisis de variables multinomiales. La primera propuesta fue la de Holgate (1964), quien presentó un trabajo sobre la distribución Poisson bivariante para variables correlacionadas. Este modelo fue tomado como insumo en las investigaciones de autores como Chiu y Kuo (2008), Lee y Branco (2009), Laungrungrong y Montgomery (2011), Epprecht, Aparisi y García-Bustos (2013).

Otra propuesta destacada es la de Lu (1998), quien desarrolló un gráfico de control tipo Shewhart para procesos multivariantes con variables cualitativas, cuando la

característica de calidad asume valores binarios, al cual denominaron gráfico *np* multivariante (MNP).

Ya en el contexto multinomial, Mukhopadhyay (2008) planteó un gráfico de control multivariante utilizando el estadístico D^2 de Mahalanobis para atributos que siguen una distribución multinomial. Además, de una propuesta para procesos multinomiales bajo el enfoque difuso (Taleb, Limam y Hirota 2006), Taleb (2009) introdujo gráficos de control para la monitorizaciónde procesos multivariantes con datos lingüísticos multidimensionales, basados en dos procedimientos: la teoría de la probabilidad y la teoría difusa; Pastuizaca-Fernández et al. (2015) presentaron un gráfico de control multivariante multinomial T^2 con un enfoque difuso.

Saltos et al. (2020) aseguran que las herramientas de control de la calidad se pueden considerar no solo para monitorizar procesos industriales sino también procesos relacionados con la educación, por ejemplo, la evaluación del desempeño estudiantil. Estos autores aplicaron el concepto de profundidad, que transforma una observación multivariante a un índice univariante, el cual es susceptible de monitorizar en una carta de control y para esto utilizaron la carta r, además utilizaron clúster medio para establecer umbrales que faciliten la conformación de grupos y establecer perfiles de estudiantes mediante medidas descriptivas.

En el estudio de los procesos que se desarrollan en el entorno social se maneja, con mucha frecuencia, variables cualitativas. No es que estén ausentes los datos cuantitativos, sino que, en las bases de datos que se utilizan para estos análisis, abundan las variables cualitativas nominales y ordinales, a veces, sobre las de tipo numérico.

López (2004) señala que al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los métodos multivariantes de reducción de la dimensión tratan de eliminarla combinando muchas variables observadas para quedarse con pocas variables ficticias que, aunque no observadas, sean combinación de las reales y sinteticen la mayor parte de la información contenida en sus datos.

En este caso se deberá tener en cuenta el tipo de variables que maneja. Si son variables cuantitativas las técnicas que le permiten este tratamiento pueden ser el Análisis de componentes principales, el Análisis factorial (Spearman, 1904; Thurstone 1947; Kaiser 1958), mientras que, si se trata de variables cualitativas, es recomendable la aplicación de un Análisis de Correspondencias Múltiples, Análisis de homogeneidad o un Análisis de Escalamiento multidimensional.

Por otra parte, el uso de herramientas computacionales ha contribuido al desarrollo de los gráficos de control estadístico multivariante. Diversos autores han implementado paquetes estadísticos que facilitan la aplicación de gráficos de control clásicos, entre ellos Curran y Hersh (2021), que desarrollaron el aplicativo para el Gráfico T^2 de Hotelling, y Scrucca, L. (2004), el paquete informático para los gráficos MCUSUM y MEWMA. Otros investigadores han publicado originalmente sus aportaciones sobre gráficos de control acompañadas de una herramienta estadística que facilita su aplicación, entre ellos Ruiz-Barzola (2013), Epprecht, Aparisi y García-Bustos (2013).

Además, como complemento del análisis estadístico de procesos multivariantes realizado mediante gráficos de control, se puede utilizar aplicaciones estadísticas basadas en

el programa R que permiten el uso de biplots y otros métodos multivariantes para dos y tres vías, como MULTBIPLOT (Vicente-Villardón, 2010), ade4 (Dray y Dufour, 2007; Bougeard y Dray, 2018), FactoMineR (Le, Josse y Husson, 2008), SparseBiplots (Cubilla-Montilla, et al., 2021), biplotbootGUI (Nieto-Librero, 2015). En el mismo sentido se puede utilizar bibliotecas de Python, como SciPy (Virtanen, et al., 2020) o statsmodels (Seabold, Skipper, and Perktold, 2010).

En el control estadístico de procesos, los aportes al desarrollo de gráficos de control para variables cualitativas todavía son incipientes, las pocas publicaciones se orientan al análisis de características de la calidad en procesos industriales, pero no a procesos sociales. Al analizar los procedimientos publicados por los autores citados en este estudio, se detectan limitaciones que podrían restringir su aplicación, por ejemplo, el análisis de pocas características de la calidad, el uso de muestras constituidas por elementos individuales en vez de grupos, la dificultad de trabajar con muchas categorías de forma simultánea.

Surge, entonces, la necesidad de un gráfico de control para la representación de p variables cualitativas, que pueda trabajar con múltiples categorías nominales y ordinales y que facilite la identificación de las causas que pueden llevar al proceso a un estado fuera de control y que pueda ser aplicado en procesos sociales.

Esta investigación atiende las limitaciones antes mencionadas en cuanto a gráficos de control para variables cualitativas y su aplicación en entornos sociales. Por tal motivo, se presenta un gráfico de control para variables cualitativas mediante el uso de metodologías estadísticas multivariantes, que contribuya a la diversificación de técnicas en la fase I del control estadístico de procesos.

2.2. T2Qv, Gráfico de Control de Procesos Estadísticos Multivariantes para variables cualitativas

En este apartado se describe la propuesta metodológica para el control de procesos estadísticos multivariantes con variables cualitativas.

2.2.1. Notación

La tabla 2.1 contiene elementos y ejemplos de la manera como se presentan los elementos algebraicos abordados en la metodología.

Tabla 2.1. Elementos algebraicos

ELEMENTOS	REPRESENTACIÓN	EJEMPLO
Escalares	Letras en minúscula	v, λ
Vectores	Letras en minúscula y negrita	$\mathbf{v}, \boldsymbol{\mu}$
Matrices	Letras en mayúscula y negrita	$\boldsymbol{V}, \boldsymbol{X}$
Matrices de tres vías (cubos de datos)	Letras con doble trazo en mayúscula	$\mathbb{C}, \mathbb{X}$

2.2.2. Análisis de Correspondencias Múltiples (MCA)

Dado que se trabaja con variables cualitativas se aplica Análisis de Correspondencias (Benzecri, 1973) analizando la similaridad entre categorías (López, 2004) basada en la distancia χ^2, siendo este un análisis similar al de componentes principales.

El MCA es la aplicación del método de Análisis de Correspondencias Simples (CA) a datos categóricos multivariantes codificados en forma de una matriz de indicadores o una matriz de Burt (Nenadic, O. y Greenacre, 2007). Se trata de una técnica de análisis factorial exploratorio para datos categóricos multivariantes que describe, en un espacio de pocas

dimensiones, la estructura de asociaciones entre un grupo de variables categóricas, así como las similitudes y diferencias entre los individuos a los cuales esas variables se aplican.

EL MCA ha sido 'reinventado' en varias ocasiones, por diversos autores, y bajo nombres o enfoques diferentes (Ledesma, 2008). En esta investigación no se aplica el enfoque francés (Michailidis y Leeuw, 1998), si no el anglosajón, donde el MCA se denomina Análisis de Homogeneidad o Escalamiento Dual, haciendo uso de la tabla de Burt (Benzecri, 1973) y partiendo de una matriz de datos con p variables cualitativas, cada una con h categorías (h >1).

La matriz está compuesta por las n filas u observaciones y las p columnas o variables, donde cada celda contine una de las categorías antes mencionadas. Esto es equivalente a la matriz disyuntiva Z, que desglosa las variables en cada una de sus modalidades y registra la ocurrencia de eventos de forma binaria.

La tabla de Burt viene dada por:

$$\boldsymbol{B} = \boldsymbol{Z}'\boldsymbol{Z} \tag{3.1}$$

La matriz $\boldsymbol{B}$ en la Ecuación 3.1 está formada por las frecuencias absolutas, éstas se transforman en frecuencias relativas, dividiendo los valores de la matriz por la frecuencia total, dando lugar a la matriz $\boldsymbol{P}$.

Se obtienen los vectores de Masas de fila ($\boldsymbol{mf}$) y columna ($\boldsymbol{mc}$), a través de las marginales de las filas y de las columnas de la matriz $\boldsymbol{P}$, respectivamente.

Se obtiene la matriz de residuos estandarizados $\boldsymbol{S}$.

$$\boldsymbol{S} = \boldsymbol{D_{fila}}^{-\frac{1}{2}}(\boldsymbol{P} - mf\ mc')\boldsymbol{D_{columna}}^{-\frac{1}{2}} \tag{3.2}$$

donde $\mathbf{D_{fila}}$ es una matriz diagonal que contiene las masas de las filas y $\mathbf{D_{columna}}$ es una matriz diagonal que contiene las masas de las columnas.

Se aplica la descomposición en valores singulares (SVD) a la matriz S (Ecuación 3.3):

$$\boldsymbol{S} = \boldsymbol{UDV'} \tag{3.3}$$

donde $\boldsymbol{U}$ y $\boldsymbol{V}$ son matrices ortogonales y $\boldsymbol{D}$ es una matriz diagonal que contiene los valores singulares.

Luego se obtienen las coordenadas estandarizadas, para lo cual se aplican las ecuaciones 3.4 y 3.5.

$$\boldsymbol{X} = \boldsymbol{D_{fila}}^{-\frac{1}{2}}\boldsymbol{U} \tag{3.4}$$

$$\boldsymbol{Y} = \boldsymbol{D_{columna}}^{-\frac{1}{2}}\boldsymbol{V} \tag{3.5}$$

2.2.2.1. Generalización a k tablas.- El gráfico T2Qv, propuesto en esta investigación, no se limita a procesar una simple tabla de datos, sino que puede manejar bases de datos con múltiples tablas tomadas en diferentes momentos (K) y representarlas como puntos en el gráfico. Esto se puede configurar como un cubo de datos de $n\ x\ p\ x\ K$. Si hay K tablas, con la misma estructura y compuestas por variables cualitativas, lo descrito en la Sección 2.2.2.1 se aplica a cada una de las K tablas, obteniendo el conjunto de K tablas con el formato inicial (Figura 2.1).

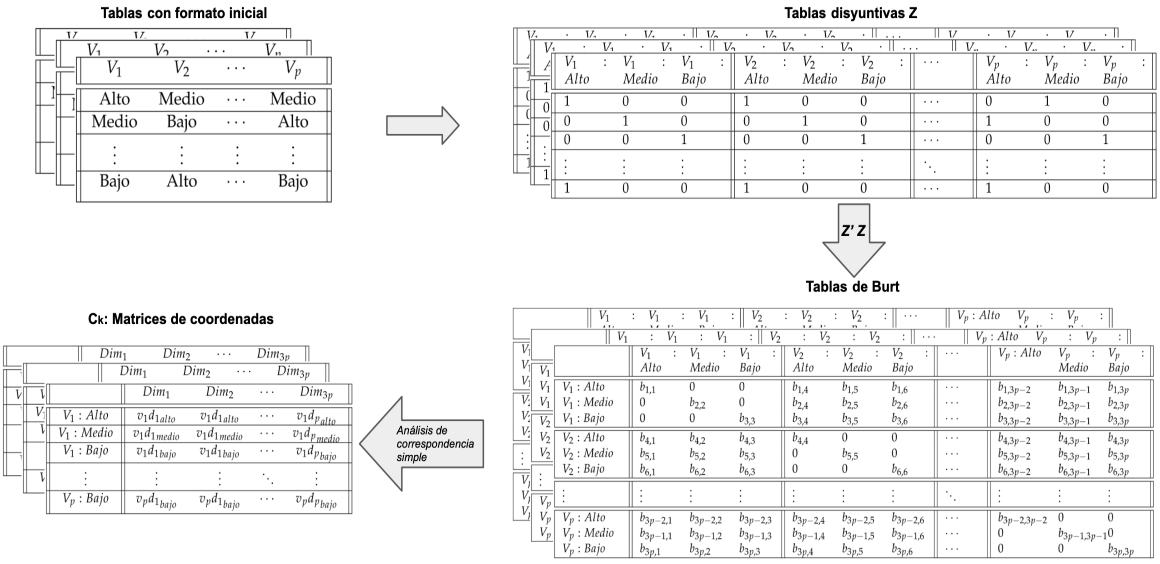


Figura 2.1. Procedimiento del MCA para *k* tablas

Cada uno de los K conjuntos de coordenadas obtenidos en el paso anterior se denota como $\boldsymbol{C}$. Para detectar la magnitud de las variables latentes, se toma el valor absoluto de los elementos de la matriz $\boldsymbol{C_k}(k = 1, \dots, K)$. De esta manera, se obtiene un conjunto de K tablas de coordenadas (cargas), cuyas filas corresponden a las variables observadas y las columnas a las variables latentes.

2.2.2.2. Normalización de tablas.- A las *k* tablas $\boldsymbol{C}$ se les aplica la normalización (Escofier y Pagès, 1994) del Análisis factorial múltiple (MFA). Sea λ_1^k el primer valor propio obtenido de la descomposición singular de la *k*-ésima tabla $\boldsymbol{C}$. Se normaliza la tabla

multiplicándola por $\frac{1}{\lambda_1^k}$. Con esto se obtiene la tabla $\boldsymbol{C'}$, que corresponde a la tabla de coordenadas normalizadas.

Individualmente, para el caso de la matriz *k,* se tendría la siguiente expresión.

$$\boldsymbol{C}'_k = \frac{1}{\lambda_k^1}\boldsymbol{C}_k \tag{3.6}$$

Hasta este punto se tiene un conjunto de matrices de coordenadas normalizadas, cuyas filas contienen las variables observadas y en las columnas a las variables latentes.

Unificando las k tablas normalizadas $\boldsymbol{C'}$ en una sola, se tiene la matriz $\mathbb{C}'$, denominada Matriz Concatenada. Esta contiene todos los elementos de las k tablas normalizadas.

$$\mathbb{C}' = [\boldsymbol{C}'_1|\boldsymbol{C}'_2|, \dots, |C'_K]' \tag{3.7}$$

La normalización a través del MFA pondera las k tablas, con el objetivo de evitar alguna descompensación al momento de realizar el análisis conjunto de las tablas.

A partir de las matrices $\mathbb{C}'$ y $\mathbf{C}'_{\boldsymbol{k}}$, se obtiene los vectores de mediana, tal como se muestra en la Figura 2.2. El vector $\tilde{\mathbf{x}}_{\mathbf{C}'_{\boldsymbol{k}}}$ explicará el comportamiento central de cada una de las k tablas y el vector $\tilde{\mathbf{x}}_{\mathbb{C}'}$ explicará el comportamiento de la matriz concatenada.

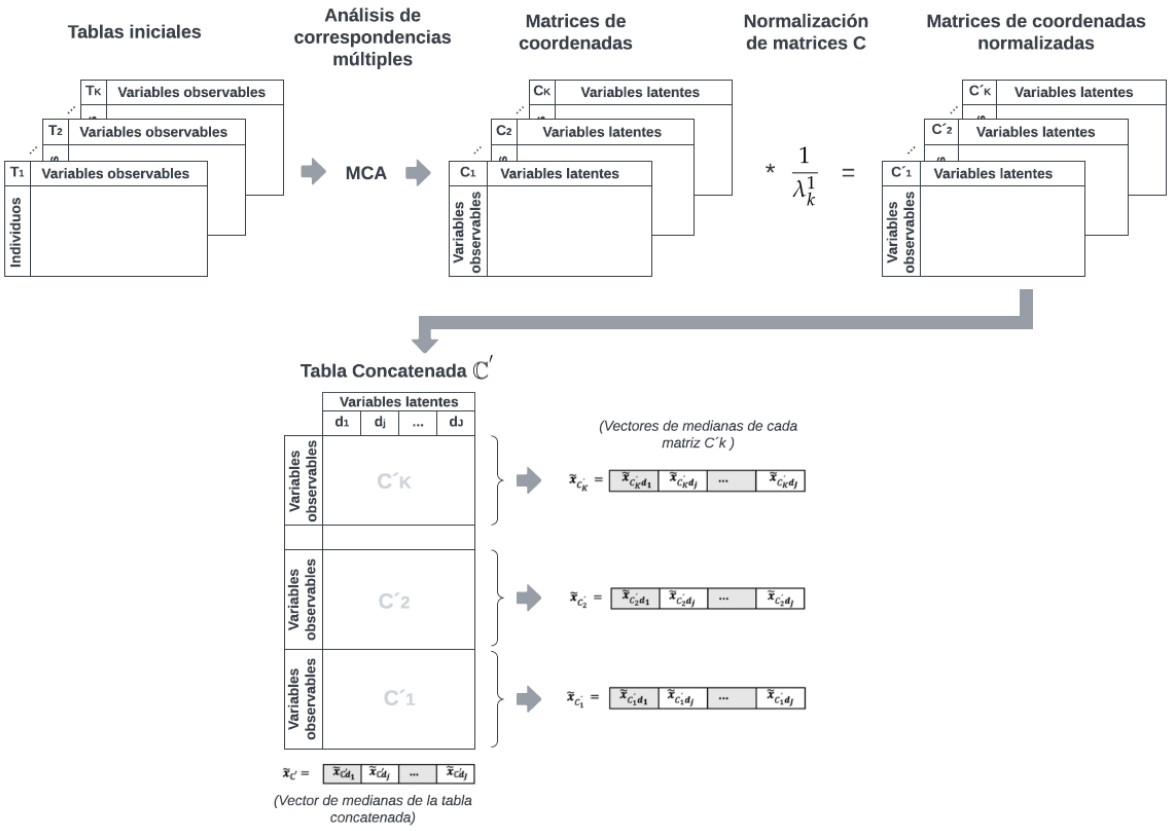


Figura 2.2. Esquema del proceso de obtención de vectores de medianas

2.2.3. Construcción del gráfico de control T2Qv

Para definir el gráfico de control T2Qv se deben tomar las siguientes consideraciones:

- La tabla $\mathbb{C}'$ (Ecuación 3.7) se denomina Concatenada, sirve como referente para definir el escenario bajo control en la fase I del control del proceso.
- El estadístico T^2 Hotelling normalmente se calcula con los vectores de media y la matriz de covarianzas del proceso bajo control. La propuesta de esta investigación es adoptar conceptos de robustez, utilizando el vector de medianas en vez de el vector de medias, en virtud de que a las medianas no les afectan los valores atípicos.

- De la matriz concatenada $\mathbb{C}'$ se obtiene $\tilde{\boldsymbol{x}}_0$ (Vector de medianas de la matriz concatenada) y $\boldsymbol{S}_0$ (Matriz de covarianzas de la matriz concatenada).
- Cada matriz $\mathbf{C'_k}$ tiene el mismo número de columnas.
- El vector de medianas $\tilde{\boldsymbol{x}}_k$ está atado a la tabla $\mathbf{C'_k}$, es decir, el gráfico de control estará en función de las diferencias entre las matrices $\mathbf{C'_k}$ y la matriz concatenada $\mathbb{C}'$.
- Las matrices $\mathbf{C'_k}$ siguen una distribución normal multivariante con vector de mediana $\tilde{\boldsymbol{x}}_k$ y matriz de covarianzas $\boldsymbol{S}_k$.

El estadístico T^2 viene dado por:

$$T^2 = n(\mu_k - \mu_0)'\boldsymbol{\Sigma}_0^{-1}(\mu_k - \mu_0) \tag{3.8}$$

Tomando en cuenta las consideraciones previas, se obtiene el estadístico T^2_{med}

$$T^2_{med} = n(\tilde{x}_k - \tilde{x}_0)'\boldsymbol{\Sigma}_0^{-1}(\tilde{x}_k - \tilde{x}_0) \tag{3.9}$$

Se sabe que la distribución T^2 converge a una distribución Chi-cuadrado con p grados de libertad (χ^2_p) cuando los datos provienen de una distribución normal multivariante (Aparisi, 1996). Además, autores como Gneri y Pimentel-Barbosa (2012) argumentan que T^2 converge a una distribución chi-cuadrado, incluso en escenarios donde los datos no provienen de una distribución normal multivariante, bajo ciertas condiciones. En este caso se puede aplicar este principio, ya que se utiliza la matriz concatenada ($\mathbb{C}'$), que representa al escenario bajo control.

Dado que este gráfico de control está basado en distancias de Mahalanobis ponderadas, su límite de control viene dado por la ecuación

$$UCL = \chi^2_{\alpha,p} \tag{3.10}$$

donde p es el número de dimensiones y α es el nivel de significancia predeterminado, se considera $\alpha = 0.0027$. Cuando el estadístico T^2_{med} de alguna muestra supera este límite, significa que el proceso está fuera de control (Figura 2.3).

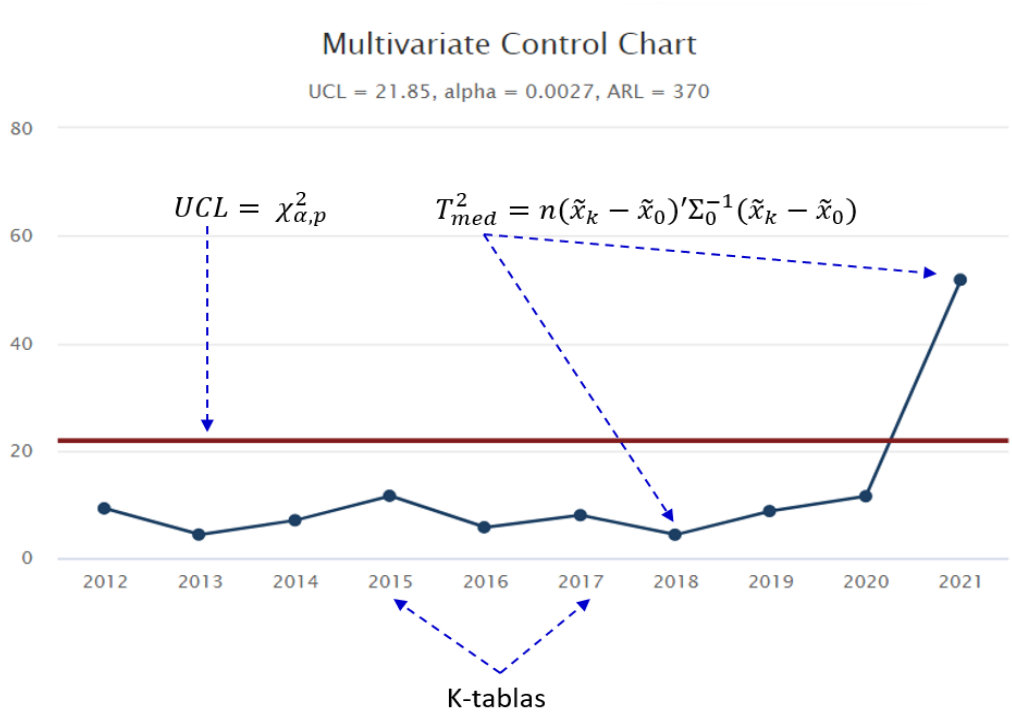


Figura 2.3. Gráfico T2Qv

2.2.4. Interpretación de puntos fuera de control

El gráfico multivariante para variables cualitativas, T2Qv, es capaz de señalar que el proceso salió de control, pero no permite reconocer las causas por las que ocurrió esto. Cada punto representado en el gráfico representa a una tabla (muestra), constituida por un grupo de individuos (observaciones) y p variables que pueden tener muchas categorías, algunas de éstas pueden mostrar un comportamiento anómalo. Por consiguiente, es necesario analizar

con detenimiento qué está pasando con los datos de las tablas reportadas para identificar la(s) variable(s) que generaron que el proceso se salga de control.

Este análisis se realiza comparando la ubicación de los puntos que representan las categorías de las variables en el MCA de la tabla concatenada y la ubicación de los puntos en los gráficos MCA de cada tabla reportada como fuera de control (tabla 2021 en la Figura 2.3). Las categorías que están incidiendo en el estado fuera de control son aquellas que muestran diferencias notorias en su ubicación al comparar ambas tablas.

Para cuantificar la magnitud de dichas diferencias o del comportamiento anómalo de estas categorías se calculan las distancias Chi-cuadrado entre las masas de columnas de la tabla reportada como fuera de control y las columnas de la tabla concatenada, tomada esta última como referente. Mientras mayor es el valor del estadístico, mayor es su incidencia en el desplazamiento de la centralidad del proceso que, finalmente, pueden llevarlo a un estado fuera de control.

Estas distancias Chi cuadrado se presentan como una tabla y como un gráfico de barras interactivo, en el que la longitud de las barras está dada por la distancia chi cuadrado. Además, para analizar con más detalle el comportamiento de las variables del proceso, las distribuciones de las categorías se representan, en porcentajes, a través de gráficos circulares para cada variable de la tabla punto y de la tabla concatenada.

De manera que la interpretación de los puntos fuera de control se realiza mediante tres técnicas: el Análisis de Correspondencias Múltiples; el análisis de la distancia chi cuadrado, expresado como tabla y como gráfico de barras; y, el uso de gráficos de sectores, para la distribución de las categorías de las variables en las tablas comparadas.

2.3. Resultados con datos simulados

Con la intención de probar la metodología propuesta en el gráfico de control para variables cualitativas, se hizo un análisis con datos simulados. Los datos están cargados en el paquete T2Qv.

2.3.1. Generación de datos simulados

Para este estudio se generó una base de datos simulados, a la que se denominó *Datak10Contaminated*. Consta de 10 tablas, cada una de ellas está constituida por 100 filas (observaciones) y 11 columnas, de las cuales, las 10 primeras corresponden a las variables analizadas ($V01, V02, \dots ; V10$), mismas que contienen 3 categorías (*High*, *Medium* y *Low*), mientras que, la columna 11, denominada *GroupLetter*, contiene el factor de clasificación de los grupos. Una sección de la base de datos *Datak10Contaminated* se presenta en la Tabla 2.2.

Para su identificación, las tablas han sido denominadas con las letras del alfabeto, desde la a hasta la *j*. La tabla *j* tiene una distribución distinta de la que tienen las otras nueve. Las 9 primeras tablas (*a, …, i*) tienen sus 10 variables con la siguiente distribución:

$$u \sim U[0,1]$$

$$t_{a,\dots,i} = \begin{cases} Bajo & si \quad u \leq \frac{1}{3} \\ Medio & si \quad \frac{1}{3} < u < \frac{2}{3} \\ Alto & si \quad u \geq \frac{2}{3} \end{cases}$$

La tabla *j*, en todas sus 10 variables, sigue la distribución presentada a continuación:

$$u \sim U[0,1]$$

$$t_j = \begin{cases} Bajo & si \quad u \leq \frac{1}{5} \\ Medio & si \quad \frac{1}{5} < u < \frac{2}{6} \\ Alto & si \quad u \geq \frac{2}{6} \end{cases}$$

Tabla 2.2. Sección de la base de datos *Datak10Contaminated*

V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	GroupLetter
Bajo	Medio	Medio	Alto	Alto	Alto	Bajo	Medio	Medio	Medio	a
Bajo	Bajo	Alto	Bajo	Medio	Alto	Alto	Alto	Bajo	Alto	a
Alto	Medio	Alto	Bajo	Alto	Medio	Medio	Alto	Medio	Bajo	a
Medio	Medio	Bajo	Alto	Bajo	Medio	Alto	Bajo	Bajo	Alto	a
Bajo	Bajo	Bajo	Alto	Bajo	Alto	Alto	Alto	Medio	Medio	a
Alto	Alto	Medio	Bajo	Alto	Bajo	Medio	Medio	Alto	Bajo	a
Alto	Alto	Bajo	Bajo	Bajo	Medio	Alto	Medio	Medio	Alto	a
Medio	Medio	Alto	Medio	Medio	Alto	Medio	Alto	Alto	Alto	a
Bajo	Bajo	Bajo	Medio	Alto	Medio	Bajo	Medio	Bajo	Bajo	a
Medio	Medio	Medio	Alto	Bajo	Medio	Alto	Bajo	Alto	Medio	a
Bajo	Alto	Bajo	Alto	Alto	Medio	Medio	Bajo	Bajo	Bajo	a
Medio	Bajo	Bajo	Alto	Bajo	Medio	Alto	Bajo	Medio	Bajo	a
Alto	Bajo	Medio	Alto	Bajo	Alto	Alto	Alto	Alto	Medio	a
Alto	Alto	Bajo	Medio	Bajo	Medio	Bajo	Alto	Alto	Medio	a

Para verificar la diferencia entre las distribuciones de la tabla *j* y las demás, se calculó el promedio de las frecuencias relativas en las tres categorías, desde la tabla *a* hasta la *i*, para las 10 variables (Tabla 2.3.), luego se calculó el promedio de las frecuencias relativas medias de las 10 variables. El resultado permite comparar la distribución de las categorías de la tabla *Datak10Contaminated* con la distribución teórica uniforme, como se observa en la Tabla 2.4.

Tabla 2.3. Promedio de frecuencias relativas medias en las tres categorías, *Datak10Contaminated*

Tabla	Categoría	V01	V02	V03	V04	V50	V06	V07	V08	V09	V10
a	**High**	0.29	0.25	0.36	0.38	0.38	0.35	0.36	0.29	0.33	0.37
a	**Medium**	0.36	0.49	0.34	0.34	0.31	0.41	0.38	0.28	0.38	0.31
a	**Low**	0.35	0.26	0.30	0.28	0.31	0.24	0.26	0.43	0.29	0.32
b	**High**	0.31	0.44	0.37	0.29	0.31	0.34	0.30	0.36	0.29	0.34
b	**Medium**	0.40	0.31	0.30	0.35	0.37	0.32	0.35	0.30	0.39	0.36
b	**Low**	0.29	0.25	0.33	0.36	0.32	0.34	0.35	0.34	0.32	0.30
c	**High**	0.34	0.33	0.25	0.35	0.32	0.30	0.39	0.40	0.41	0.43
c	**Medium**	0.36	0.33	0.25	0.32	0.32	0.32	0.27	0.35	0.32	0.35
c	**Low**	0.30	0.34	0.50	0.33	0.36	0.38	0.34	0.25	0.27	0.22
d	**High**	0.32	0.34	0.34	0.38	0.41	0.33	0.35	0.46	0.34	0.45
d	**Medium**	0.35	0.30	0.28	0.31	0.27	0.35	0.30	0.24	0.33	0.24
d	**Low**	0.33	0.36	0.38	0.31	0.32	0.32	0.35	0.30	0.33	0.31
e	**High**	0.32	0.32	0.36	0.26	0.36	0.31	0.29	0.28	0.32	0.41
e	**Medium**	0.34	0.40	0.34	0.40	0.38	0.37	0.27	0.37	0.32	0.23
e	**Low**	0.34	0.28	0.30	0.34	0.26	0.32	0.44	0.35	0.36	0.36
f	**High**	0.31	0.29	0.27	0.32	0.36	0.32	0.26	0.41	0.34	0.26
f	**Medium**	0.41	0.29	0.36	0.31	0.31	0.38	0.36	0.33	0.30	0.37
f	**Low**	0.28	0.42	0.37	0.37	0.33	0.30	0.38	0.26	0.36	0.37
g	**High**	0.27	0.39	0.34	0.38	0.28	0.31	0.35	0.38	0.27	0.34
g	**Medium**	0.42	0.27	0.32	0.35	0.37	0.32	0.35	0.36	0.41	0.26
g	**Low**	0.31	0.34	0.34	0.27	0.35	0.37	0.30	0.26	0.32	0.40
h	**High**	0.32	0.47	0.34	0.38	0.47	0.34	0.32	0.35	0.35	0.31
h	**Medium**	0.28	0.31	0.29	0.27	0.27	0.43	0.39	0.35	0.36	0.40
h	**Low**	0.40	0.22	0.37	0.35	0.26	0.23	0.29	0.30	0.29	0.29
i	**High**	0.32	0.42	0.29	0.30	0.26	0.28	0.38	0.38	0.36	0.36
i	**Medium**	0.35	0.34	0.29	0.33	0.47	0.38	0.25	0.29	0.33	0.31
i	**Low**	0.33	0.24	0.42	0.37	0.27	0.34	0.37	0.33	0.31	0.33
j	**High**	0.75	0.71	0.78	0.71	0.70	0.73	0.69	0.66	0.73	0.78
j	**Medium**	0.08	0.10	0.01	0.06	0.10	0.12	0.11	0.12	0.12	0.10
j	**Low**	0.17	0.19	0.21	0.23	0.20	0.15	0.20	0.22	0.15	0.12
$\bar{x}$ a: i	**High**	**0.31**	**0.37**	**0.33**	**0.33**	**0.35**	**0.32**	**0.33**	**0.37**	**0.33**	**0.36**
$\bar{x}$ a: i	**Medium**	**0.37**	**0.34**	**0.31**	**0.33**	**0.34**	**0.36**	**0.33**	**0.32**	**0.35**	**0.32**
$\bar{x}$ a: i	**Low**	**0.32**	**0.30**	**0.36**	**0.33**	**0.31**	**0.32**	**0.34**	**0.32**	**0.32**	**0.32**

Tabla 2.4. Comparación de la distribución de las categorías de la tabla *Datak10Contaminated* con la distribución teórica uniforme.

Categorías	Distribución teórica uniforme	Media de la distribución de las variables en las Tablas a, b, …, i	Media de la distribución de las variables en la Tabla j
High	0.333	0.340	0.724
Medium	0.333	0.336	0.092
Low	0.333	0.324	0.184

Se aplicaron pruebas Chi cuadrado de bondad de ajuste para confirmar la distribución de los datos generados. Se hizo la comparación de la distribución de la tabla *j* con las demás tablas, confirmándose las diferencias significativas entre las distribuciones (p-valor < 0.05), como se observa en la tabla 2.5.

Tabla 2.5. Estadísticos de prueba para la comparación de las distribuciones de las categorías de las 10 variables entre la tabla j y las demás, *Datak10Contaminated*

Group Letter	Estadísticos	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
a	Chi-cuadrado	0.860	11.060	0.560	1.520	0.980	4.460	2.480	4.220	1.220	0.620
	p-valor	0.651	0.004	0.756	0.468	0.613	0.108	0.289	0.121	0.543	0.733
b	Chi-cuadrado	2.060	5.660	0.740	0.860	0.620	0.080	0.500	0.560	1.580	0.560
	p-valor	0.357	0.059	0.691	0.651	0.733	0.961	0.779	0.756	0.454	0.756
c	Chi-cuadrado	0.560	0.020	12.500	0.140	0.320	1.040	2.180	3.500	3.020	6.740
	p-valor	0.756	0.990	0.002	0.932	0.852	0.595	0.336	0.174	0.221	0.034
d	Chi-cuadrado	0.140	0.560	1.520	0.980	3.020	0.140	0.500	7.760	0.020	6.860
	p-valor	0.932	0.756	0.468	0.613	0.221	0.932	0.779	0.021	0.990	0.032
e	Chi-cuadrado	0.080	2.240	0.560	2.960	2.480	0.620	5.180	1.340	0.320	5.180
	p-valor	0.961	0.326	0.756	0.228	0.289	0.733	0.075	0.512	0.852	0.075
f	Chi-cuadrado	2.780	3.380	1.820	0.620	0.380	1.040	2.480	3.380	0.560	2.420
	p-valor	0.249	0.185	0.403	0.733	0.827	0.595	0.289	0.185	0.756	0.298
g	Chi-cuadrado	3.620	2.180	0.080	1.940	1.340	0.620	0.500	2.480	3.020	2.960
	p-valor	0.164	0.336	0.961	0.379	0.512	0.733	0.779	0.289	0.221	0.228
h	Chi-cuadrado	2.240	9.620	0.980	1.940	8.420	6.020	1.580	0.500	0.860	2.060
	p-valor	0.326	0.008	0.613	0.379	0.015	0.049	0.454	0.779	0.651	0.357
i	Chi-cuadrado	0.140	4.880	3.380	0.740	8.420	1.520	3.140	1.220	0.380	0.380
	p-valor	0.932	0.087	0.185	0.691	0.015	0.468	0.208	0.543	0.827	0.827
j	Chi-cuadrado	79.340	65.060	95.780	68.180	62.000	70.940	58.460	49.520	70.940	89.840
	p-valor	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

0 casillas (0.0%) han esperado frecuencias menores que 5. La frecuencia mínima de casilla esperada es 33.3.

2.3.2. Aplicación del aplicativo T2Qv con datos simulados

El primer resultado es el gráfico de control T2Qv (Figura 2.4), está basado en el estadístico T^2 de Hotelling ajustado (T^2_{med}), aplicado a la detección de anomalías en

cualquiera de las *k* tablas analizadas. Cada una de las tablas de la base de datos está representada por los puntos en el gráfico, por esta razón, de manera general reciben la denominación de "Tabla Punto", seguida del nombre específico de la tabla, así, por ejemplo, la tabla *Punto a*, la *Punto b*, la *Punto j*. En el gráfico se observa también una línea horizontal que representa al límite de control superior. El límite de control inferior es igual a cero.

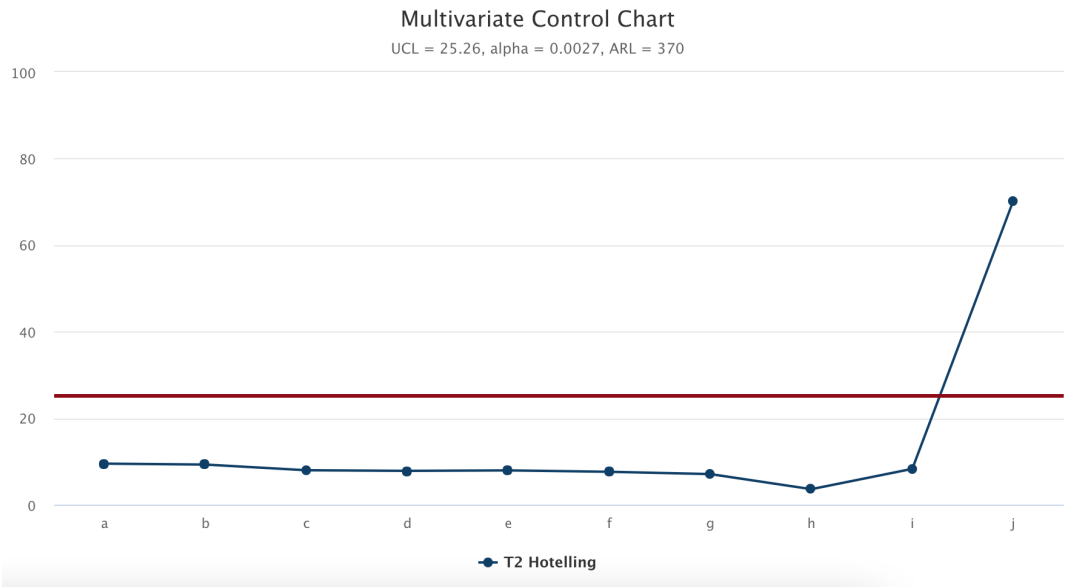


Figura 2.4. Gráfico de control multivariante T2Qv aplicable a variables cualitativas, *Datak10Contaminated*

Por otra parte, el punto que representa a la tabla *j* se ubica por encima del límite de control superior, lo que quiere decir que la tabla Punto j ha sido identificada como un valor fuera de control. Por consiguiente, es necesario analizar con detenimiento qué está pasando con los datos de esta tabla, a fin de identificar las causas de variación y tomar las acciones pertinentes. Para hacer un análisis del punto fuera de control se realiza un gráfico del MCA

de la tabla *j* y se lo compara con el gráfico del MCA de la tabla concatenada, como se presenta en la Figura 2.5.

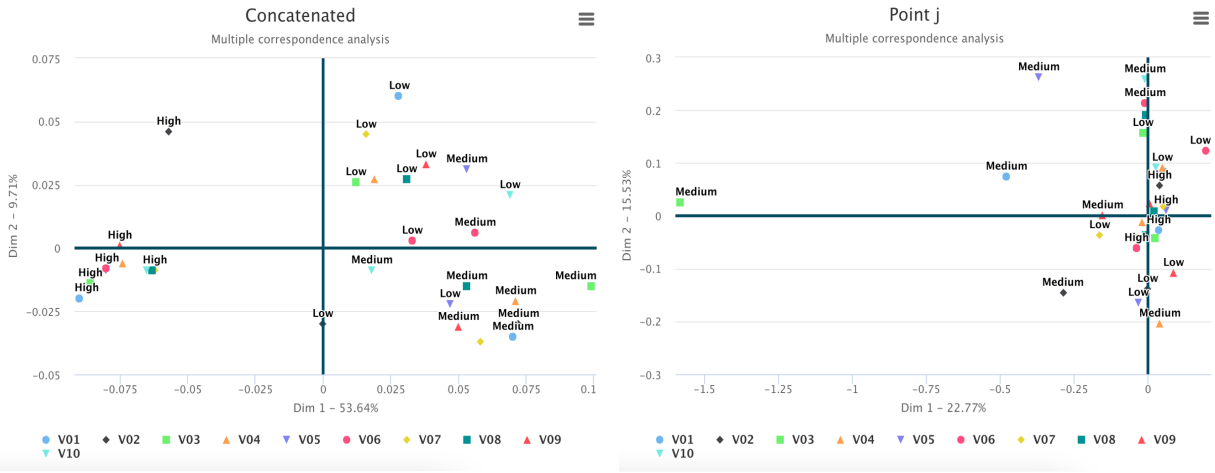


Figura 2.5. Gráfico del MCA de la tabla Concatenada y la Tabla Punto, *Datak10Contaminated*.

Otro resultado es el gráfico del MCA aplicado a la tabla Concatenada (Figura 2.6). Esta tabla es considerada el referente visual para el escenario bajo control en el análisis posterior de las tablas que resultan reportadas como puntos fuera de control por el gráfico T2Qv.

Los autores Hoffman y De Leeuw (1992) emiten directrices para la interpretación de un gráfico de MCA. Indican que las categorías de las variables se representan en un gráfico bidimensional mediante puntos, donde la distancia entre ellos indica la medida de homogeneidad de los perfiles, los patrones idénticos de respuesta se trazan como puntos muy cercanos entre sí. La posición de los puntos en el plano indica la frecuencia marginal de cada categoría, un punto de categoría con baja frecuencia marginal se trazará hacia el borde del

mapa, mientras que una categoría con alta frecuencia marginal se trazará más cerca al origen del gráfico.

Además, estos autores destacan la importancia de asociar los ejes del gráfico con características relevantes del proceso para poder etiquetar las dimensiones uno y dos y así discriminar los casos en función de esas características. Aunque los valores propios pueden sugerir que se necesiten más de dos dimensiones, se utilizan solo las dos primeras para simplificar la explicación.

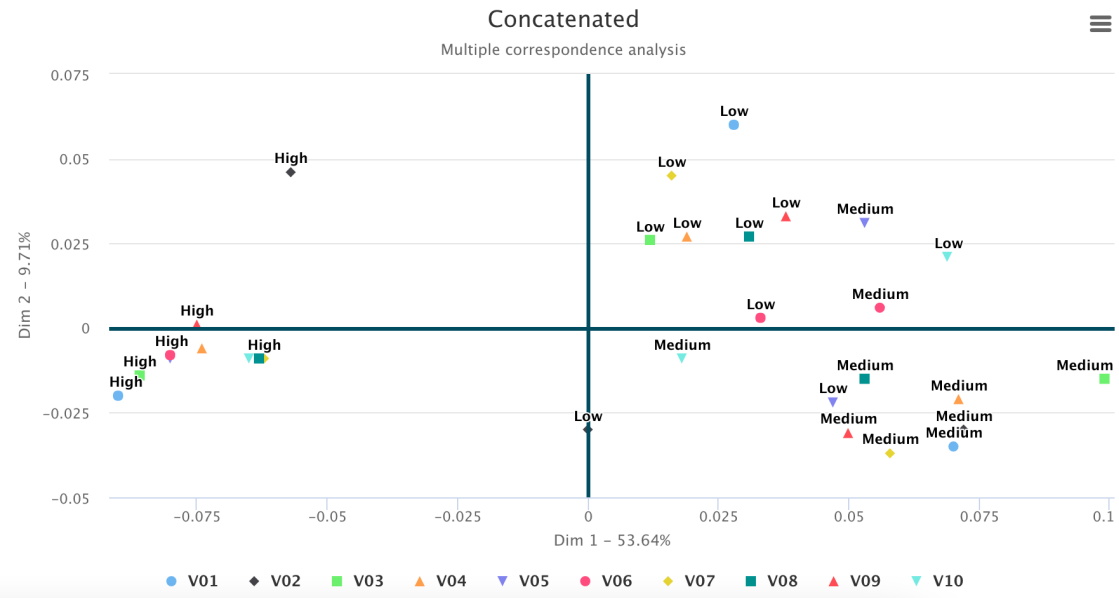


Figura 2.6. Análisis de Correspondencias Múltiples aplicado a la tabla concatenada

La Figura 2.6 reporta una inercia total del 63.35%, la dimensión 1 representa el 53.64% de la información, mientras que la dimensión 2, el 9.71%. Los puntos del gráfico muestran la ubicación de las categorías de cada una de las 10 variables en sus tres niveles: *High*, *Medium* y *Low*. En el gráfico, las observaciones que se ubiquen en el centro del gráfico

representan a las categorías que se presentan con mayor frecuencia, mientras que las más alejadas del centro son las que pocas veces aparecen, los casos raros. En este sentido, en la tabla concatenada no hay observaciones ubicadas en el centro del gráfico, sino que están repartidas en grupos, rodeando el centro, lo que se explica por la distribución uniforme de las categorías de las variables en la mayoría de las tablas, ninguna prevalece.

Salta a la vista que las categorías *High* de las variables representadas se han agrupado a la izquierda en el primer eje principal, mientras que, *Low* se ha organizado aproximadamente del centro hacia la derecha, y la categoría *Medium* encontró su lugar a la derecha del eje horizontal.

Se interpreta que, de manera general, hay una relación inversa entre las observaciones de *High* y *Medium*, esto significa que a medida que se incrementan las frecuencias de una categoría, disminuyen las de la otra. Este es el caso de las categorías *High* de la variable V01 y *Medium* de la variable V03, que forman un ángulo obtuso con relación al centro del gráfico. Asimismo, se observan categorías *High* que se ubican de forma opuesta a las *Low*, pero, también hay otras que forman ángulos más cerrados, casi rectos, lo que denota escasa o nula asociación entre categorías. Es el caso de las categorías *High* de la variable V02 y *Low* de la variable V01.

Como en el gráfico T2Qv (Figura 2.4) se muestra que el punto *j* se ubica más arriba del límite UCL, se interpreta que la tabla Punto j se encuentra fuera de control. Es necesario, entonces, analizar con más detalle qué está sucediendo, cuáles pueden ser las causas que originaron ese comportamiento anómalo en la tabla señalada. Por consiguiente, procede un análisis comparativo del gráfico del MCA de la tabla Concatenada y de la tabla Punto j. El

aplicativo T2Qv puede presentar por separado estos dos gráficos y también puede presentarlos juntos (Figura 2.5).

La Figura 2.5 presenta la distribución de las observaciones de las tablas Concatenada y Punto j mediante gráficos del MCA. El gráfico de la tabla concatenada, que sirve de referente, ya se analizó en la Figura 2.6. El gráfico MCA de la tabla *j* muestra Categorías *Medium* de algunas variables que se ubican al lado izquierdo del eje principal, alejándose del centro del gráfico, lo que indica que son poco frecuentes. Es el caso de las variables V01, V02, V05, pero especialmente la variable V03, que registra una observación para el nivel medio con el valor más alejado del grupo.

Las demás observaciones *Medium* y todas las *Low* se han situado alrededor del segundo eje, dejando la ubicación central del gráfico a las categorías *High* , lo que significa que esta categoría presenta mayor frecuencia que las demás. Esto tiene sentido si se considera que la distribución de la tabla *j* es *High* 0.724, *Medium* 0.092 y *Low* 0.184 (Tabla 2.4).

Para configurar una visión más completa del comportamiento de las variables en las diferentes tablas de la base de datos se puede realizar el gráfico del MCA de cualquier tabla Punto, que sea diferente de la tabla reportada como anómala. En este ejemplo se revisa el gráfico de la tabla Punto b (Figura 2.7).

En la Figura 2.7 es notorio que las categorías *High*, *Medium* y *Low* de todas las variables están distribuidas aleatoriamente en todos los cuadrantes del gráfico, no se puede precisar un patrón específico de agrupación.

Algo parecido ocurriría si se hiciera el análisis de cualquiera de las otras tablas de la base de datos *Datak10Contaminated*, porque comparten la misma distribución uniforme, a excepción de la tabla *j*, que fue diseñada con una distribución diferente.

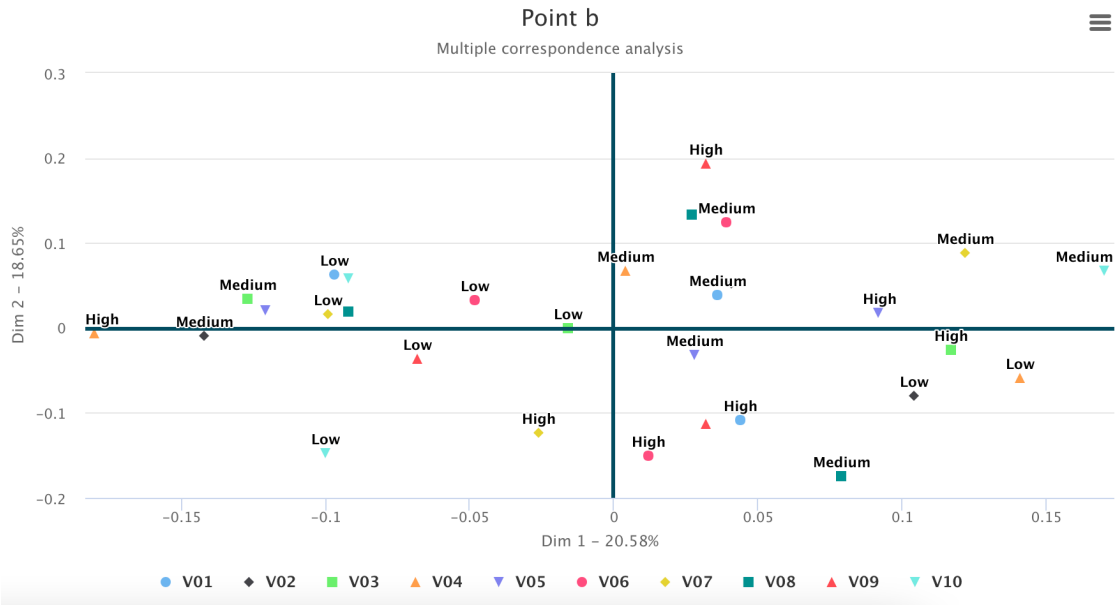


Figura 2.7. Análisis de Correspondencias Múltiples aplicado a la tabla Punto b

Al comparar los gráficos es evidente que la distribución de los datos en el gráfico de la tabla *j* es diferente de las distribuciones de las demás tablas, y en especial, es diferente de la distribución de los datos en el gráfico de la tabla concatenada, lo que explica por qué el punto *j* ha sido identificado como fuera de control en el gráfico T2Qv.

Esta diferencia se explica en la Tabla 2.6, que muestra la distancia Chi cuadrado entre las observaciones de la tabla concatenada y la tabla *j*.

Tabla 2.6. Distancia Chi cuadrado entre las masas de columnas de la tabla *j* y la concatenada, *Datak10Contaminated*.

Variables	ChiSq
V01	0.06968
V02	0.04293
V03	0.07700
V04	0.05209
V05	0.04385
V06	0.05938
V07	0.04521
V08	0.03046
V09	0.05440
V10	0.05840

Otra manera de visualizar esta información es a través un gráfico de barras que genera el aplicativo T2Qv (Figura 2.8).

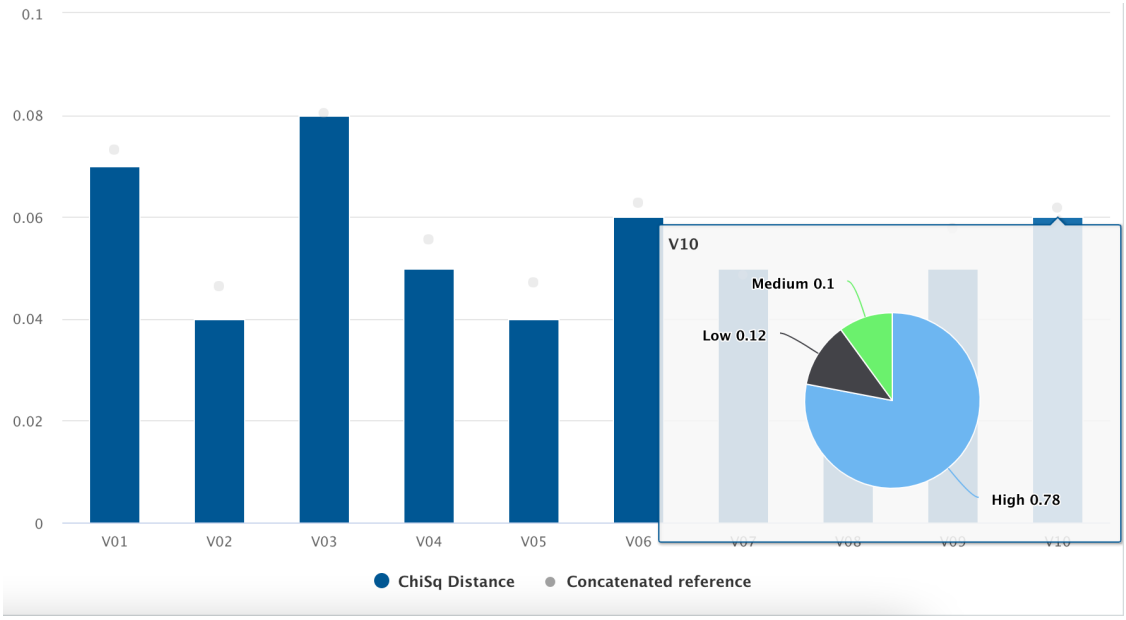


Figura 2.8. Distancia Chi cuadrado entre las masas de la tabla concatenada y la tabla j, *Datak10Contaminated*.

El gráfico de barras de la Figura 2.8 expresa también la distancia χ^2 entre las masas de la tabla Concatenada y las de las k tablas de la base de datos *Datak10Contaminated*, en este caso la *j*. En la Tabla 6 se observa que las variables V03, V01 y V06 manifiestan las mayores distancias Chi cuadrado entre las masas de la tabla concatenada y la tabla *j* (0.07700, 0.06968, 0.05938), lo que en la Figura 2.8 se representa con las barras más altas. Estas variables son las que están introduciendo más variabilidad al modelo, lo que genera en mayor grado cambios en la mediana del proceso y, en consecuencia, tienen mayor contribución a la salida de control del punto *j*.

La interactividad de este gráfico facilita la observación de la distribución de las categorías de las variables en la tabla Punto, y su comparación con la distribución de las categorías de las variables en la tabla concatenada, como se observa en la Figura 2.9.

La Figura 2.9 presenta, en gráficos circulares, la distribución de las categorías de las variables V03, V01 y V06, que registraron mayores distancias Chi cuadrado entre las masas de la tabla concatenada y la tabla *j*. Los gráficos que corresponden a la tabla concatenada presentan sectores con áreas equivalentes entre sí, lo que se explica por la distribución uniforme de las variables, mientras que, los de la tabla *j* muestran áreas con tamaños variados, donde la categoría *High* tiene una frecuencia relativa alta en los tres casos, y *Low*, baja frecuencia.

Al comparar estos gráficos se hace evidente que la distribución de las categorías presenta grandes diferencias entre la tabla concatenada y la tabla *j*.

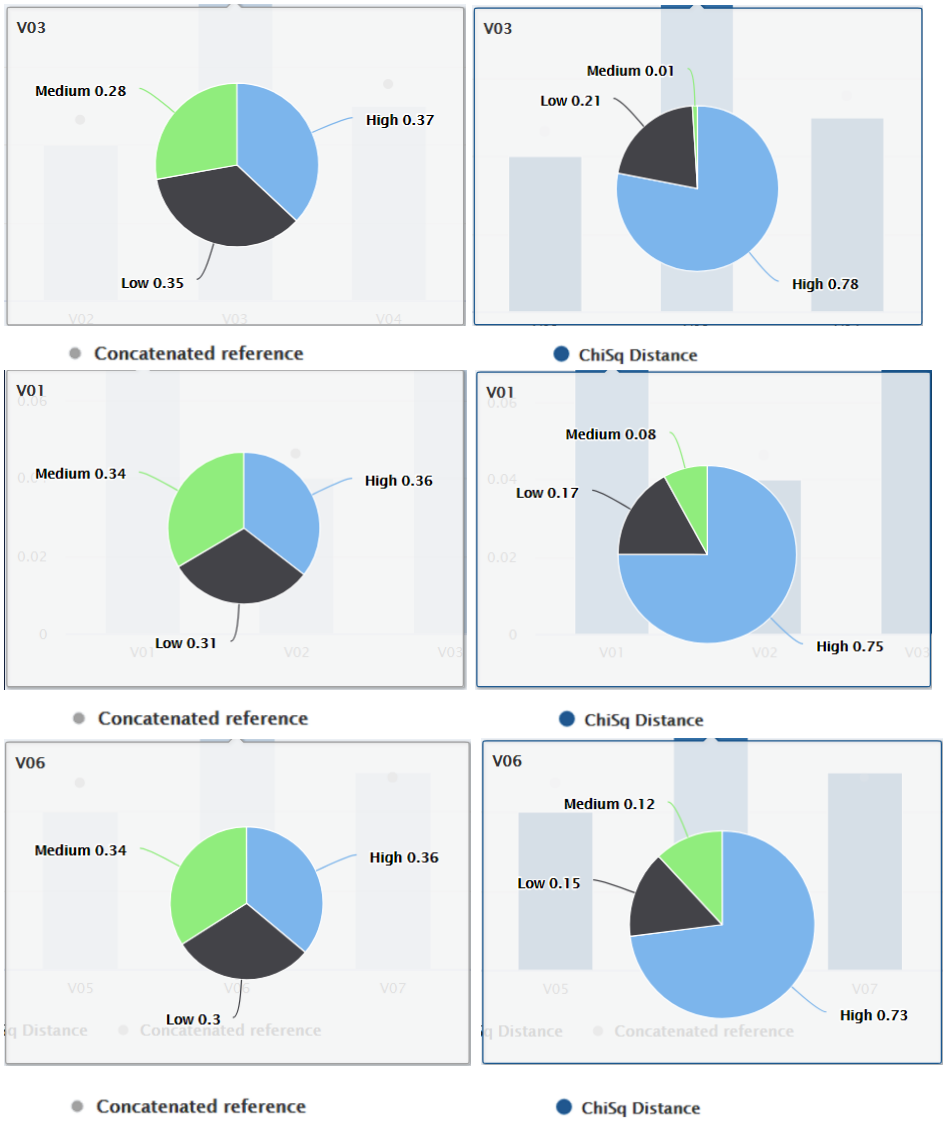


Figura 2.9. Distribución de las categorías de las variables V03, V01 y V06 en la tabla Concatenada y la tabla j en el aplicativo T2Qv

Se corrobora que el comportamiento de las variables V03, V01 y V06 tiene mayor incidencia en el desplazamiento de la tendencia central del proceso que, al final, lo lleva a un estado fuera de control. Sin embargo, al tratarse de un contexto multivariante, todas las variables contribuyen en mayor o menor medida a explicar el comportamiento del proceso, de manera

que la salida de control no se puede atribuir a la acción individual de una variable, o a la acción por separado de un grupo de ellas, sino al efecto combinado de las variables correlacionadas.

2.4. Análisis de sensibilidad

Como se ha establecido, en el gráfico T2Qv, un punto fuera de control se interpreta como una tabla (k_i) que incluye una cantidad o proporción de variables contaminadas. En estos casos, se espera que los puntos en el gráfico T2Qv generalicen el comportamiento de estas diferencias en su distribución y, por lo tanto, superen el límite de control superior. La ubicación de este límite de control varía según el número de dimensiones representadas, ya que se basa en el gráfico T^2 de Hotelling cuyo límite de control superior depende del número de variables consideradas (Ruiz-Barzola, 2013).

En el caso de T2Qv, este es el número de dimensiones latentes consideradas (Ecuación 10), y por lo tanto, una alta dimensionalidad logra un rendimiento óptimo, mientras que disminuir el número de dimensiones representables introduce inestabilidad y reduce la fiabilidad de los resultados.

El gráfico de control propuesto puede detectar un punto fuera de control incluso con un bajo número de variables contaminadas al trabajar con un alto número de dimensiones. Se recomienda usar $p - 1$, donde p es el número total de dimensiones en la matriz inicial (Figura 2.1). Cuando el número de dimensiones disminuye, la altura del límite de control superior (UCL) también disminuye, resultando en un mayor número de puntos fuera de control, aunque las variables no necesariamente expresen diferencias significativas en sus valores, aumentando la probabilidad de obtener un error de tipo I.

Por lo tanto, surge la pregunta de cuántas dimensiones se pueden reducir en el análisis sin perder fiabilidad en el resultado. La importancia de esta pregunta radica en la necesidad de un gráfico confiable que identifique puntos fuera de control incluso si se ha aplicado una técnica de reducción de dimensionalidad a los datos, sin caer en casos de falsos positivos.

El análisis de sensibilidad utiliza gráficos de contorno y gráficos de superficie de respuesta (Figura 2.10), para los cuales se utilizaron las funciones *persp3D* y *contour2D* del paquete *plot3D* (Soetaert, 2021). Se desarrolló una simulación de bases de datos con diferentes parámetros, considerando una variación en el porcentaje de variables contaminadas en la tabla k_i y el número de dimensiones representadas.

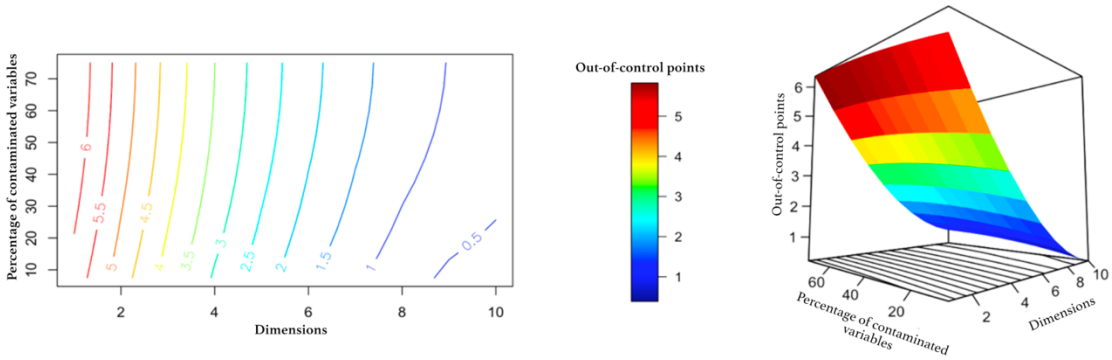


Figura 2.10. Gráficos de sensibilidad, gráficos de contorno y gráficos de superficie de respuesta a partir de la medida del comportamiento de la gráfica T2Qv se obtienen

El gráfico T2Qv se aplicó a cada simulación para evaluar el comportamiento del gráfico de control bajo pequeñas perturbaciones en el caso de pocas variables contaminadas y grandes perturbaciones en el caso de la mayoría de las variables contaminadas.

Los datos de prueba utilizados en el modelo se registran en 10 tablas, cada una de las cuales incluye 10 variables y cada variable tiene tres categorías: *High*, *Medium* y *Low*. La Tabla 10 (o tabla *j*) tiene una distribución diferente a las demás, siendo esta la tabla contaminada.

Se observa que el modelo puede identificar un punto fuera de control al trabajar con $p - 1$ dimensiones (9), incluso con un bajo porcentaje de variables contaminadas. Cuando el número de dimensiones disminuye a $p - 2$ (8) y el porcentaje de variables contaminadas está cerca del 100%, detecta correctamente 1 punto fuera de control. También se observa que cuando el número de dimensiones es menor, se pierde estabilidad y se reduce el poder de la prueba. En consecuencia, el análisis de sensibilidad confirma que el gráfico de control T2Qv funciona bien cuando se trabaja con dimensiones altas.

2.5. Contribuciones del Capítulo 2

En este apartado se presenta una propuesta metodológica para el control estadístico de procesos multivariantes cuando se trata de variables cualitativas. El análisisis de Correspondencias Múltiples (MCAy el gráfico T^2 de Hotelling se identifican como las técnicas centrales en que se basa la propuesta.

El MCA es una extensión del Análisis de Correspondencias Simples (CA) y es apropiado para analizar datos categóricos multivariantes. Su principal objetivo es describir, en un espacio de dimensiones reducidas, la estructura de asociaciones entre un conjunto de variables categóricas y las similitudes/diferencias entre los individuos a los cuales se aplican estas variables.

El MCA se basa en matrices de indicadores, siendo la matriz de Burt un ejemplo central. Esta matriz, que se define como $\boldsymbol{B} = \boldsymbol{Z}'\boldsymbol{Z}$, es derivada de la matriz disyuntiva Z. Esta última registra ocurrencias de eventos de manera binaria, desglosando las variables en sus modalidades individuales. La metodología implica la transformación de las frecuencias absolutas en la matriz $\boldsymbol{B}$ en frecuencias relativas, dando lugar a la matriz $\boldsymbol{P}$. Luego, se introducen conceptos como vectores de masas de fila y columna, y la matriz de residuos estandarizados, preparando el terreno para la descomposición en valores singulares (SVD).

La SVD, técnica esencial en el análisis multivariante, se aplica a la matriz de residuos estandarizados y la descompone en tres componentes fundamentales: U, V y D. Posteriormente, se obtienen coordenadas estandarizadas a partir de estas matrices, que serán utilizadas en análisis posteriores.

Esta metodología no se limita a analizar una única tabla de datos. Puede manejar bases de datos con múltiples tablas de dimensiones $n\ x\ p$ tomadas en K diferentes tiempos, lo que se puede configurar como un cubo de datos.

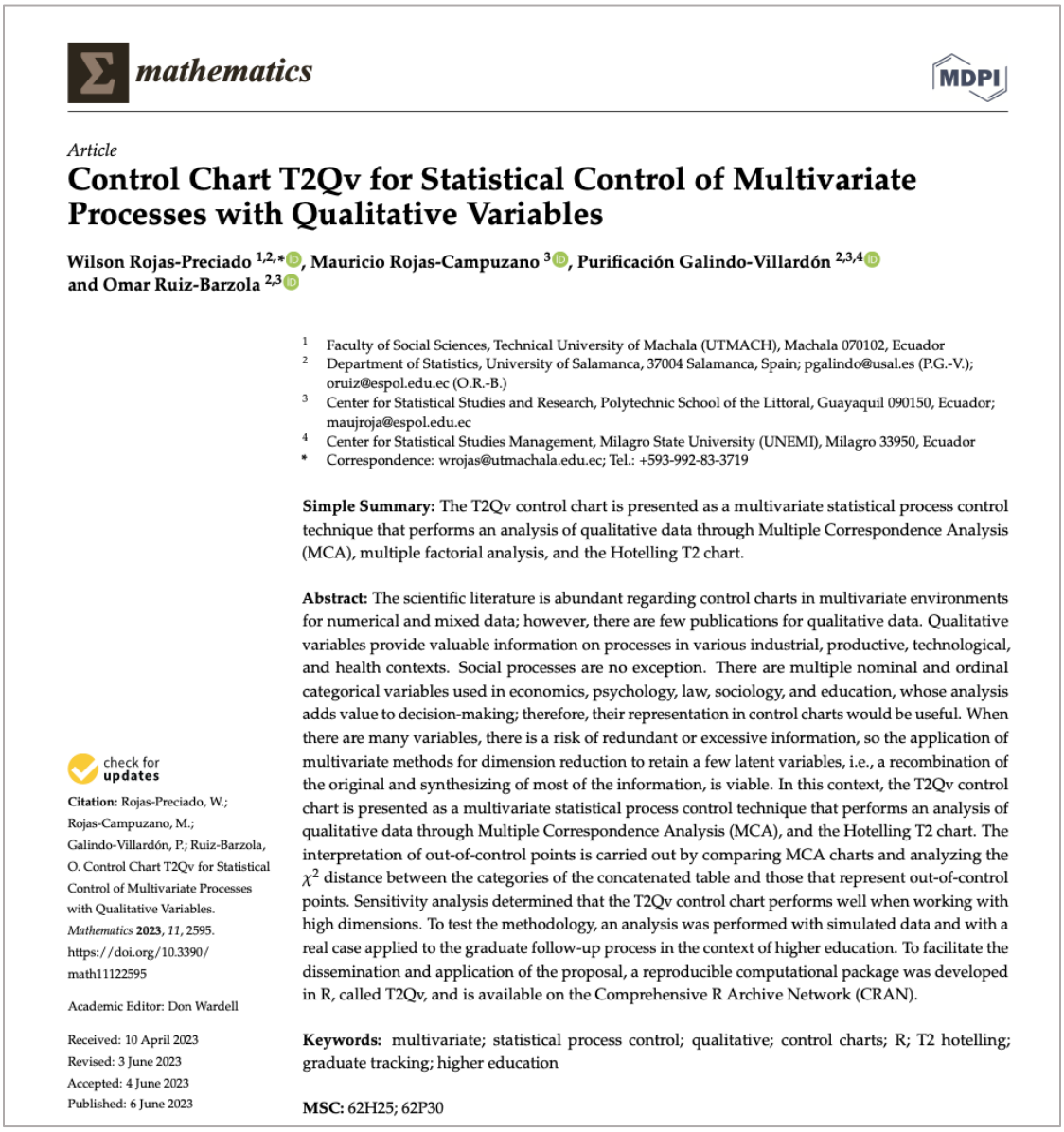
Las tablas resultantes de la generalización se someten a un proceso de normalización, teniendo en cuenta valores propios específicos. Todas estas tablas normalizadas se unifican en una "Matriz Concatenada", que integra toda la información.

Para la monitorización y el control del proceso, se propone el gráfico de control T2Qv, que está basado en distancias de Mahalanobis ponderadas. Un aspecto clave es que, en lugar de utilizar vectores de media tradicionales, el nuevo estadístico T^2_{med} adopta un enfoque de robustez utilizando vectores de medianas, ya que las medianas son menos susceptibles a valores atípicos.

Si el proceso se desvía del control, el gráfico T2Qv puede identificar esta anomalía, pero no las causas específicas detrás de ella. Por lo tanto, es esencial un análisis adicional, que implica comparar la ubicación de puntos que representan categorías de variables en el MCA de la tabla concatenada y en las tablas identificadas como fuera de control. Se utilizan técnicas adicionales, como la distancia chi cuadrado entre las masas de columnas de las tablas comparadas y se complementa con gráficos circulares para representar las distribuciones de categorías y así cuantificar mejor las anomalías.

El resultado del análisis de sensibilidad determinó que el gráfico T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones, pero, pierde estabilidad a bajas dimensiones.

Un resumen de estas contribuciones fue presentado en el ***II Congreso Internacional de Estadística Aplicada***, evento científico organizado por con la Facultad de Recursos Naturales de la Escuela Politécnica de Chimborazo (ESPOCH), que se realizó los días 19, 20 y 21 de abril de 2023, con el trabajo titulado "*T2Qv, aplicación para el control estadístico de procesos multivariantes con variables cualitativas*". Además, respecto de este tema fue publicado un artículo científico en la revista *Mathematics* (JCR 2022 category rank Q1: Mathematics-SCIE; Current impact factor 2.4; Scopus CiteScore 3.5). La Figura 2.11 muestra la primera página del artículo "Control Chart T2Qv for Statistical Control of Multivariate Processes with Qualitative Variables".



mathematics

MDPI

Article

Control Chart T2Qv for Statistical Control of Multivariate Processes with Qualitative Variables

Wilson Rojas-Preciado [1,2,*], Mauricio Rojas-Campuzano [3], Purificación Galindo-Villardón [2,3,4] and Omar Ruiz-Barzola [2,3]

[1] Faculty of Social Sciences, Technical University of Machala (UTMACH), Machala 070102, Ecuador
[2] Department of Statistics, University of Salamanca, 37004 Salamanca, Spain; pgalindo@usal.es (P.G.-V.); oruiz@espol.edu.ec (O.R.-B.)
[3] Center for Statistical Studies and Research, Polytechnic School of the Littoral, Guayaquil 090150, Ecuador; maujroja@espol.edu.ec
[4] Center for Statistical Studies Management, Milagro State University (UNEMI), Milagro 33950, Ecuador
* Correspondence: wrojas@utmachala.edu.ec; Tel.: +593-992-83-3719

Simple Summary: The T2Qv control chart is presented as a multivariate statistical process control technique that performs an analysis of qualitative data through Multiple Correspondence Analysis (MCA), multiple factorial analysis, and the Hotelling T2 chart.

Abstract: The scientific literature is abundant regarding control charts in multivariate environments for numerical and mixed data; however, there are few publications for qualitative data. Qualitative variables provide valuable information on processes in various industrial, productive, technological, and health contexts. Social processes are no exception. There are multiple nominal and ordinal categorical variables used in economics, psychology, law, sociology, and education, whose analysis adds value to decision-making; therefore, their representation in control charts would be useful. When there are many variables, there is a risk of redundant or excessive information, so the application of multivariate methods for dimension reduction to retain a few latent variables, i.e., a recombination of the original and synthesizing of most of the information, is viable. In this context, the T2Qv control chart is presented as a multivariate statistical process control technique that performs an analysis of qualitative data through Multiple Correspondence Analysis (MCA), and the Hotelling T2 chart. The interpretation of out-of-control points is carried out by comparing MCA charts and analyzing the χ^2 distance between the categories of the concatenated table and those that represent out-of-control points. Sensitivity analysis determined that the T2Qv control chart performs well when working with high dimensions. To test the methodology, an analysis was performed with simulated data and with a real case applied to the graduate follow-up process in the context of higher education. To facilitate the dissemination and application of the proposal, a reproducible computational package was developed in R, called T2Qv, and is available on the Comprehensive R Archive Network (CRAN).

Keywords: multivariate; statistical process control; qualitative; control charts; R; T2 hotelling; graduate tracking; higher education

MSC: 62H25; 62P30

check for updates

Citation: Rojas-Preciado, W.; Rojas-Campuzano, M.; Galindo-Villardón, P.; Ruiz-Barzola, O. Control Chart T2Qv for Statistical Control of Multivariate Processes with Qualitative Variables. *Mathematics* **2023**, *11*, 2595. https://doi.org/10.3390/math11122595

Academic Editor: Don Wardell

Received: 10 April 2023
Revised: 3 June 2023
Accepted: 4 June 2023
Published: 6 June 2023

Figura 2.11. Publicación en la revista Mathematics - JCR 2022 category rank Q1: Mathematics-SCIE; Current impact factor 2.4; Scopus CiteScore 3.5

Rojas-Preciado, W., M. Rojas-Campuzano, P. Galindo-Villardón, y O. Ruiz-Barzola. 2023. «Control Chart T2Qv for Statistical Control of Multivariate Processes with Qualitative Variables». *Mathematics* 11 (12). https://doi.org/10.3390/math11122595.

Capítulo 3

Complemento computacional T2Qv

3.1. Introducción

En este apartado se presenta el complemento computacional T2Qv (Rojas-Preciado et al., 2022), paquete estadístico que, para facilitar la difusión y aplicación del método propuesto, se ha desarrollado en lenguaje R Core Team (2023). El nombre de este paquete, T2Qv, es una referencia al gráfico de control multivariante T^2 de Hotelling aplicable para variables cualitativas.

En este capítulo se incluye una descripción del paquete, que explica cómo funciona el T2Qv; se presenta cada una de las funciones con la respectiva documentación requerida por el CRAN y los resultados que se generan; finalmente, se registran las contribuciones de este capítulo, referidas a la disponibilidad del paquete. Para facilitar la ejemplificación de las funciones del paquete T2Qv, se ha incluido un conjunto de datos simulados, denominado *Datak10Contaminated*, que contiene 10 tablas de datos cualitativos. Una descripción detallada de esta base de datos se presenta en el apartado 2.3.

3.2. Descripción del paquete T2Qv

El paquete estadístico T2Qv realiza Análisis de Correspondencias Múltiples a las tablas originales ($\boldsymbol{T_k}$), generando matrices de variables latentes ($\boldsymbol{C_k}$) cuyas coordenadas se someten a un proceso de normalización, multiplicándolas por $\frac{1}{\lambda_1^k}$. Las matrices de coordenadas normalizadas ($\mathbf{T}_k'$) se ordenan una debajo de otra, para conformar una tabla

concatenada ($\mathbb{C}'$), de la que se extrae su vector de medianas $\tilde{\tilde{x}}_{\mathbb{C}'}$ así como los vectores de medianas de cada matriz $\tilde{x}_{\mathbf{C}_k'}$ que la conforman.

Con estos vectores se obtienen los estadísticos $T^2_{med} = n(\tilde{x}_k - \tilde{x}_0)'\mathbf{\Sigma_0^{-1}}(\tilde{x}_k - \tilde{x}_0)$ para cada una de las tablas analizadas, los que se representan como puntos en el gráfico de control T2Qv. Los puntos que en el gráfico se ubiquen fuera del límite ($UCL = \chi^2_{\alpha,p}$) son reportados fuera de control.

El paquete estadístico T2Qv permite la interpretación del comportamiento anómalo de los puntos fuera de control a través de la comparación de los gráficos MCA de una tabla *TC*$_k$, que resulta de concatenar las matrices iniciales, y cada tabla inicial $\boldsymbol{T_k'}$. El paquete permite la selección de las $\boldsymbol{T_k'}$ tablas, de manera que el investigador pueda enfocar su análisis en las que se identifiquen como fuera de control.

Además, el paquete T2Qv genera un gráfico interactivo de barras que representa las distancias χ^2 entre las masas de columnas de las variables de la tabla $\boldsymbol{TC_k}$ y la tabla $\boldsymbol{T_k'}$. Las barras que denotan mayor altura identifican a las variables que están provocando, con mayor fuerza, la salida de control de la *k*-ésima tabla. Este gráfico interactivo incluye, mediante un gráfico circular anidado, una representación de la distribución de las categorías de la variable observada, correspondiente a la *k*-ésima tabla, así como un gráfico circular de la distribución de las categorías de la tabla concatenada ($\boldsymbol{TC_k}$), lo que facilita la identificación de los cambios en la distribución de las categorías.

Los gráficos se pueden mostrar de forma plana o interactiva, de la misma manera todas las salidas se pueden mostrar en un panel interactivo de Shiny y sus resultados gráficos y numéricos pueden ser exportados.

De esta manera el paquete T2Qv consolida la metodología propuesta en esta investigación y permite la explicación de cuándo y por qué el proceso salió de control.

3.3. Funciones y documentación del paquete T2Qv

El paquete T2Qv contiene 6 funciones, las cuales se describen en la Tabla 3.1.

Tabla 3.1. Funciones del paquete T2Qv

Función	Descripción
T2 qualitative	Gráfico de control multivariante T^2 de Hotelling aplicable a variables cualitativas.
MCAconcatenated	Análisis de correspondencias múltiples aplicado a una tabla concatenada.
MCApoint	Análisis de correspondencias múltiples aplicado a una tabla específica.
ChiSq variable	Contiene la distancia Chi-cuadrado entre las masas de columna de la tabla especificada en PointTable y la tabla concatenada. Permite la identificación de qué modo es responsable de la anomalía en la tabla en la que se encuentra.
Datak10Contaminated	Conjunto de datos simulados de 10 tablas
Full Panel	Un panel Shiny completo con el gráfico de control multivariante para variables cualitativas, los dos gráficos MCA y la tabla de distancia de modalidad. Dentro del tablero, se pueden modificar argumentos como el error de tipo I y la dimensionalidad

La documentación del paquete, como lo exige el CRAN, incluye la descripción de las funciones, la definición de los argumentos, las referencias y un ejemplo que se ejecuta con una base de datos simulados que se ha precargado (`Datak10Contaminated`). Se puede acceder a esta documentación utilizando el comando "`help( )`" o "`?`". A continuación se caracteriza cada una de las funciones del paquete T2Qv.

Tabla 3.2. Función *T2_qualitative* {T2Qv}

T2_qualitative {T2Qv}	
Descripción	
Gráfico de control multivariante T2 de Hotelling aplicable para variables cualitativas.	
Uso	
`T2_qualitative (base, IndK, dim, interactive = TRUE, alpha = 0.0027)`	
Argumentos	
`base`	Set de datos
`IndK`	Carácter con el nombre de la columna que especifica la partición del conjunto de datos en *k* tablas.
`dim`	Dimensión tomada para la reducción. La Dimensión inicial recomendada es - 1.
`interactive`	Si es VERDADERO, el gráfico se mostrará de forma interactiva. Si es FALSO, el gráfico se muestra plano. FALSO es el valor predeterminado.
`alpha`	Error tipo I, se recomienda alcanzar este valor utilizando el ARL.
Valor	
Un gráfico de control elaborado con el estadístico Hotelling T2, aplicado para detectar anomalías en cualquiera de las K tablas obtenidas con la especificación de IndK. El límite de control del gráfico se obtiene a partir del número de dimensiones dim y del error alfa tipo I.	
Ejemplos	
`data(Datak10Contaminated)`	
`T2_qualitative(Datak10Contaminated,"GroupLetter",9, TRUE,0.0027)`	

El resultado de esta función (tabla 3.2) es el gráfico intercativo T2Qv (Figura 3.1).

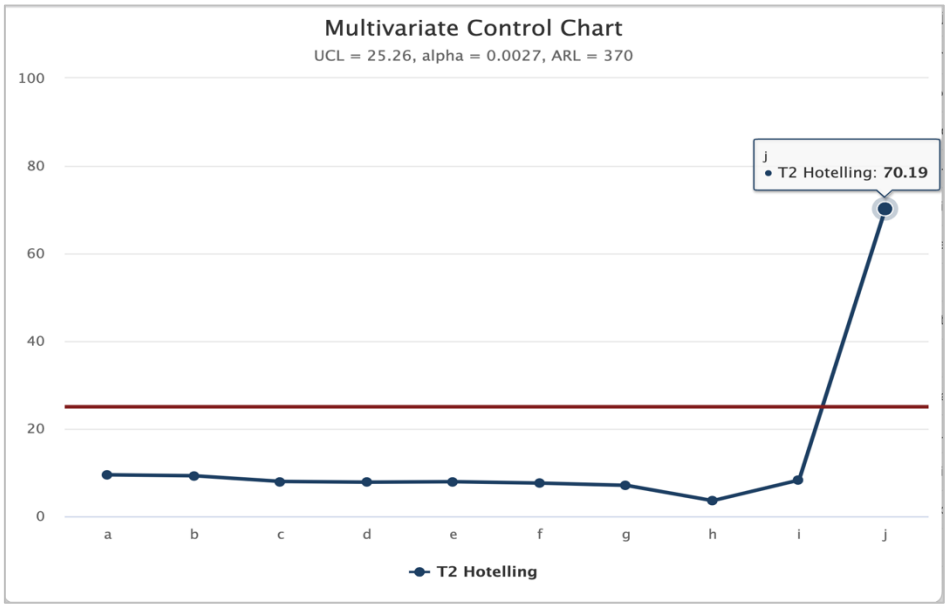


Figura 3.1. Gráfico de control T2Qv

Tabla 3.3. Función *ACMconcatenated* {T2Qv}

ACMconcatenated {T2Qv}	
Descripción	
Análisis de correspondencias múltiples aplicado a una tabla concatenada.	
Uso	
`ACMconcatenated(base, IndK, interactive = TRUE)`	
Argumentos	
base	Set de datos
IndK	Carácter con el nombre de la columna que especifica la partición del conjunto de datos en *k* tablas.
interactive	Si es VERDADERO, el gráfico se mostrará de forma interactiva. Si es FALSO, el gráfico se muestra plano. FALSO es el valor predeterminado.
Valor	
Un gráfico de Análisis de correspondencias múltiples de la tabla concatenada.	
Ejemplos	
`data(Datak10Contaminated)`	
`ACMconcatenated(Datak10Contaminated,"GroupLetter", interactive = TRUE)`	

El resultado de esta función (Tabla 3.3) es un gráfico del Análisis de Correspondencias Múltiples de la tabla Concatenada, como se muestra en la Figura 3.2.

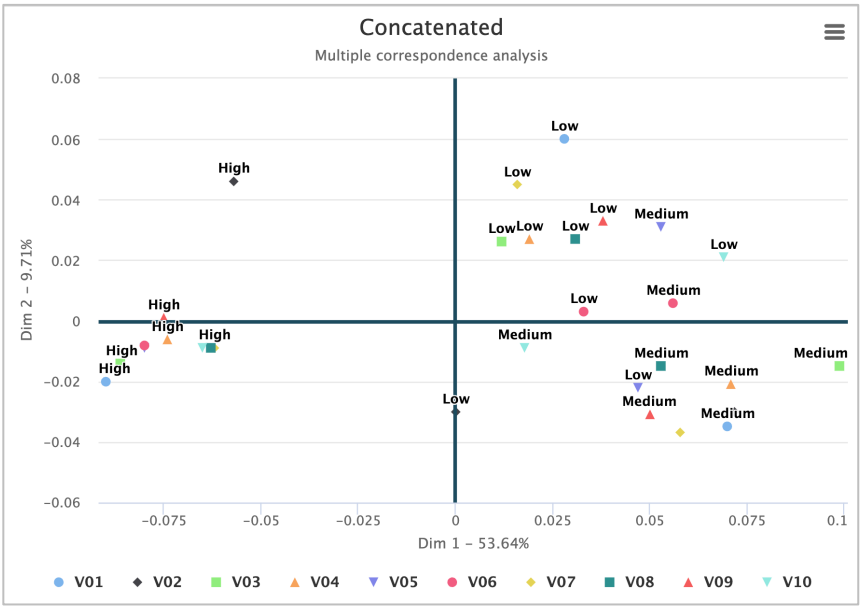


Figura 3.2. Gráfico MCA de la tabla Concatenada del paquete T2Qv

Tabla 3.4. Función *ACMpoint* {T2Qv}

ACMpoint {T2Qv}	
Descripción	
Análisis de correspondencias múltiples aplicado a una tabla específica.	
Uso	
`ACMpoint(base, IndK, PointTable, interactive = TRUE)`	
Argumentos	
base	Set de datos
IndK	Carácter con el nombre de la columna que especifica la partición del conjunto de datos en *k* tablas.
PointTable	Indicador de mesa. Un carácter o número que forma parte de los registros del IndK. Este argumento especifica la tabla en la que se realizará el análisis.
interactive	Si es VERDADERO, el gráfico se mostrará de forma interactiva. Si es FALSO, el gráfico se muestra plano. FALSO es el valor predeterminado.
Valor	
Un gráfico de Análisis de correspondencias múltiples de la tabla especificada en la Tabla Punto.	
Ejemplos	
`data(Datak10Contaminated)`	
`ACMpoint(Datak10Contaminated,"GroupLetter", PointTable="j", interactive=TRUE)`	

El resultado de esta función (Tabla 3.4) es un gráfico interactivo del MCA de la tabla Punto (Figura 3.3), reportada como fuera de control por el gráfico T2Qv (Figura 3.1).

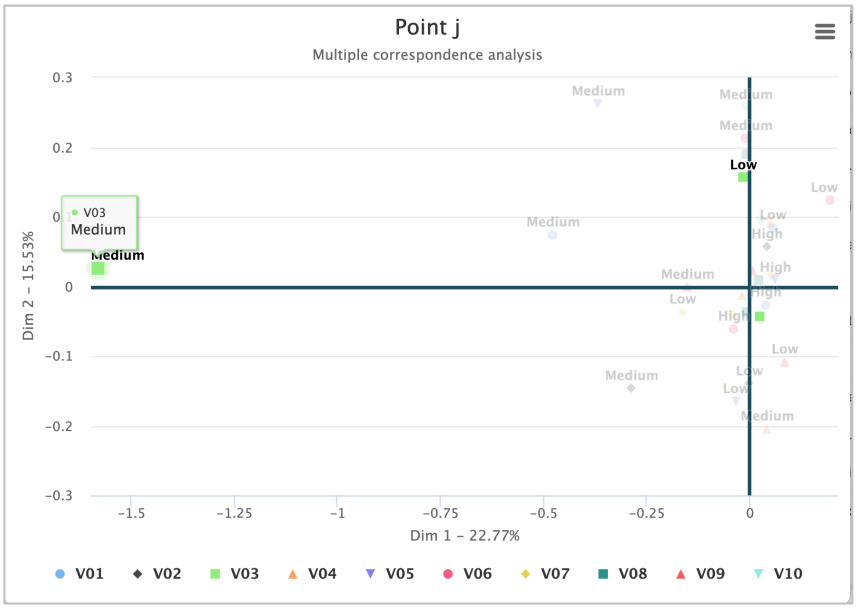


Figura 3.3. Gráfico MCA de la tabla Punto del paquete T2Qv

Tabla 3.5. Función *ChiSq_variable* {T2Qv}

ChiSq_variable {T2Qv}	
Descripción	
Contiene la distancia Chi cuadrado entre las masas de las columnas de la tabla especificada en la tabla Punto y la tabla Concatenada. Permite identificar qué modo es el responsable de la anomalía en la tabla en la que se encuentra.	
Uso	
`ChiSq_variable(base, IndK, PointTable, interactive = TRUE, ylim = 0.09)`	
Argumentos	
base	Set de datos
IndK	Carácter con el nombre de la columna que especifica la partición del conjunto de datos en k tablas.
PointTable	Indicador de tabla. Un carácter o número que forma parte de los registros del *IndK*. Este argumento especifica la tabla en la que se realizará el análisis.
interactive	Si es VERDADERO, el gráfico se mostrará de forma interactiva. Si es FALSO, el gráfico se muestra plano. FALSO es el valor predeterminado.
ylim	Límite del eje *Y*.
Valor	
Una tabla con distancias Chi cuadrado entre las masas de las columnas de la tabla especificada en PointTable y la tabla concatenada.	
Ejemplos	
`data(Datak10Contaminated)`	
`ChiSq_variable(Datak10Contaminated, "GroupLetter", PointTable="j", ylim=5, interactive=TRUE)`	

Cuando se selecciona la opción `interactive=FALSE`, esta función (Tabla 3.5) genera, como resultado, una tabla con las distancias Chi cuadrado entre las masas de las columnas de la tabla especificada en PointTable y la tabla Concatenada (Figura 3.4).

	Variables	ChiSq
	<chr>	<dbl>
1	V01	0.0697
2	V02	0.0429
3	V03	0.0770
4	V04	0.0521
5	V05	0.0439
6	V06	0.0594
7	V07	0.0452
8	V08	0.0305
9	V09	0.0544
10	V10	0.0584

Figura 3.4. Distancias Chi cuadrado entre las masas de las columnas de la tabla Punto y la tabla Concatenada, del paquete T2Qv

Pero, si se selecciona la opción `interactive=TRUE`, la función *ChiSq_variable* genera un gráfico de barras interactivo de las distancias Chi cuadrado. Las distancias se expresan, en este gráfico, mediante la altura de las barras, mientras mayores son las distancias Chi cuadrado, más altas son las barras. Además, al pasar el cursor sobre cualquiera de las barras se muestra un gráfico de sectores que representa la distribución de las categorías de las variables correspondientes a la tabla Punto. En la Figura 3.5 se observa la distribución de las categorías (High: 0.7, Medium: 0.1, Low: 0.2) de la variable V05.

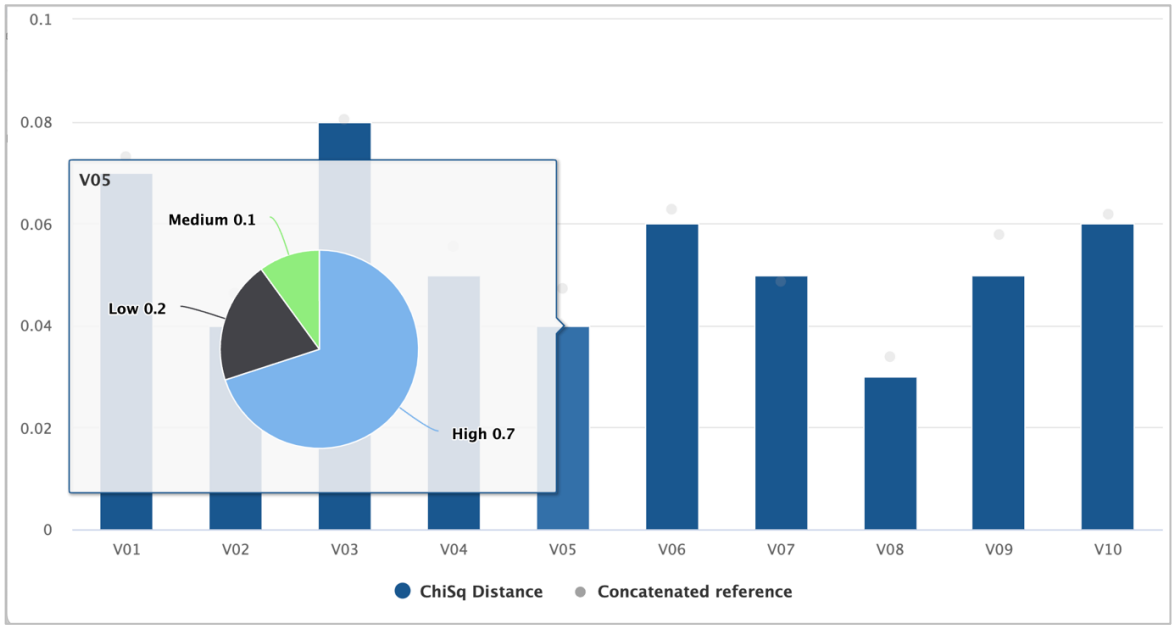


Figura 3.5. Gráfico de barras de las distancias Chi cuadrado entre las masas de las columnas de la tabla Punto y la tabla Concatenada, paquete T2Qv.

Si se pasa el cursor por alguno de los puntos de color gris que aparecen sobre las barras, se muestra un gráfico de sectores de las categorías de las variables correspondientes a la tabla Concatenada. La Figura 3.6 permite observar la distribución de las categorías (High: 0.38, Medium: 0.32, Low: 0.3) de la variable V05.

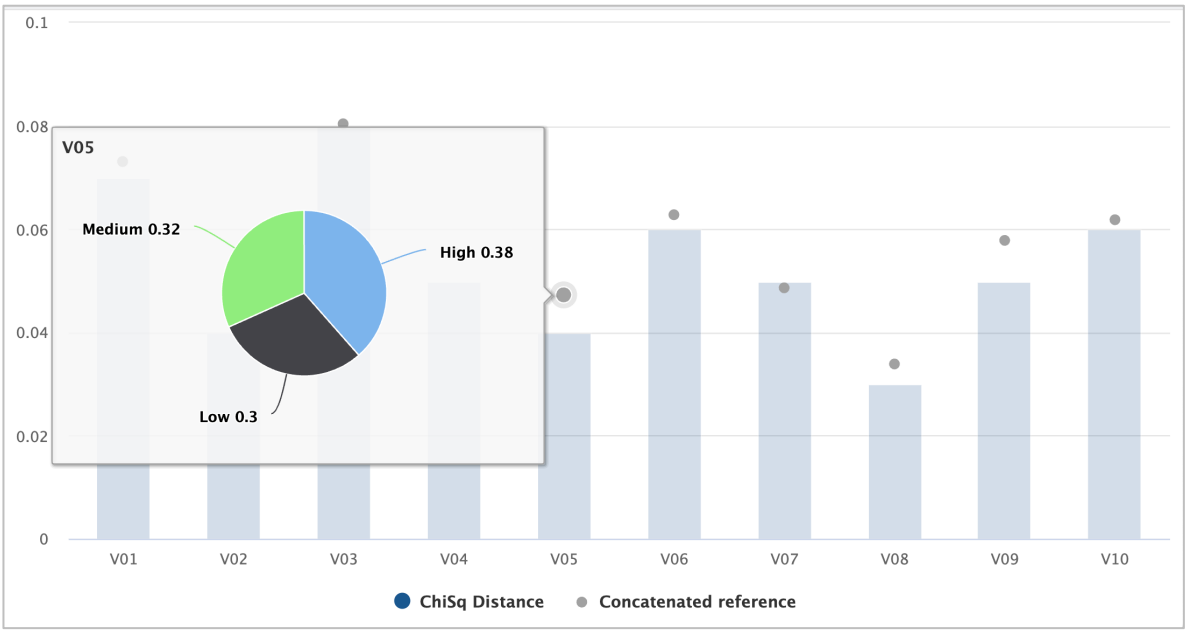


Figura 3.6. Distribución de la variable V05 de la tabla Concatenada, del paquete T2Qv.

Tabla 3.6. Función Datak10Contaminated {T2Qv}

Datak10Contaminated {T2Qv}	
Descripción	
Datos de 10 tablas con 10 variables categóricas, los datos de la tabla 10 se generaron con una distribución diferente a las demás.	
Uso	
`Datak10Contaminated`	
Formato	
Un marco de datos:	
V01	Contiene 3 modos "High", "Medium", "Low".
V02	Contiene 3 modos "High", "Medium", "Low".
V03	Contiene 3 modos "High", "Medium", "Low".
V04	Contiene 3 modos "High", "Medium", "Low".
V05	Contiene 3 modos "High", "Medium", "Low".
V06	Contiene 3 modos "High", "Medium", "Low".
V07	Contiene 3 modos "High", "Medium", "Low".
V08	Contiene 3 modos "High", "Medium", "Low".
V09	Contiene 3 modos "High", "Medium", "Low".
V10	Contiene 3 modos "High", "Medium", "Low".
`GroupLetter`	Las letras de la "*a*" a la "*j*" identifican las *k* tablas.

La función registrada en la Tabla 3.6 devuelve la base de datos *Datak10Contaminated,* que se describe en el apartado 3.2 de este documento y está precargada en el paquete T2Qv.

Tabla 3.7. Función *Full_Panel* {T2Qv}

Full_Panel {T2Qv}	
Descripción	
Un panel de Shiny completo con el gráfico de control multivariante para variables cualitativas, los dos gráficos ACM y la tabla de modalidades de distancia. Dentro del tablero, se pueden modificar argumentos como el error tipo I y la dimensionalidad.	
Uso	
`Full_Panel(base, IndK)`	
Argumentos	
base	Set de datos
IndK	Caracter con el nombre de la columna que especifica la partición del conjunto de datos en *k* tablas.
Valor	
Un panel completo con la gráfica de control multivariante para variables cualitativas, las dos gráficas ACM y la tabla de modalidades de distancia.	
Ejemplos	
`data(Datak10Contaminated)`	
`Full_Panel(Datak10Contaminated, "GroupLetter")`	

La función *Full Panel* (Tabla 3.7) genera una vista que presenta, de forma integrada e interactiva los resultados de las otras funciones, es decir, el gráfico T2Qv; los gráficos de ACM de la tabla concatenada y la tabla punto, uno al lado del otro, para favorecer el ejercicio de comparación; el gráfico de barras de las distancias Chi cuadrado; y el gráfico circular, que representa la distribución de las categorías de las variables analizadas en ambas tablas comparadas (Figura 3.7).

La función *Full panel* permite seleccionar a cualquiera de las tablas de la base de datos para su análisis. El usuario debe seleccionar una tabla y el aplicativo realiza un MCA

de la tabla Punto que puede ser comparado con el gráfico de MCA de la tabla Concatenada. Luego, se puede interpretar el comportamiento de las variables por la ubicación de las categorías en el plano de ambos gráficos.

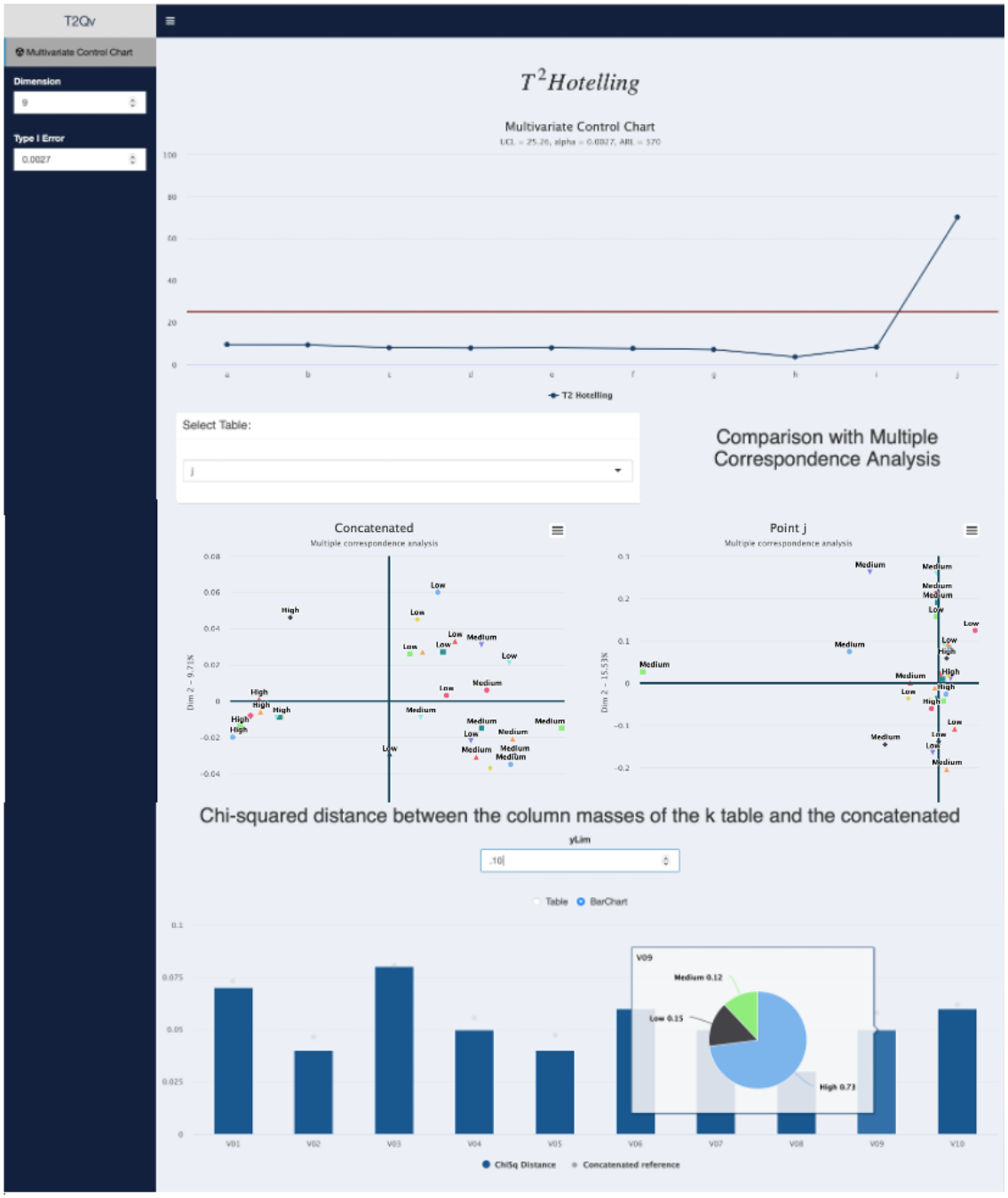


Figura 3.7. Vista de la pantalla de la función Full Panel del aplicativo T2Qv

3.4. Contribuciones realizadas en el Capítulo 3

El paquete T2Qv, en su versión 0.2.0, está disponible en el repositorio oficial de R, The Comprehensive R Archive Network (CRAN), en el enlace https://cran.r-project.org/package=T2Qv. La descarga se la puede realizar utilizando el comando `install.packages("T2Qv")`. Después de instalado, el paquete se carga mediante el comando `library(T2Qv)`. La descripción técnica del paquete, registrada en el CRAN, se enuncia en la Figura 3.8.

```
Package: T2Qv
Type: Package
Title: Control Qualitative Variables
Version: 0.2.0
Authors@R: c(person("Wilson", "Rojas-Preciado", role = c("aut", "cre"),
                    email = "wrojas@utmachala.edu.ec"),
             person("Mauricio", "Rojas-Campuzano", role = c("aut","ctb"),
                    email="maujroja@espol.edu.ec"),
             person("Purificación", "Galindo-Villardón", role =
c("aut","ctb"),
                    email = "oruiz@espol.edu.ec"),
             person("Omar", "Ruiz-Barzola", role = c("aut","ctb"),
                    email = "oruiz@espol.edu.ec"))
Maintainer: Wilson Rojas-Preciado <wrojas@utmachala.edu.ec>
Description: Covers k-table control analysis using multivariate control charts
for qualitative variables using fundamentals of multiple correspondence
analysis and multiple factor analysis. The graphs can be shown in a flat or
interactive way, in the same way all the outputs can be shown in an interactive
shiny panel.
License: MIT + file LICENSE
Encoding: UTF-8
LazyData: true
RoxygenNote: 7.1.1
Depends: R (>= 3.5)
Imports: shiny, shinydashboardPlus, shinydashboard, shinycssloaders,
        dplyr, ca, highcharter, stringr, tables, purrr, tidyr,
        htmltools (>= 0.5.1.1)
Suggests: testthat (>= 3.0.0)
Config/testthat/edition: 3
NeedsCompilation: no
Packaged: 2023-10-11 16:02:22 UTC; javierrojas
Author: Wilson Rojas-Preciado [aut, cre],
  Mauricio Rojas-Campuzano [aut, ctb],
  Purificación Galindo-Villardón [aut, ctb],
  Omar Ruiz-Barzola [aut, ctb]
```

Figura 3.8. Descripción técnica del paquete T2Qv, cargada en el CRAN

Capítulo 4

Aplicación del paquete T2Qv en el contexto de la Educación Superior

4.1. Introducción

En este ejemplo se realiza la aplicación del aplicativo T2Qv, al análisis de los resultados del proceso de Seguimiento a Graduados de la carrera de Ciencias Médicas de la Universidad Técnica de Machala (UTMACH), Ecuador.

El Consejo de Aseguramiento de la Calidad de la Educación Superior del Ecuador (CACES, 2023) define al seguimiento a graduados como el proceso a través del cual las universidades y escuelas politécnicas gestionan la información referente a la empleabilidad pertinente, campos ocupacionales y los niveles de satisfacción de los graduados de las carreras o programas ofertados. Estos procesos permiten orientar la toma de decisiones institucionales en torno a la oferta académica, gestión del currículo, los niveles de inserción laboral de sus graduados y fortalecer los procesos de aseguramiento interno de la calidad y de mejora continua.

4.2. Base de datos *CMSG*

Para este fin se utiliza una base de datos, denominada *CMSG*, tomada de reportes que están disponibles para autoridades de la universidad en su Sistema Informático SIUTMACH. La base de datos *CMSG* contiene 166 observaciones y 16 variables cualitativas tomadas de los resultados de una encuesta de seguimiento a graduados de la carrera de Ciencias Médicas, desde 2017 hasta 2021. Los datos se organizaron en cuatro tablas que corresponden a cuatro

periodos en los que aplicaron las encuestas: 2021, 2020, 2019, 2018-2017. Este último periodo agrupa los datos de seguimiento a graduados de los dos primeros años (Apéndice A).

Hay otros resultados de seguimiento a graduados de la carrera de Ciencias Médicas de la UTMACH, pero, corresponden a periodos anteriores a 2017 y fueron tomados con otra encuesta, en consecuencia, se trata de otras variables, por eso sus resultados no constan en este análisis.

Las variables registradas en la base de datos *CMSG,* con sus respectivas categorías, son las siguientes:

- **Año_graduación.-** Esta variable sirve como clasificador de tablas de la base de datos, divide al conjunto de datos en grupos que corresponden a los 4 periodos de estudio: *2017-2018, 2029, 2020 y 2021.*
- **Relación trabajo y Perfil profesional.-** Es una variable categórica ordinal que expresa los niveles de asociación entre las actividades laborales más relevantes que desempeñan los médicos graduados en la UTMACH y el Perfil profesional, entendido como el conjunto de características esenciales referidas a conocimientos, habilidades y valores que debe tener un médico para su desempeño en el país. Estos niveles son: *Muy alta, Alta, Media, Medio baja y Baja.*
- **Satisfacción con conocimientos y habilidades adquiridas.-** Mide el grado de satisfacción que tienen los graduados respecto de los conocimientos y habilidades que adquirieron durante su proceso de formación profesional. Sus niveles son: *Alta; Medio alta; Media; Medio baja y Baja.*

- **Satisfacción Malla curricular.-** Valora el nivel de satisfacción que tienen los graduados respecto de la malla curricular que estudiaron en su proceso académico. Sus niveles son: *Muy satisfactorio, Satisfactorio, Poco satisfactorio e Insatisfactorio.*
- **Satisfacción Estrategias Aprendizaje.-** Evalúa el nivel de satisfacción que tienen los graduados respecto de las estrategias de enseñanza aprendizaje que se aplicaron en los procesos académicos durante su periodo formativo. Sus niveles son: *Muy satisfactorio, Satisfactorio, Poco satisfactorio e Insatisfactorio.*
- **Satisfacción Investigación Formativa.-** Evalúa el nivel de satisfacción que tienen los graduados respecto de los procesos de investigación realizados para promover y enriquecer el aprendizaje y la formación de los estudiantes. Sus niveles son: *Muy satisfactorio, Satisfactorio, Poco satisfactorio e Insatisfactorio.*
- **Satisfacción aplicación de investigación - Vinculación.-** Mide el nivel de satisfacción que tienen los graduados respecto de la aplicación de procesos de investigación en programas y proyectos de vinculación con la sociedad durante su etapa de formación profesional. Sus niveles son: *Muy satisfactorio, Satisfactorio, Poco satisfactorio e Insatisfactorio.*
- **Satisfacción difusión resultados investigación.-** Mide el nivel de satisfacción que tienen los graduados respecto de la divulgación y comunicación de los hallazgos, resultados y conclusiones de procesos de investigación con la comunidad académica y científica, así como con el público en general, durante su etapa de formación profesional. Sus niveles son: *Muy satisfactorio, Satisfactorio, Poco satisfactorio e Insatisfactorio.*

- **Tipo de postgrado que estudia.-** Esta variable categórica nominal identifica los graduados que están cursando programas de postgrado. Las categorías de esta variable son: *Ninguno, Diplomado, Maestría, Especialización.*
- **Interés por especialidad.-** Esta variable complementa a la anterior, determina el interés que tienen los graduados por un área específica para cursar en programas de postgrado. Sus clases son: *Cirugía, Clínica, Educación Médica Superior, Gineco-Obstetricia, Pediatría y Salud comunitaria*.
- **Estrategia para conseguir empleo**.- Esta variable se dirige al análisis de demanda ocupacional y los campos de actuación profesional de los graduados, para facilitar la visión clara y actualizada de las perspectivas de empleo y las posibilidades de desarrollo profesional en el campo de la salud. Las opciones de respuesta son: *Bolsa de empleo UTMACH, Prensa_Radio_TV, Preferencias personales, Red Socioempleo* y *Redes sociales*.
- **Situación laboral.-** Esta variable determina en qué está trabajando el graduado. Sus categorías son: *Profesional independiente, Empleado público, Empleado privado, Desempleado*.
- **Nivel jerárquico laboral.-** Posición y responsabilidad que tiene un empleado dentro de la estructura organizativa de una empresa o institución. Las categorías de esta variable son: *Administrativo, Médico (MD) privado, MD residente, MD rural, MD tratante, Otro, Ninguno*.
- **Antigüedad en el trabajo.-** Es el tiempo ha transcurrido desde se incorporó a un trabajo. Las opciones son: *Más de 2 años, 1 a 2 años, 6 meses a 1 año, menos de 6 meses, Ninguno*.

- **Ingreso mensual.-** Ingreso económico percibido en un mes. Las opciones son: *2101 a 2400, 1801 a 2100, 1501 a 1800, 1201 a 1500, 901 a 1200, 601 a 900, 376 a 600, Menos de 375, Ninguno*.
- **Cohorte.-** Grupo de graduados que ingresaron juntos a la carrera en el mismo año académico y son seguidos a lo largo de su trayectoria educativa para analizar su rendimiento, logros o comportamiento: las cohortes consideradas en este estudio son: *Anterior a 2009, 2009, 2010, 2011, 2012, 2013, 2014*.

4.3. Gráfico de control multivariante T2Qv

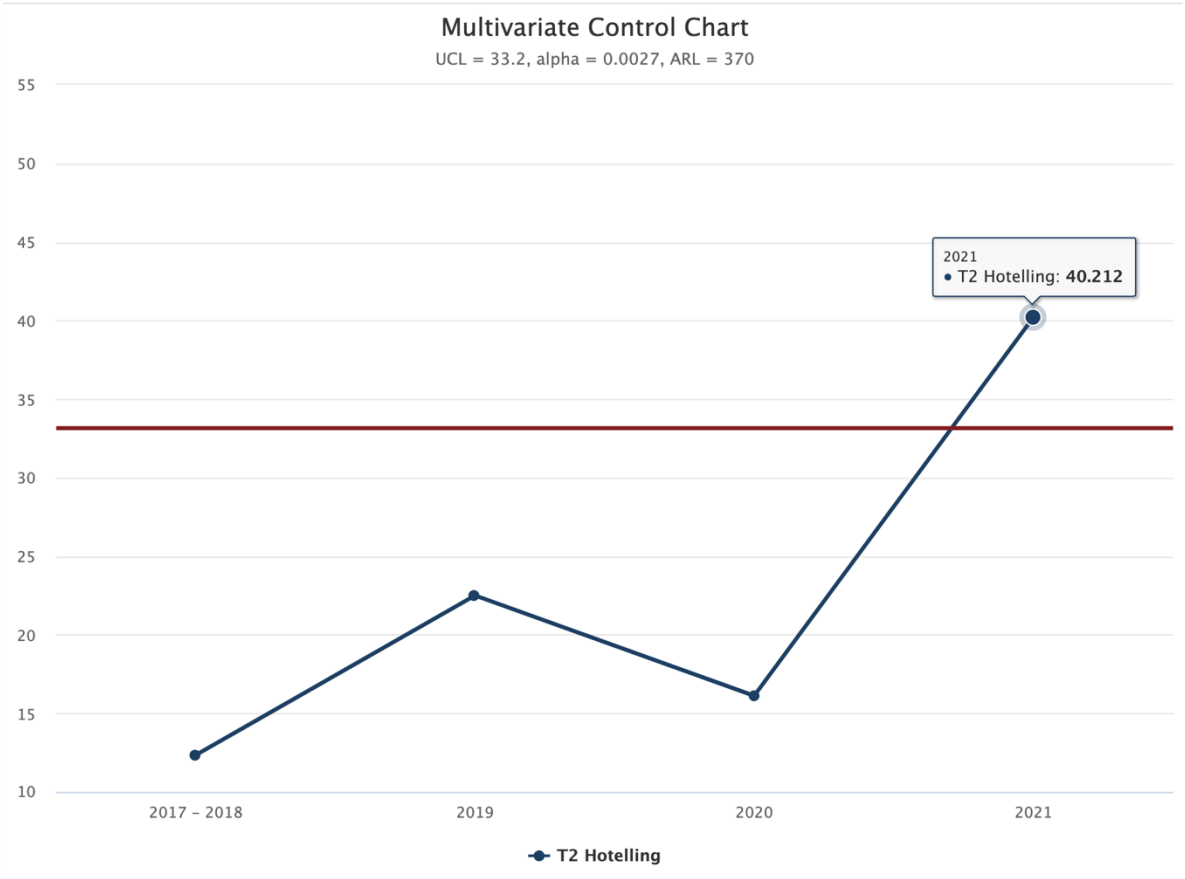


Figura 4.1. Gráfico T2Qv aplicado a la base de datos *CMSG*

La Figura 4.1 muestra el gráfico de control T2Qv para la representación de las $K = 4$ tablas que conforman la base de datos *CMSG.* Cada una de estas tablas se representa mediante puntos en el gráfico y corresponden a los cuatro periodos considerados en este estudio de seguimiento a graduados. Los tres primeros puntos se ubicaron por debajo del límite de control (UCL = 33.2), mientras que el cuarto, que corresponde a la tabla del año 2021, registra un valor mayor que UCL ($T^2_{med} = 40.21$), por lo tanto, se puede decir que la tabla 2021 está llevando al proceso fuera de control. Luego, será necesario un análisis comparativo de esta tabla (tabla punto) versus la concatenada, que sirve de referente, para identificar las causas de la variación y facilitar la toma de decisiones que permitan corregir las desviaciones encontradas. Para ello se utilizan gráficos del MCA de las tablas señaladas.

4.4. Análisis de correspondencias múltiples en la interpretación de los puntos fuera de control

La aplicación de esta técnica estadística multivariante implica una comparación de los gráficos del Análisis de Correspondencias Múltiples (ACM) de la tabla Concatenada y a la tabla Punto, que en este caso corresponde a la tabla del año 2021. El análisis se busca semejanzas y diferencias entre la ubicación de las categorías de las variables de ambas tablas.

4.4.1. Análisis de Correspondencias Múltiples aplicado a la Tabla Concatenada

La Figura 4.2 permite observar el gráfico del Análisis de Correspondencias Múltiples de la Tabla Concatenada. La inercia total es de 44.3%. El gráfico muestra la nube de puntos, que representa la ubicación en el plano de las diferentes categorías de las variables analizadas. El primer eje es el que conserva la mayor varianza explicada y se asocia con la satisfacción de los graduados respecto de los distintos aspectos consultados. A la izquierda se encuentran

las categorías que expresan bajos niveles de satisfacción, mientras que, a la derecha, los de mayor satisfacción. Se observa que se formaron tres grupos que se caracterizan a continuación.

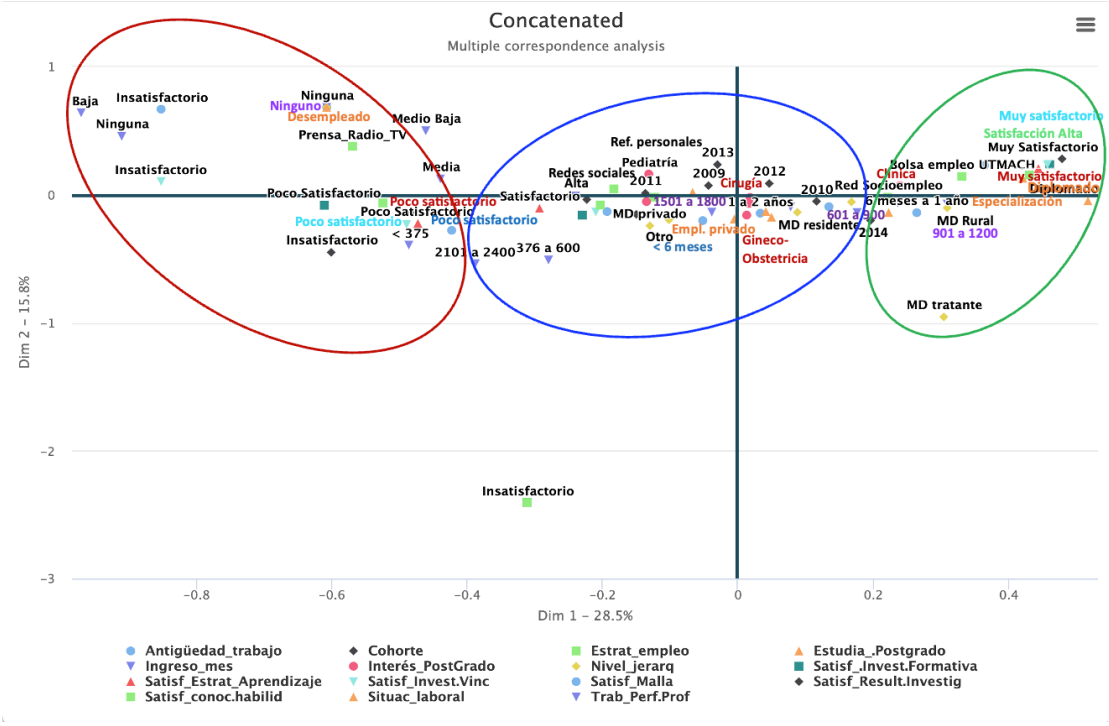


Figura 4.2. Gráfico del MCA de la tabla concatenada, *CMSG*

El primer grupo, a la izquierda del gráfico, contiene los niveles más bajos de satisfacción de los graduados con los conocimientos y las habilidades adquiridas, la malla curricular, las estrategias de aprendizaje aplicadas, la investigación formativa, la aplicación de Investigación a la Vinculación con la sociedad y la difusión de los resultados de investigación. Estos niveles bajos se asocian con escenarios de desempleo, pero, se observa también pocos esfuerzos de los graduados por vincularse a algún trabajo, ellos sólo han esperado obtener información en la prensa, radio y televisión. Los graduados de este grupo

manifiestan que hay escasa o ninguna relación entre trabajo y perfil profesional del médico, así como, desinterés por cursar programas de postgrado.

El segundo grupo, al otro extremo del eje horizontal, muestra una fuerte asociación entre las categorías que expresan la más alta satisfacción entre las variables relacionadas con la formación profesional de los médicos graduados de la UTMACH. Se trata de profesionales que han puesto empeño en dos estrategias para conseguir trabajo: la Bolsa de empleo de la UTMACH y la Red Socioempleo, impulsada el gobierno central. Como resultado, estos profesionales están trabajando, en su mayoría como médicos tratantes o rurales, vinculados al Ministerio de Salud Pública y sus ingresos económicos están entre 901 y 1200 dólares mensuales. Asimismo, estos graduados, principalmente de la cohorte 2014, aseguran que hay alta relación entre trabajo y perfil profesional del médico y tienen interés por cursar programas de postgrado, algunos ya están estudiando programas en especialidades médicas, especialmente en el área clínica.

El tercer grupo asocia las categorías que representan valores medios y medio altos de satisfacción respecto de las variables evaluadas. Aquí está la mayoría de los graduados de las distintas cohortes. Ellos han basado su estrategia para conseguir empleo en el uso de redes sociales y los contactos personales. Como respuesta, están trabajando en la empresa privada y como médicos independientes, médicos residentes y en labores administrativas en casas de salud. Sus ingresos económicos están entre 1201 y 1800 dólares. El interés por cursar programas de postgrado se manifiesta alrededor de áreas de Cirugía, Educación Médica Superior, Gineco-Obstetricia, Pediatría y Salud comunitaria. Aunque, dentro de este grupo de graduados también hay quienes no se animan a estudiar un postgrado, seguramente por

considerar que el empleo en empresas privadas no es tan estable como en instituciones del estado.

4.4.2. Análisis de Correspondencias Múltiples aplicado a la Tabla Punto

La Figura 4.3 muestra el gráfico del MCA de la tabla punto, es decir, la tabla que contiene los datos de la encuesta aplicada a los graduados del año 2021. La inercia total es de 47.26%. Haciendo un análisis comparativo de esta tabla con la Concatenada, se observa que aquí hay una distribución más amplia de los puntos en el plano formado por las dos dimensiones, lo que quiere decir que en la representación de los puntos hay mayor incidencia de otras variables latentes, mientras que, en la tabla Concatenada los puntos se ubicaron prácticamente alrededor del primer eje, que representaba la satisfacción.

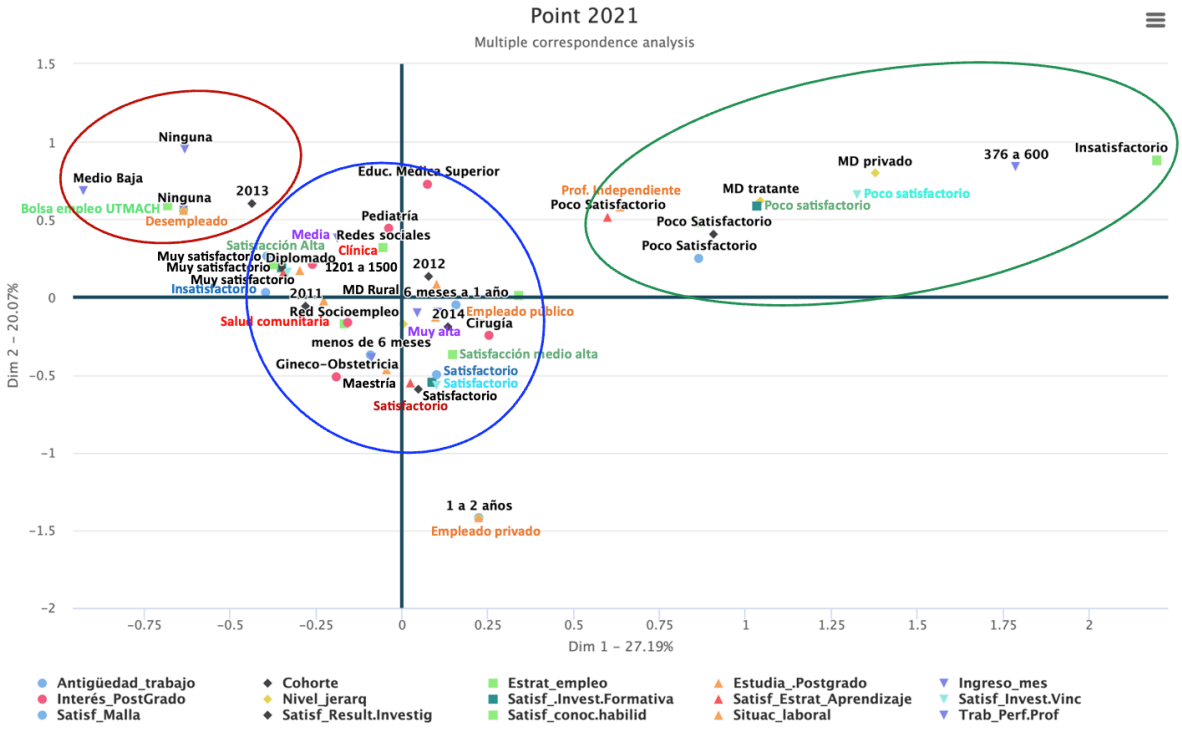


Figura 4.3. Gráfico del MCA de la tabla 2021, *CMSG*

En la Figura 4.3 también se forman tres grupos que presentan una configuración parecida a la de la tabla concatenada, pero hay diferencias en la asociación de distintas categorías de variables que a continuación se describen.

El primer grupo, ubicado en el primer cuadrante del plano, contiene los niveles más bajos de satisfacción de los graduados con casi todas las variables evaluadas. Pero, a diferencia de lo que ocurre en la tabla concatenada, estos niveles bajos no se asocian con escenarios de desempleo. Esto es preocupante. Los graduados de este grupo manifiestan inconformidad con elementos relevantes de su formación y son profesionales que están laborando en su campo del conocimiento, como profesionales independientes o en el rol de médicos tratantes. Esto podría deberse a una percepción de poca estabilidad laboral e ingresos económicos bajos, que no sobrepasan los $600.

Esta inconformidad y percepción de inestabilidad se puede relacionar con su negación o escaso interés por crecer en su formación profesional a nivel de postgrado, por lo menos hasta lograr una situación laboral y posición económica más sólida, quizás cuando se vinculen a instituciones de salud del estado. Quizás en ese momento se pueda despertar, en este grupo, el interés por emprender en la formación de postgrado, pero por el momento, no.

Al otro lado del gráfico, en el segundo cuadrante, se formó el segundo grupo, representado por la cohorte 2013, cuyos graduados perciben una relación Medio baja o Ninguna relación entre el trabajo y el Perfil Profesional del médico. Al parecer, su única estrategia para conseguir trabajo, la Bolsa de empleo UTMACH, no ha dado los resultados esperados, en consecuencia, están desempleados y se dedican a otras actividades alejadas de

su perfil profesional. Es comprensible que en este grupo todavía no se manifieste interés por cursar programas de postgrado.

El tercer grupo, en el centro del gráfico, está representado por los graduados de las cohortes 2011, 2012 y 2014, quienes valoran como alta o muy alta la relación entre el trabajo y perfil profesional del médico. Los graduados de este grupo le apostaron a encontrar trabajo utilizando redes sociales, referencias personales y la Red Socio-empleo, promovida por el Estado, en consecuencia, desde hace un año como máximo, lograron ubicarse como médicos rurales en el Ministerio de Salud, aunque esto es temporal. Estos nuevos profesionales se desempeñan como médicos rurales, es decir, empleados públicos y sus ingresos económicos van desde 601 a 1500 dólares por mes.

Se observa que este grupo manifiesta fuerte asociación entre las categorías que expresan valores medio altos y altos satisfacción con todas las variables evaluadas, pero, llama la atención que en medio de esta nube de puntos se ubicó una categoría que revela insatisfacción de los graduados de la cohorte 2011 con la malla curricular. El interés por la formación de postgrado gira en torno a los campos de la Cirugía, Gineco-Obstetricia, Salud comunitaria, Clínica y Pediatría.

4.4.3. Distancia chi cuadrado entre las entre de columnas de la tabla concatenada y la tabla punto

Las categorías de variables que han cambiado de forma notable su ubicación en las tablas comparadas, así como sus niveles de asociación, pueden estar ocasionando el estado fuera de control del proceso. La identificación de estas variables se facilita cuando se analiza la Figura 4.4.

La Figura 4.4 presenta, en un gráfico de barras, la distancia Chi cuadrado entre las entre las masas de columna de la tabla concatenada y la tabla punto, que en el gráfico T2Qv (Figura 4.1) se mostró con un valor mayor que el límite de control. Mientras más altas son las barras de las variables, mayor es esta distancia. Las variables con mayores distancias Chi cuadrado son las que con mayor fuerza están provocando el desplazamiento de la centralidad del proceso y llevándolo a un estado fuera de control.

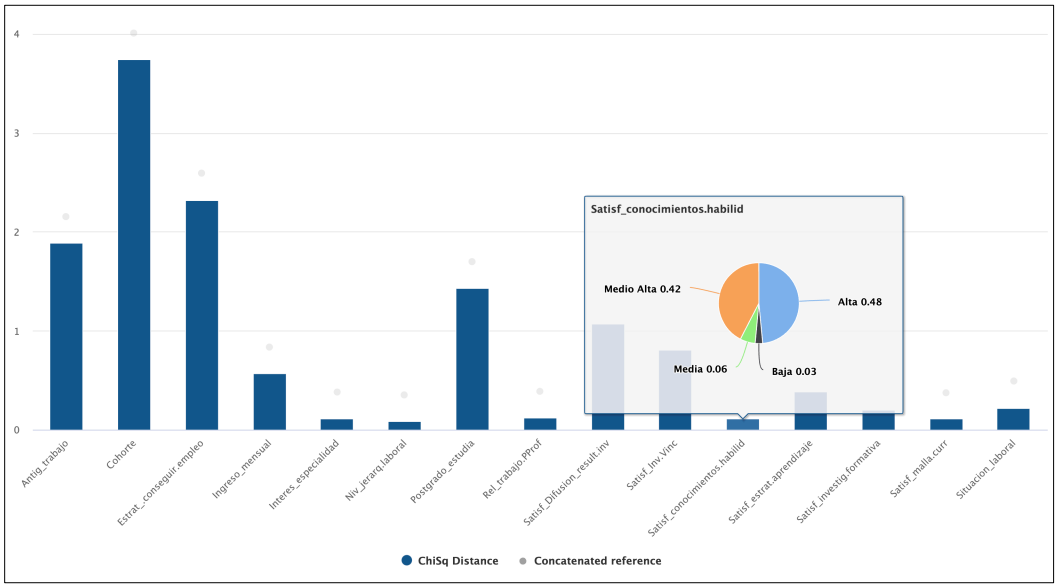


Figura 4.4. Distancia $\chi 2$ entre las masas de la tabla concatenada y la 2021, *CMSG*

Las mayores distancias χ^2 corresponden a las variables Cohorte (3.74), Estrategia para conseguir empleo (2.32) y Antigüedad en el trabajo (1.89). Esto es consistente con los hallazgos encontrados en el análisis comparativo de la tabla concatenada y la tabla punto mediante MCA. Las distancias menores son Nivel jerárquico laboral (0.09), Satisfacción Malla Curricular (0.11) y Satisfacción con conocimientos y habilidades adquiridas (0.11), como se observa en la Tabla 4.1.

Tabla 4.1. Distancia χ^2 entre las masas de la tabla Concatenada y la 2021, *CMSG*.

Variables	ChiSq
Cohorte	3.74188
Estrategia para conseguir empleo	2.32318
Antigüedad en el trabajo	1.88773
Tipo de postgrado que estudia	1.43195
Satisfacción Difusión de resultados de investigación	1.07024
Satisfacción Aplicación de investigación - Vinculación	0.80582
Ingreso mensual	0.56563
Satisfacción Estrategias Aprendizaje	0.38948
Situación laboral	0.22173
Satisfacción Investigación Formativa	0.20029
Relación Trabajo y Perfil Profesional	0.12210
Interés especialidad	0.11470
Satisfacción con conocimientos y habilidades adquiridas	0.11299
Satisfacción Malla curricular	0.10760
Nivel jerárquico laboral	0.08577

4.4.4. Distribución de las categorías de las variables en la tabla concatenada y la tabla punto

Una vez que se ha identificado las variables que con mayor fuerza están provocando el desplazamiento de la centralidad del proceso, se analiza la distribución de las categorías de cada una de estas variables, en las dos tablas comparadas. Para este fin, el aplicativo T2Qv genera gráficos interactivos de sectores, de las variables analizadas, que aparecen al pasar el cursor sobre el gráfico de barras. La Figura 4.5 muestra cómo se han distribuido las diferentes categorías de la variable Cohorte, en la tabla punto (Tabla 2021) y la Tabla Concatenada, y expresa esta distribución en proporciones. La comparación denota importantes diferencias.

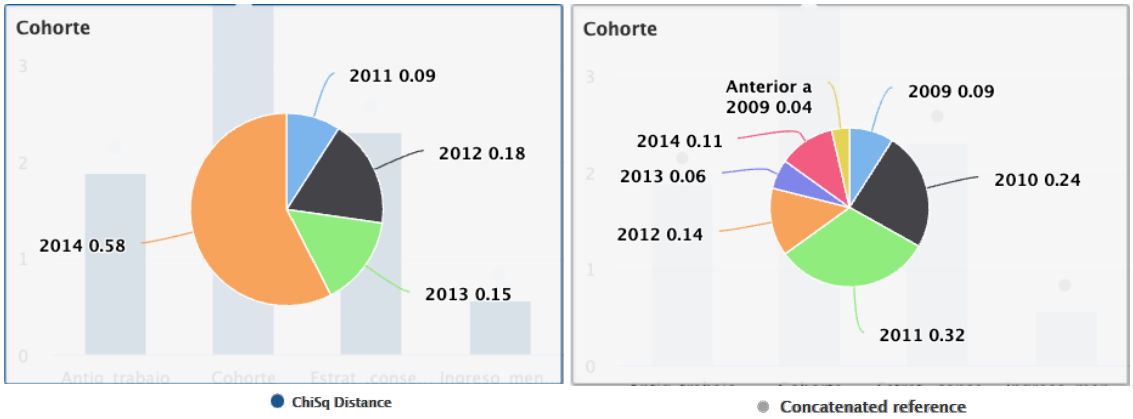


Figura 4.5. Distribución de las categorías de la variable Cohorte en la tabla Concatenada y la Tabla punto en el aplicativo T2Qv

Para empezar, en la tabla punto están ausentes las categorías correspondientes a las cohortes 2010, 2009 y anteriores a 2009, que, en cambio, sí aparecen en la tabla concatenada, quizás porque ésta recoge información de un periodo más amplio de años. Pero, esto también significa que en los procesos de graduación más recientes hay menos casos de cohortes rezagadas. Por ejemplo, mientras que el año 2021 sólo hubo graduados de las cuatro últimas cohortes, entre los de 2020 hubo de cinco cohortes, lo mismo que entre los graduados de 2019. En procesos de titulación anteriores a 2019 ya se registran graduados de hasta seis cohortes. Está mejorando el proceso de titulación.

Otra de las variables que están incidiendo con fuerza en la salida de control del proceso es Estrategia para conseguir empleo. La Figura 4.6 permite observar que los graduados de 2021 no acuden a medios de comunicación formales como prensa, radio y televisión, categoría que sí está presente en la tabla concatenada. Los graduados de 2021 utilizaron la red estatal Socio empleo en un 45% para vincularse laboralmente a instituciones

de salud, esto es 11% más que lo registrado en la tabla concatenada. Esta categoría muestra en 2021 una diferencia, a favor, de 11% comparada con la tabla concatenada. El uso de referencias personales decrece en 2 puntos porcentuales, pero se incrementa el uso de redes sociales, de un 8% en la tabla concatenada a un 12% en la tabla punto. La bolsa de empleo UTMACH muestra un incremento de un 2% en los graduados de 2021, pero en general tiene poca incidencia en la vinculación laboral de los graduados.

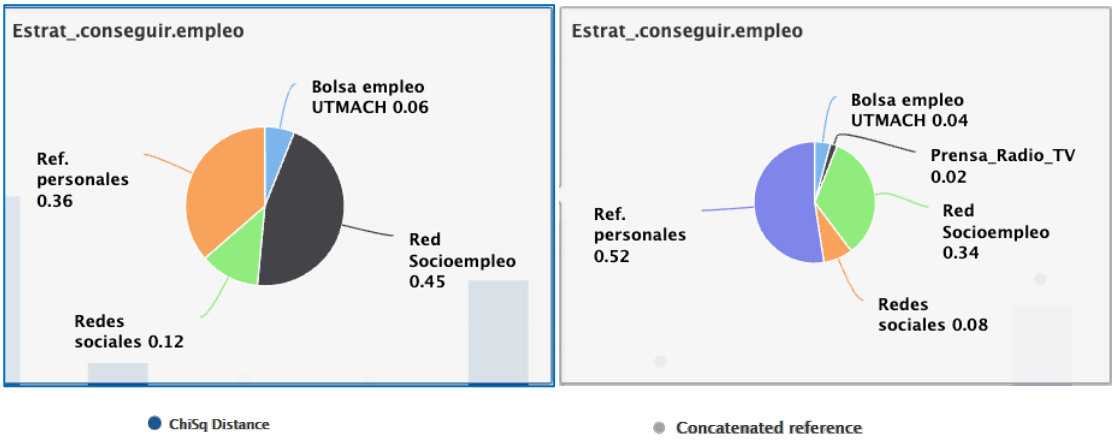


Figura 4.6. Distribución de las categorías de la variable Estrategia para conseguir empleo, en el aplicativo T2Qv

Antigüedad en el trabajo es otra de las variables importantes en esta parte del análisis (Figura 4.7). La categoría que tiene mayor representación en la tabla 2021 es la de 6 meses a 1 año de trabajo (73%), comparada con la tabla concatenada que llega al 45%. Esto se explica porque el 70% de los graduados del año 2021 están laborando su "año rural". Se refiere a los médicos recién graduados que ejercen en áreas rurales o remotas durante un periodo (usualmente un año), como requisito para luego ejercer la profesión en el país. Por esta razón, el 79% de los graduados de 2021 tienen, como máximo, una antigüedad de 1 año de trabajo.

La tabla concatenada recoge información de graduados de años anteriores, desde 2017, por esta razón exhibe la categoría de más de 2 años (4%), que no está en la tabla punto y la categoría de 1 a 2 años (8%) es superior que en la tabla 2021 (3%).

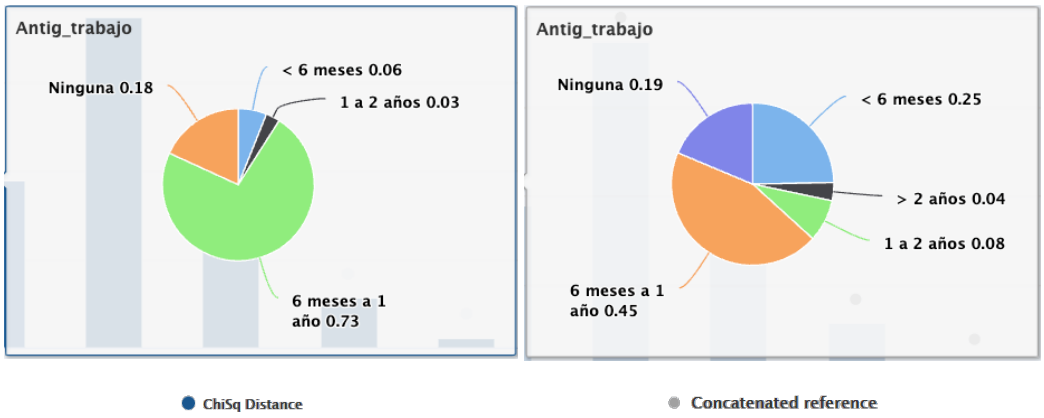


Figura 4.7. Distribución de las categorías de la variable Antigüedad en el trabajo, en el aplicativo T2Qv

Como ya se había notado en la Figura 4.4, después del Antigüedad en el trabajo, las demás variables tienen menores distancias Chi cuadrado entre las entre las masas de columna de la tabla concatenada y la tabla punto, lo que indica menor incidencia en el desplazamiento de la mediana del proceso y, en consecuencia, inciden menos en la salida de control del proceso. Como ejemplo se analiza a continuación el comportamiento de la variable Relación entre el mundo del trabajo y el Perfil Profesional del médico (Figura 4.8). Esta variable denota una Distancia χ^2 entre las masas de la tabla concatenada y la 2021, con un valor de 0.09, la más baja de entre las variables analizadas.

Esta variable busca determinar el nivel de satisfacción de los graduados con la Relación existente entre el mundo del trabajo y el Perfil Profesional del médico. Se observa

en la Figura 4.8 que la proporción de las categorías de la variable en la tabla punto y en la tabla Concatenada no es muy diferente, si acaso, la categoría *Baja* no aparece en la tabla 2021, pero, en la Concatenada sí está, aunque con un valor bajo (1%). Sin embargo, el potencial de incidir en la salida de control del proceso, además de la proporción de las categorías, reside en los tipos y niveles de asociación que las categorías de esta variable tienen con otras, como se observa en la ubicación de los puntos en los gráficos del MCA en las tablas Concatenada y 2021.

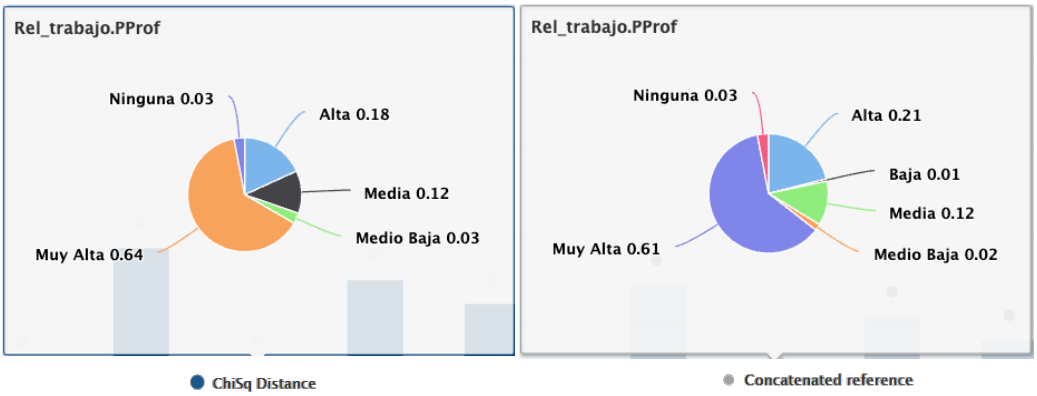


Figura 4.8. Distribución de las categorías de la variable Relación entre el mundo del trabajo y el Perfil Profesional en la tabla Concatenada y la 2021 en el aplicativo T2Qv

Esto es propio del enfoque multivariante en el control de procesos estadísticos, que considera que todas las variables contribuyen en mayor o menor medida al comportamiento del proceso. La salida de control no se atribuye a la acción individual de una variable o grupo de ellas, sino al efecto combinado de las variables correlacionadas. Este enfoque permite comprender los factores que afectan el proceso, identificar las diversas causas de su variación y adoptar medidas que permitan corregir las desviaciones aún en situaciones complejas donde

múltiples variables interactúan entre sí y pueden tener efectos indirectos en el comportamiento del proceso.

4.5. Contribuciones realizadas en el Capítulo 4

Luego de aplicar la metodología para el control de procesos estadísticos multivariantes con variables cualitativas al proceso de seguimiento a graduados de la carrera de Ciencias Médicas de la Universidad Técnica de Machala (UTMACH), mediante el aplicativo T2Qv, se presenta el siguiente análisis con sus respectivas contribuciones.

El Gráfico de Control Multivariante T2Qv indica que el proceso de seguimiento a graduados en 2021 difiere significativamente de los años anteriores. El valor del estadístico T^2_{med} de la tabla 2021 supera el límite de control, lo que sugiere una variación o anomalía importante en ese año, que requiere un análisis en profundidad.

El ACM permitió la visualización y comparación de las relaciones y variabilidades entre las categorías de las variables de la tabla Concatenada y de la tabla Punto (2021). Se destacan diferencias notables en la ubicación y asociación de categorías de variables entre estas dos tablas, lo que sugiere cambios en las dinámicas subyacentes, en las percepciones, experiencias y realidades de los graduados a lo largo del tiempo.

El análisis comparativo de la tabla concatenada y la tabla punto (2021), mediante el MCA, revela que hay una mayor dispersión de puntos en el gráfico MCA para la tabla punto (2021) en comparación con la tabla concatenada. Esto sugiere que en la tabla 2021 hay una mayor variabilidad y posiblemente una mayor influencia de variables latentes que afectan la representación. Por otro lado, la tabla concatenada muestra una distribución más concentrada

alrededor del eje principal, indicando que la satisfacción fue la variable dominante en este conjunto de datos.

Las categorías relacionadas con la satisfacción laboral y la relación entre el perfil profesional y el empleo presentan diferencias entre las dos tablas. En 2021, hay una categoría con graduados que, a pesar de tener empleo, expresa niveles bajos de satisfacción con varios aspectos de su formación, lo que no se observa en la tabla concatenada. Esta incongruencia entre empleo y satisfacción podría ser indicativa de problemas emergentes en la formación o en el mercado laboral que no se manifestaban en periodos anteriores.

Aunque ambos gráficos MCA identifican tres grupos distintos, hay diferencias notables en cómo están conformados y en las asociaciones de las categorías de variables. Por ejemplo, en la tabla 2021, el primer grupo expresó bajos niveles de satisfacción, pero no necesariamente asociados con desempleo, lo que contrasta con la tabla concatenada. Esto puede indicar un cambio en las percepciones o realidades de los graduados en 2021 en comparación con los graduados de años anteriores.

Además, hay diferencias en las asociaciones específicas entre las categorías dentro de los grupos identificados en las tablas. Por ejemplo, la relación entre la satisfacción con la malla curricular y las cohortes específicas varía entre las tablas. Estas diferencias pueden señalar cambios en la percepción de la calidad de la formación o ajustes en la malla curricular a lo largo del tiempo.

Las diferencias en las distancias Chi cuadrado entre las tablas comparadas indican que las variables Cohorte, Estrategia para conseguir empleo y Antigüedad en el trabajo, tienen una influencia significativa en el desplazamiento observado en la tabla punto (2021).

Esto sugiere que factores como las estrategias de búsqueda de empleo y las características de la cohorte de graduados se han vuelto más relevantes en 2021.

En cuanto a la estrategia que los graduados emplean para encontrar trabajo, se tiene que, mientras que en la tabla concatenada las referencias personales tienen mayor representación y aún hay registros de uso de medios tradicionales como prensa, radio y televisión, los graduados de 2021 parecen depender más de la red estatal Socio empleo y redes sociales. Esto puede reflejar un cambio en el panorama laboral, la eficacia de los recursos disponibles, o una adaptación a las tendencias actuales y plataformas tecnológicas para la búsqueda de empleo. La bolsa de empleo UTMACH todavía tiene una incipiente influencia en la vinculación laboral de los graduados.

La comparación de la distribución de cohortes entre las tablas revela una mejora en los procesos de graduación. La ausencia de cohortes más antiguas en la tabla de 2021 sugiere que las cohortes más recientes están avanzando y graduándose a un ritmo más regular y sin rezagos significativos, mientras que, en la tabla concatenada, que abarca un período más extenso, hay registros de graduados de hasta seis cohortes diferentes.

Los resultados de este análisis resaltan la importancia de monitorear y analizar de manera continua la experiencia y trayectoria de los graduados para adaptar las prácticas educativas y de apoyo laboral a las necesidades cambiantes del mercado y de los propios graduados.

La aplicación de técnicas de estadística multivariante es fundamental para identificar y entender las variaciones y desviaciones en conjuntos de datos complejos como el presentado. Estos hallazgos y conclusiones deben guiar futuras intervenciones y decisiones

estratégicas. La herramienta T2Qv proporciona visualizaciones interactivas que permiten comparaciones detalladas entre las tablas Concatenada y Punto. Estos gráficos interactivos ofrecen perspectivas sobre las categorías de las variables que han cambiado su distribución a lo largo del tiempo. Estas distribuciones revelan patrones y tendencias que pueden ser esenciales para comprender y gestionar el proceso de seguimiento a los graduados.

Capítulo 5

Análisis y discusión

En el control estadístico de procesos todavía no son muchas las propuestas publicadas sobre gráficos de control para variables cualitativas. Las diferencias entre procedimientos para la determinación de los estadísticos y los gráficos de control en este campo hacen difícil su comparación.

La estructura de la base de datos que se requiere para la aplicación del gráfico de control T2Qv, implica un conjunto de tablas superpuestas, donde cada una de ellas constituye una muestra. Las tablas deben tener las mismas variables, las cuales se ubican en columnas. Una de estas variables registra los datos que sirven para la identificación de las tablas, por ejemplo el año, mientras que, las demás aportan las categorías que operan en el MCA. De estas variables surgen las variables latentes que son las dimensiones que intervienen en el análisis.

Considerando que el MCA es una técnica de análisis multivariante que involucra una reducción de dimensiones y que el procesamiento de los datos para el gráfico T2Qv funciona con $p - 1$ dimensiones, desde el comienzo se requiere una base de datos con p variables ($p > 3$) para el análisis, además de la variable de identificación de las tablas, es decir, el gráfico no podría funcionar con un conjunto de datos que tenga menos de cuatro variables, incluyendo la variable de clasificación de tablas. Esta característica es una restricción en el uso del T2Qv, especialmente cuando el resultado del análisis de sensibilidad determinó que

el gráfico pierde estabilidad a bajas dimensiones y que cuando trabaja con altas dimensiones tiene un buen rendimiento.

Sobre este tema, en varios estudios de gráficos de control multivariante revisados en la literatura, los ejemplos analizan sólo dos o tres variables como casos de aplicación, esto se observa en las publicaciones de Epprecht, Aparisi y García-Bustos (2013); Raza y Aslam (2019); Jiang et al. (2002); Pastuizaca-Fernández, Carrión-García y Ruiz-Barzola (2015); Taleb (2009); Taleb, Limam, y Hirota (2006). Estos casos no podrían ser tratados con el T2Qv, porque tienen menos dimensiones que las requeridas.

Mientras tanto, en el ejemplo de aplicación con datos simulados que se presenta en esta investigación, el T2Qv analiza el comportamiento de 10 variables, la base de datos *Datak10Contaminated* tiene 11 columnas (tabla 2.2). Se podría asegurar, entonces, que una fortaleza del gráfico de control multivariante T2Qv es su buen rendimiento cuando trabaja con altas dimensiones, mientras que su debilidad se asocia al trabajo con bajas dimensiones y que con menos de cinco variables no puede funcionar.

Otra de las características del gráfico propuesto en esta investigación es que no se limita a procesar una simple tabla de datos de dimensiones n (filas) x p (variables), sino que se dirige al trabajo con bases de datos que tiene K tablas, donde cada k_i tabla es una muestra constituida por n observaciones (filas) y p variables (columnas), tomada en K diferentes momentos de análisis y se representa como un punto en el gráfico T2Qv. El análisis con el T2Qv, propuesto en esta investigación, se puede configurar como un cubo de datos $(n\ x\ p\ x\ k)$.

En publicaciones revisadas en la literatura se puede constatar que sus ejemplos de aplicación analizan una sola tabla (n x p) a la vez, donde cada n_i fila es una muestra. Por ejemplo, el gráfico de control MNP, de Lu (1998) contiene en su artículo una tabla de datos simulados de 30 muestras, donde cada una de ellas es un único individuo (objeto) que registra el conteo de defectos para tres características de la calidad. Asimismo, la ejemplificación que Chiu y Kuo (2008) presentaron de su gráfico de control MP se hizo con una tabla de datos simulados de 26 muestras, donde cada muestra representa a un individuo al que se le registra el D número de defectos o no conformidades asociadas a tres características de calidad.

En este sentido, la base de datos simulados, *Datak10Contaminated,* mediante la cual se ejemplifica la propuesta de esta investigación incluye un conjunto de 10 tablas y 11 variables, donde cada tabla, que se representa como un punto en el gráfico T2Qv, es una muestra conformada por 100 observaciones, en total 1000 filas. En consecuencia, una de las características más valiosas del gráfico de control T2Qv es su capacidad para trabajar con bases de datos que contengan K tablas, las cuales pueden estar constituidas por muchos individuos y con múltiples variables.

En la propuesta presentada en este artículo, cada uno de los individuos (filas) que conforman las diferentes nuestras pueden tener distintas configuraciones en función del número de categorías de las múltiples variables. Variables sociodemográficas, muy comunes en investigaciones de contexto social, tienen diverso número de categorías, por ejemplo, sexo, grupo etario, nivel de estudios, estado civil, estado o provincia de residencia, tipo de vivienda, presencia de discapacidad, tipo de discapacidad, autodenominación étnica, actividad laboral, nivel de ingresos económicos, entre otros. Los individuos registrados (filas) en una base de datos pueden tener diversas configuraciones en función de las categorías

seleccionadas para cada una de las variables que los caracterizan. El gráfico de control T2Qv, así como los demás gráficos que complementan esta propuesta, representan bien el comportamiento de las variables aunque entre éstas haya dicotómicas o politómicas de tres, 10 o más categorías, esta es otra de sus fortalezas.

Mientras tanto, hay publicaciones sobre gráficos de control multivariante para datos de atributos que, aunque en su análisis consideran varias características de calidad, al final clasifican a cada individuo por una sola de las variables analizadas. Es el caso de la propuesta de Mukhopadhyay (2008), que se demuestra con un caso de aplicación que controla 7 características de calidad en 24 muestras cuyo tamaño varía entre 20 y 404 individuos. Las variables responden a 6 tipos de defectos de la pintura en la cubierta de ventiladores de techo: cobertura deficiente, desbordamiento, defecto de empanada, burbujas, defectos de pintura, defectos de pulido. La séptima característica es la ausencia de defectos. A cada individuo se lo clasifica por su defecto más predominante, por consiguiente, en su registro sólo aparece un tipo de defecto o ausencia de defectos, lo que resulta en una pérdida de información sobre el efecto combinado de las variables sobre el proceso.

Siguiendo con el análisis, se observa numerosas propuestas publicadas sobre gráficos multivariantes que corresponden a la fase II del control estadístico de procesos y que, en consecuencia, pueden realizar ajustes a su rendimiento con miras a su optimización. Esto contribuye a la mejora de la eficacia en la detección de pequeños cambios en la media del proceso (Crosier, 1988; Pignatiello y Runger, 1990; Lowry et al., 1992); también al incremento de la eficiencia, optimizando tamaños de muestra (Aparisi, 1996), disminuyendo costos de muestreo (Ruiz-Barzola, 2013), minimizando el tiempo promedio de detección de cambios fuera del control (Epprecht, Aparisi y García-Bustos, 2013), utilizando estadísticos

de seguimiento basados en algoritmos de clasificación en datos complejos de alta dimensión (Liu, Liu y Jung, 2020), mejorando la calidad de la data al limpiar valores atípicos (Ahsan, et al., 2021).

El T2Qv es un gráfico que maneja variables cualitativas en fase I del control estadístico de procesos, en consecuencia, no se ha considerado aún la evaluación de su eficacia, ni su eficiencia. Esta característica del gráfico propuesto constituye una debilidad si se hace una comparación con los gráficos de fase II, pero, también es una oportunidad de mejora para considerar en futuros estudios que se dirijan a su optimización, estableciendo límites de control que se ajusten a los parámetros específicos del análisis, o, siguiendo a Aparisi (1996), Ruiz-Barzola (2013) y Soriano (2017), profundizado en el análisis de la relación entre el tamaño de muestra (n) y el ARL_1 en fase II, dado que si se aumenta n se reduce el ARL_1.

El ACM es una técnica factorial que busca asociaciones de variabilidad y hace que la información común a la mayoría de los casos tenga poca discriminación en los ejes principales del gráfico. Las categorías con baja frecuencia marginal se ubicarán en el borde del plano, mientras que las de alta frecuencia marginal se ubicarán más cerca del origen (Hoffman y De Leeuw, 1992). En el aplicativo T2Qv, si una variable se expresa en sólo una categoría en una tabla de la base de datos no se puede representar, se reporta un error, dado que no se podría medir su asociación con otras categorías de las demás variables, pero, si hay al menos un caso diferente, este se representará como un punto muy alejado en algún extremo de los ejes principales. Esta es una limitación propia de la metodología del MCA que la hereda el aplicativo T2Qv.

Una oportunidad para futuras investigaciones relacionadas con el control multivariante para variables cualitativas sería el desarrollo de una metodología que además del MCA incorpore técnicas multivariantes de tres vías. Podría ser viable la incorporación, por ejemplo, del JK-Meta Biplot (Galindo-Villardón et al., 2001) o el STATIS Dual (Escoufier, 1985; L'Hermier des Plantes, 1976), técnicas que facilitan la comprensión de la estructura interna del cubo de datos, previo trabajo de codificación cualitativa realizado en una fase de análisis cualitativo clásico (Caballero-Juliá, Villardón y García, 2017).

Otra oportunidad de desarrollo para futuras investigaciones es el uso de control estadístico de procesos multivariantes en el contexto del big data. Este tema todavía tiene un desarrollo incipiente en la literatura científica. El análisis de big data requiere la monitorizacióndel proceso secuencial subyacente de los datos observados para monitorear el desempeño longitudinal de los procesos (Qiu, 2017), o para determinar cómo cambia su distribución con el tiempo (Qiu, 2020). Los gráficos de control tradicionales de SPC en la actualidad tienen dificultades en actividades de diseño, reconocimiento e interpretación de patrones en medios de aprendizaje automático. Los algoritmos de *Machine Learning* se pueden integrar a los gráficos de control de SPC para resolver estos problemas (Tran et al., 2022) utilizando nuevos métodos, detectar anomalías tempranas y tomar mejores decisiones.

Se necesitan nuevas técnicas y herramientas de análisis de datos que puedan integrarse efectivamente en los procesos de control de calidad existentes y abordar desafíos como la validación de modelos y aplicaciones informáticas. Entre las aplicaciones potenciales que tendrían estas nuevas investigaciones están la detección dinámica de enfermedades, el reconocimiento de perfiles o imágenes en tiempo real.

Conclusiones

1. El desarrollo de gráficos de control para procesos multivariantes comprende cinco etapas evolutivas. La primera etapa corresponde al Gráfico T^2 de Hotelling, la segunda a las Mejoras al rendimiento de este gráfico. Las últimas tres recorren una trayectoria casi simultánea, que a veces se traslapa, y da lugar al desarrollo de tres propuestas para la monitorizaciónde procesos: Gráficos de control en entornos no paramétricos, Gráficos de control para variables cualitativas y Gráficos para el control de procesos con técnicas estadísticas multivariantes.
2. Al partir del estadístico T^2 de Hotelling, se ha establecido el T^2_{med}, un estadístico que adopta conceptos de robustez, utilizando el vector de medianas en vez de el vector de medias, para que contribuya el control de procesos para variables cualitativas en el contexto estadístico multivariante.
3. Se ha presentado una herramienta para el SPC multivariantes que realiza análisis de datos cualitativos, denominada gráfico de control T2Qv, basada en el Análisis de Correspondencias Múltiples y el gráfico T^2 de Hotelling. Es apropiada para una variedad de sectores, incluidos el productivo, industrial, ambiental, administrativo, de salud y social, donde predominan las variables nominales y ordinales.
4. El gráfico T2Qv es una herramienta eficaz para la detección e interpretación de anomalías en procesos estadísticos, evaluando el impacto de las variables en condiciones fuera de control. Se complementa con gráficos MCA de referencia de la tabla Concatenada y análisis Chi-cuadrado para comparar distribuciones en estados

anómalos. Gráficos interactivos adicionales facilitan el examen detallado de la distribución porcentual de las categorías de las variables implicadas.

5. El análisis de sensibilidad determinó que el gráfico de control T2Qv funciona bien cuando se trabaja con dimensiones altas, pero pierde estabilidad en dimensiones bajas.
6. Para facilitar la difusión y aplicación del método propuesto, se ha desarrollado un paquete estadístico computacional reproducible en R, llamado T2Qv, que está disponible en CRAN. Este paquete permite visualizar los resultados en un formato plano o interactivo e incluye un tablero Shiny que contiene todas las funciones integradas en un solo espacio.
7. El gráfico T2Qv tiene ventajas como su adaptabilidad a bases de datos cualitativas para n individuos, con p variables en K momentos distintos, como un cubo de datos. Funciona bien al trabajar con dimensiones altas, es estable en presencia de valores potencialmente atípicos, representa bien el comportamiento de las variables dicotómicas o politómicas con diferentes números de categorías. Es fácil de aplicar gracias a su complemento computacional.
8. Una de las limitaciones del T2Qv es la necesidad de una base de datos con un mínimo de cuatro variables, incluida la variable de clasificación de la tabla. Las variables en las tablas deben tener al menos una variación para que la asociación sea medible (al menos dicotómicas). El método pierde estabilidad al trabajar con dimensiones bajas.
9. Como oportunidad de futuras contribuciones respecto de esta propuesta están la optimización del gráfico para su fase II y la incorporación de técnicas multivariantes de tres vías como Statis Dual en futuras investigaciones y actualizaciones de software. Además, el T2Qv podría optimizarse para su inclusión en entornos de *big data*.

10. En un contexto multivariante, todas las variables contribuyen en mayor o menor medida a explicar el comportamiento del proceso, por lo que la salida fuera de control no puede atribuirse a la acción individual de una variable, o a la acción separada de un grupo de ellas, sino al efecto combinado de variables correlacionadas. Es por eso por lo que, un enfoque multivariante es necesario en el control estadístico de procesos.

Bibliografía

Adegoke, N, J Ajadi, A Mukherjee, y S Abbasi. 2022. «Nonparametric multivariate covariance chart for monitoring individual observations». *Computers & Industrial Engineering* 167(Article 108025: 360-8352. https://doi.org/10.1016/j.cie.2022.108025.

Ahsan, Muhammad, Muhammad Mashuri, y Hidayatul Khusna. 2022. «Comparing the performance of Kernel PCA Mix Chart with PCA Mix Chart for monitoring mixed quality characteristics». *Scientific Reports* 12 (1). https://doi.org/10.1038/s41598-022-20122-w.

Ahsan, Muhammad, Muhammad Mashuri, Heri Kuswanto, Dedy Dwi Prastyo y Hidayatul Khusna. 2021. «Outlier detection using PCA mix based T2 control chart for continuous and categorical data». *Communications in Statistics: Simulation and Computation* 50 (5): 1496-1523. https://doi.org/10.1080/03610918.2019.1586921.

Ahsan, Muhammad, Muhammad Mashuri, Heri Kuswanto, Dedy Dwi Prastyo, y Hidayatul Khusna. 2018. «Multivariate control chart based on PCA mix for variable and attribute quality characteristics». *Production and Manufacturing Research* 6 (1): 364-84. https://doi.org/10.1080/21693277.2018.1517055.

Alfaro, J, J Mondéjar, y M Vargas. 2010. *Gráficos multivariantes aplicados al control estadístico de la calidad*. La Coruña: Netbiblio.

Aparisi, F. 1996. «Hotelling's T2 Control Chart with Adaptive Sample Sizes». *International Journal of Production Research* 34 (10): 2853-62. https://doi.org/10.1080/00207549608905062.

Aparisi, F, y C García-Díaz. 2001. «Aumento de la potencia del gráfico de control multivariante T2 de Hotelling utilizando señales adicionales de falta de control». *Estadística española* 43 (148): 171-88.

Aparisi, Francisco, y Cesar L. Haro. 2001. «Hotelling's T2 control chart with variable sampling intervals». *International Journal of Production Research* 39 (14): 3127-40. https://doi.org/10.1080/00207540110054597.

Aslam, M, M Azam, y C Jun. 2014. «New Attributes and Variables Control Charts under Repetitive Sampling». *Industrial Engineering & Management Systems* 13 (1): 101-6. https://doi.org/10.7232/iems.2014.13.1.101.

Benzécri, J. 1973. *OL'analyse Des Correspondances. En l'analyse Des Données: Leçons Sur l'analyse Factorielle Et La Reconnaissance Des Formes Et Travaux*. Editado por Dunod. Paris: s.n.

Bougeard, S, y S Dray. 2018. «Supervised multiblock analysis in R with the ade4 package». *J. Stat. Softw* 86: 1-17.

Caballero-Juliá, Daniel, Ma Purificación Galindo Villardón, y Marie Carmen García. 2017. «JK-Meta-Biplot y STATIS Dual como herramientas de análisis de tablas textuales múltiples». *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao* 2017 (25): 18-33. https://doi.org/10.17013/risti.25.18-33.

CACES. 2023. *Modelo de Evaluación Externa con fines de Acreditación para el Aseguramiento de la Calidad de las UEP – 2023*. *Consejo de Aseguramiento de la Calidad de la Educación Superior*. Ecuador: Obtenido del sitio web del CACES: www.caces.gob.ec.

Chakraborti, S, P Laan, y S Bakir. 2001. «Nonparametric Control Charts: An Overview and Some Results». *Journal of Quality Technology* 33 (3): 304-15. https://doi.org/10.1080/00224065.2001.11980081.

Chiñas, Pamela, Ismael López, y José Vásquez. 2014. «Reconocimiento de variables multivariantes empleando el estadístico T2 Hotelling y MEWMA mediante las RNA's». *Ingeniería Investigación y Tecnología* 15 (1): 125-38.

Chiu, Jing Er, y Tsen I. Kuo. 2008. «Attribute control chart for multivariate Poisson distribution». *Communications in Statistics - Theory and Methods* 37 (1): 146-58. https://doi.org/10.1080/03610920701648771.

Crosier, R. 1988. «Multivariate Generalizations of Cumulative Sum Quality-Control Schemes». *Technometrics* 30 (3): 291-303. https://doi.org/http://www.jstor.org/stable/1270083.

Cubilla_ Montilla, M, A Nieto-Librero, P Galindo-Villardón, y C Torres-Cubilla. 2021. «Sparse HJ Biplot: A new methodology via elastic net». *Mathematics* 9: 1298.

Curran, J, y T Hersh. 2021. «Hotelling: Hotelling's T2 Test and Variants». Available online: https: //cran.r-project.org/web/packages/Hotelling/Hotelling.pdf .

Das, N. 2009. «A New Multivariate Non-Parametric Control Chart Based on Sign Test». *Quality Technology & Quantitative Management*, 155-69. https://doi.org/10.1080/16843703.2009.11673191.

Doganaksoy, N, F Faltin, y W T Tucker. 1991. «Identification of Out-of-Control Quality Characteristics in a Multivariate Manufacturing Environment». *Communications in Statistics - Theory and Methods* 20 (9): 2775– 2790. https://doi.org/10.1080/03610929108830667.

Dray, S, y A.B Dufour. 2007. «The ade4 package: Implementing the duality diagram for ecologists». *Journal of Statistical Software* 22: 1-20.

Edwards, W, y L Cavalli-Sforza. 1965. «A method for cluster analysis». *Biometrics* 21 (2): 362-75. https://doi.org/10.2307/2528096.

Efron, B. 1979. «Bootstrap Methods: Another Look at the Jackknife». *Ann. Statist* 7 (1). https://doi.org/10.1214/aos/1176344552.

Epprecht, Eugenio K., Francisco Aparisi, y Sandra García-Bustos. 2013. «Optimal linear combination of Poisson variables for multivariate statistical process control». *Computers and Operations Research* 40 (12): 3021-32. https://doi.org/10.1016/j.cor.2013.07.007.

Escofier, B, y J Pagès. 1994. «Multiple factor analysis (AFMULT package)». *Computational Statistics & Data Analysis*. Vol. 18.

Escoufier, Y. 1985. «Objectifs et procédures de l'analyse conjointe de plusieurs tableaux de données». *Statistique et analyse des données* 10 (1): 1-10. http://www.numdam.org/item?id=SAD_1985__10_1_1_0.

Escoufier, Y. 1987. «Three-mode data analysis: the STATIS method». *Methods for multidimensional data analysis*, 259-72.

Evans, J, y W Lindsay. 2015. *Administración y control de la calidad*. Editado por Cengage Learning Editores. 9° Ed. México D.F.

Faraz, Alireza, y Ahmad Parsian. 2006. «Statistical Papers Hotelling's T2 control chart with double warning lines». *Statistical Papers*. Vol. 47.

Farokhnia, M, y S Niaki. 2020. «Principal Component Analysis-Based Control Charts Using Support Vector Machines for Multivariate Non-Normal Distributions». *Communications in Statistics - Simulation and Computation* 49 (7): 1815-38. https://doi.org/10.1080/03610918.2018.1506032.

Filho, D, y L Luna. 2015. «Multivariate Quality Control of Batch Processes Using STATIS». *The International Journal of Advanced Manufacturing Technology* 82 (5): 867 – 875.

Gabriel, K. 1971. «The biplot graphic display of matrices with application to principal component analysis». *Biometrika* 58 (3): 453-67. https://doi.org/10.2307/2334381.

Galindo-Villardón, M. 1986. «Una alternativa de representación simultánea: HJ-Biplot». *Questiio* 10 (1): 13-23. https://doi.org/http://hdl.handle.net/2099/4523.

Galindo-Villardón, P, J Vicente-Villardón, C Zarza, M Fernandez-Gómez, y J Martın. 2001. « JK-META-BIPLOT: Una alternativa al método STATIS para el estudio espacio temporal de ecosistemas». En *Proceedings of the Conferencia Internacional de Estadística en Estudios Medioambientales*. Cádiz, España: Departamento de Estadística e Investigación OperativaUniversidad de Cádiz.

García, S. 2014. «Control multivariante estadístico de variables discretas tipo Poisson». Tesis Doctoral, Valencia, España: Universidad de Valencia.

Gneri, M, y E Pimentel. 2012. «Robustez Asintótica de la Estadística de Hotelling». *Revista De Educación Matemática* 27 (2). https://revistas.unc.edu.ar/index.php/REM/article/view/10204.

González, David. 2015. «Análisis de tablas de tres entradas como herramienta en las Ciencias Atmosféricas. Estudio de las depresiones del Golfo de Génova y del Golfo de Cádiz mediante el STATIS DUAL y el Análisis Parcial Triádico». Salamanca, España.

Gutiérrez Pulido, Humberto, y Román De la Vara Salazar. 2013. *Control estadístico de la calidad y Seis Sigma*. McGraw-Hill Education.

Hawkins, D. 1993. «Regression Adjustment for Variables in Multivariate Quality Control». *Journal of Quality Technology* 25 (3): 170-82. https://doi.org/10.1080/00224065.1993.11979451.

Hawkins, Douglas M., y Qiqi Deng. 2010. «A nonparametric change-point control chart». *Journal of Quality Technology* 42 (2): 165-73. https://doi.org/10.1080/00224065.2010.11917814.

Hazewinkel, Michiel. 2001. «Generalized overlapping shuffle algebras». *Journal of Mathematmil Scienca*. Vol. 106.

Herrera-Acosta, Roberto José, Richard Michael Wasinski-Zuñiga, y Indira Dayana Romero-Cabrera. 2018. «Contraste entre las cartas de control MR de Shewhart y Cusum Varianza en el monitoreo del potencial de hidrógeno en protectores de planta». *ITECKNE* 15 (2): 88-98. https://doi.org/10.15332/iteckne.v15i2.2070.

Hoffman, Donna L, y Jan De Leeuw. 1992. «Interpreting multiple correspondence analysis as a multidimensional scaling method». *Marketing Letters* 3 (3): 259-72. https://doi.org/10.1007/BF00994134.

Holgate, P. 1964. «Estimation for the Bivariate Poisson Distribution». *Biometrika* 51 (1-2): 241-87. https://doi.org/10.2307/2334210.

Hotelling, H. 1933. «Analysis of a Complex of Statistical Variables into Principal Components». *Journal of Educational Psychology* 24 (6): 417-41. https://doi.org/10.1037/h0071325.

Hotelling, Harold. 1947. *Multivariate quality control. In Techniques of Statistical Analysis*. McGraw-Hill. New York, USA.

Inselberg, A., y B Dimsdale. 1990. «Parallel coordinates: a tool for visualizing multi-dimensional geometry». En *1990*, editado por Proceedings of the First IEEE Conference on Visualization: Visualization'90, 361-78. San Francisco, EEUU. https://doi.org/10.1109/VISUAL.1990.146402.

Jackson, E. 1991. *A Use's Guide to Principal Components*. New York: John Wiley & Sons. https://doi.org/10.1002/0471725331.

Jensen, W A, L A Jones-Farmer, C W Champ, y W H Woodall. 2006. «Effects of Parameter Estimation on Control Chart Properties». *A Literature Review. ournal of Quality Technology* 38 (4): 349-64. https://doi.org/10.1080/00224065.2006.11918623.

Jiang, W., S. Au, K. Tsui, y M. Xie. 2002. «Process monitoring with univariate and multivariate c-charts».

Jin, J, y G Loosveldt. 2022. «Nonparametric multivariate control chart for numerical and categorical variables». *Communications in Statistics - Simulation and Computation*. https://doi.org/10.1080/03610918.2021.2023572.

Jin, Jiayun, y Geert Loosveldt. 2022. «Nonparametric multivariate control chart for numerical and categorical variables». *Communications in Statistics - Simulation and Computation* 0 (0): 1-19. https://doi.org/10.1080/03610918.2021.2023572.

Jones, L, W Woodall, y M Conerly. 1999. «Exact Properties of Demerit Control Charts». *Journal of Quality Technology* 31 (2): 207-16. https://doi.org/10.1080/00224065.1999.11979915.

Kaiser, H. 1958. «The varimax criterion for analytic rotation in factor analysis». *Psychometrika* 23: 187-200.

Kourti, Theodora, y John F. MacGregor. 1996. «Multivariate SPC methods for process and product monitoring». *Journal of Quality Technology* 28 (4): 409-28. https://doi.org/10.1080/00224065.1996.11979699.

Kumar, A, y P Mohapatra. 2012. «A New Approach to Design of Fuzzy Multi-Attribute Control Charts». *Computers & Industrial Engineering*. Chicago, Illinois, U.S.A., s.n. https://doi.org/10.1016/j.cie.2008.06.015.

Laungrungrong, Busaba, Connie M Borror, y Douglas C Montgomery. 2011. «EWMA control charts for multivariate Poisson-distributed data». *Int. J. Quality Engineering and Technology*. Vol. 2.

Lavit, C. 1988. *Analyse conjointe de tableaux quantitatifs*. Paris: Masson.

Lavit, Christine, Yves Escoufier, Robert Sabatier, y Pierre Traissac. 1994. «The ACT (STATIS method)». *Computational Statistics & Data Analysis* 18 (1): 97-119. https://doi.org/10.1016/0167-9473(94)90134-1.

Lê, S, J Josse, y F Husson. 2008. «FactoMineR: An R package for multivariate analysis». *Stat. Softw.* 25: 1-18.

Ledesma, R. 2008. «Metodología de Encuestas SOFTWARE DE ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES: UNA REVISIÓN COMPARATIVA». *Metodol. Encuestas* 10: 59-75.

Lee, L, y A Branco. 2009. «Control Charts for Individual Observations of a Bivariate Poisson Proces». *The International Journal of Advanced Manufacturing Technology* 43 (7-8): 744-55. https://doi.org/10.1007/s00170-008-1746-4.

Li, J, F Tsung, y C Zou. 2012. «Directional Control Schemes for Multivariate Categorical Processes». *Journal of Quality Technology* 2 (44): 136-45. https://doi.org/10.1080/00224065.2012.11917889.

Li, Yanting, Dezhao Pei, y Zhenyu Wu. 2020. «A multivariate non-parametric control chart based on run test». *Computers and Industrial Engineering* 149 (noviembre). https://doi.org/10.1016/j.cie.2020.106839.

Liu, Yiqi, Yumin Liu, y Uk Jung. 2020. «Nonparametric multivariate control chart based on density-sensitive novelty weight for non-normal processes». *Quality Technology and Quantitative Management* 17 (2): 203-15. https://doi.org/10.1080/16843703.2019.1577345.

Lopez, C.P. 2004. *Técnicas de análisis Multivariante de Datos*. Editado por Pearson Educación. London, UK.

Lowry, C, y D Montgomery. 1995. «Una revisión de gráficos de control multivariados». *IIE Transactions* 27 (6): 800-810. https://doi.org/10.1080/07408179508936797.

Lowry, Cynthia A, William H Woodall, Charles W Champ, y Steven E Rigdon. 1992. «A Multivariate Exponentially Weighted Moving Average Control Chart». *Technometrics*. Vol. 34.

Lu, X. 1998. «Control Chart for Multivariate Attribute Processes». *International Journal of Production Research* 36 (12): 3477-89. https://doi.org/10.1080/002075498192166.

Mahalanobis, P. 1936. «On the Generalised Distance in Statistics». En *Proceedings of the National Institute of Science of India*, Volumen 12:49-55.

Marín, E. 2016. «Evaluación de estrategias docentes universitarias: una aplicación práctica del control estadístico de procesos en estudios de empresas.» Thesis de Doctorado, Huelva: Universidad de Huelva.

Martínez J, Palacios G, y Oliva D. 2023. «Guía para la Revisión y el Análisis Documental: Propuesta desde el Enfoque Investigativo». *RA XIMHAI* 19 (Núm. 1): 67-83. https://doi.org/10.35197/rx.19.01.2023.03.jm.

Mashuri, Muhammad, Muhammad Ahsan, Muhammad Hisyam Lee, Dedy Dwi Prastyo, y Wibawati. 2021. «PCA-based Hotelling's T2 chart with fast minimum covariance determinant (FMCD) estimator and kernel density estimation (KDE) for network intrusion detection». *Computers and Industrial Engineering* 158 (agosto). https://doi.org/10.1016/j.cie.2021.107447.

Mason, Robert L., Nola D. Tracy, y John C. Young. 1995. «Decomposition of T2 for multivariate control chart interpretation». *Journal of Quality Technology* 27 (2): 99-108. https://doi.org/10.1080/00224065.1995.11979573.

Merlo, J, A Cordero-Franco, y V Tercero-Gómez. 2022. «Nonparametric multivariate processes monitoring with guaranteed in-control performance for changes in location». *Computers & Industrial Engineering* 166 (Article 107940, ISSN 0360-8352.). https://doi.org/10.1016/j.cie.2022.107940.

Michailidis, George, y Jan de Leeuw. 1998. «The Gifi System of Descriptive Multivariate Analysis». *Statistical Science* 13 (4): 307-36. http://www.jstor.org/stable/2676814.

Montgomery, D. 2012. *Introduction to Statistical Quality Control*. 7th ed. NJ, USA: Wiley Global Education: Hoboken.

Mukhopadhyay, Arup Ranjan. 2008. «Multivariate attribute control chart using Mahalanobis D2 statistic». *Journal of Applied Statistics* 35 (4): 421-29. https://doi.org/10.1080/02664760701834980.

Nenadi´c, Oleg Nenadi´c, Universitat Pompeu, y Fabra Barcelona. 2007. «Journal of Statistical Software Correspondence Analysis in R, with Two-and Three-dimensional Graphics: The ca Package Michael Greenacre». http://www.jstatsoft.org/.

Nieto-Librero, A. 2015. «Package 'BiplotbootGUI'». Available online: http://cran.nexr.com/web/packages/biplotbootGUI/index. html .

Pacella, Massimo, Quirico Semeraro, y Alfredo Anglani. 2004. «Manufacturing quality control by means of a Fuzzy ART network trained on natural process data». *Engineering Applications of Artificial Intelligence* 17 (1): 83-96. https://doi.org/10.1016/j.engappai.2003.11.005.

Page, E S. 1954. «Continuous Inspection Schemes». *Biometrika* 41 (1).

Pastuizaca Fernández, María Nela, Andrés Carrión García, y Omar Ruiz Barzola. 2015. «Multivariate multinomial T sup2/sup control chart using fuzzy approach». *International Journal of Production Research* 53 (7): 2225-38. https://doi.org/10.1080/00207543.2014.983617.

Patel, H. 1973. «Quality Control Methods for Multivariate Binomial and Poisson Distributions». *Technometrics* 15 (1): 103-12. https://doi.org/10.1080/00401706.1973.10489014.

Pearson, K. 1901. «On lines and planes of closest fit to systems of points in space». *Philosophical Magazine* 2 (11): 559-72. https://doi.org/10.1080/14786440109462720.

Phaladiganon, P. 2011. «Bootstrap-Based T2 Multivariate Control Charts». *Communications in Statistics - Simulation and Computation* 40 (5): 645-62. https://doi.org/10.1080/03610918.2010.549989.

Pignatiello, J, y G Runger. 1990. «Comparisons of multivariate CUSUM charts». *Journal of Quality Technology* 22 (3): 173– 186. https://doi.org/10.1080/00224065.1990.11979237.

Plantes, L.'Hermier Des. 1976. «Structuration des Tableaux à trois Indices de la Statistique». *Théorie et applications d'une méthode d'analyse conjointe*. Ph.D. Thesis, Montpellier, France.: Universite des Sciences et Techniques du Languedoc.

Qiu, Peihua. 2013. *Introduction to Statistical Process Control*. New York: CRC press.

Qiu, Peihua. 2008. «Distribution-free multivariate process control based on log-linear modeling». *IIE Transactions (Institute of Industrial Engineers)* 40 (7): 664-77. https://doi.org/10.1080/07408170701744843.

Qiu, Peihua. 2017. «Statistical Process Control Charts as a Tool for Analyzing Big Data». En *Big and Complex Data Analysis: Methodologies and Applications* , editado por S Ejaz Ahmed, 123-38. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-41573-4_7.

Qiu, Peihua. 2020. «Big Data? Statistical Process Control Can Help!» *The American Statistician* 74 (4): 329-44. https://doi.org/10.1080/00031305.2019.1700163.

R Core Team. 2023. «R: A Language and Environment for Statistical Computing». Vienna, Austria: R Foundation for Statistical Computing.

Ramos, Miriam, José Ascencio, Miriam Vanessa Hinojosa, Francisco Vera, Omar Ruiz, María Isabel Jimenez-Feijoó, y Purificación Galindo. 2021. «Multivariate statistical process control methods for batch production: a review focused on applications». *Production and Manufacturing Research* 9 (1): 33-55. https://doi.org/10.1080/21693277.2020.1871441.

Ramos-Barberán, M. 2017. «Una Alternativa a los Métodos Clásicos de Control de Procesos Basada en Coordenadas Paralelas». *Métodos Biplot y Statis*. Salamanca.

Ramos-Barberán, Miriam, Miriam Vanessa Hinojosa-Ramos, José Ascencio-Moreno, Francisco Vera, Omar Ruiz-Barzola, y María Purificación Galindo-Villardón. 2018. «Batch process control and monitoring: a Dual STATIS and Parallel Coordinates (DS-PC) approach». *Production and Manufacturing Research* 6 (1): 470-93. https://doi.org/10.1080/21693277.2018.1547228.

Raza, Muhammad Ali, y Muhammad Aslam. 2019. «Design of control charts for multivariate Poisson distribution using generalized multiple dependent state sampling». *Quality Technology and Quantitative Management* 16 (6): 629-50. https://doi.org/10.1080/16843703.2018.1497935.

Robert, P, y Y Escoufier. 1976. «A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient Author(s». *Source: Journal of the Royal Statistical Society. Series C (Applied Statistics)*. Vol. 25.

Roberts, S. W. 1959. «Control Chart Tests Based on Geometric Moving Averages». *Technometrics* 1 (3): 239-50. https://doi.org/10.1080/00401706.1959.10489860.

Rodríguez, A, y A Pérez. 2017. «Métodos científicos de indagación y de construcción del conocimiento». *Revista Escuela de Administración de Negocios*, n.º 82 (julio): 175-95. https://doi.org/10.21158/01208160.n82.2017.1647.

Rojas-Preciado, W, M Rojas-Campuzano, P Galindo-Villardón, y O Ruiz-Barzola. 2022. «T2Qv: Control Qualitative Variables». https://cran.r-project.org/web/packages/T2Qv/index.html.

Rojas-Preciado, W., M. Rojas-Campuzano, P. Galindo-Villardón, y O. Ruiz-Barzola. 2023. «Control Chart T2Qv for Statistical Control of Multivariate Processes with Qualitative Variables». *Mathematics* 11 (12). https://doi.org/10.3390/math11122595.

Ross, Gordon J., y Niall M. Adams. 2012. «Two nonparametric control charts for detecting arbitrary distribution changes». *Journal of Quality Technology* 44 (2): 102-16. https://doi.org/10.1080/00224065.2012.11917887.

Ruelas, E A, J Cruz, E Ruelas, y D Ieee. 2020. «Statistical Control of Multivariant Processes Through the Artificial Neural Network Multilayer Perceptron and the MEWMA Graphic Analysis». Vol. 18.

Ruiz-Barzola, O. 2013. «Gráficos de Control de Calidad Multivariantes con Dimension Variable». *2013*. Ph. D. Thesis, Valencia, España: Universitat Politécnica de Valéncia.

Saltos, G, M Flores, L Horna, y K Morales. 2020. «New methodologies applied to multivariate monitoring of student performance using control charts and threshold systems». *Perfiles* 1: 68-74.

Scrucca, L. 2004. «qcc: An R package for quality control charting and statistical process control». *R News* 4 (1): 11-17.

Seabold, S, y J Perktold. 2010. «statsmodels: Econometric and statistical modeling with python». En *In Proceedings of the 9th Python in Science Conference*. Austin, TX, USA.

Shabbak, Ashkan, y Habshah Midi. 2012. «An improvement of the Hotelling T2 statistic in monitoring multivariate quality characteristics». *Mathematical Problems in Engineering* 2012. https://doi.org/10.1155/2012/531864.

Soetaert, K. 2021. «plot3D: Plotting Multi-Dimensional Data». Available online: https://cran.r-project.org/ web/packages/plot3D/index.html.

Song, Z, A Mukherjee, P Qiu, y M Zhou. 2023. «Two Robust Multivariate Exponentially Weighted Moving Average Charts to Facilitate Distinctive Product Quality Features Assessment». *Computers & Industrial Engineering* 183 (https://doi.org/10.1016/j.cie.2023.109469). https://doi.org/https://doi.org/10.1016/j.cie.2023.109469.

Soriano Martínez, Eva. 2017. «Estudio de la influencia de la Fase I en el desempeño de la Fase II en el gráfico T2 de Hotelling». Valencia: Universitat Politécnica de Valéncia.

Sparks, R, A Adolphson, y A Phatak. 1997. «Multivariate process monitoring using the dynamic biplot». *International Statistical Review* 65 (3): 325-49. https://doi.org/10.2307/1403375.

Spearman, Charles. 1904. «General intelligence objectively determined and measured.» *Am. J. Psychol.* 15: 201-93.

Sullivan, J, y W Woodall. 1999. «A Comparison of Multivariate Control Charts for Individual Observations». *Journal of Quality Technology* 28 (4): 398-408. http://asq.org/qic/display-.

Sun, R, y F Tsung. 2003. «A Kernel-distance-based multivariate control chart using support vector methods». *International Journal of Production Research* 41 (13): 2975-89. https://doi.org/10.1080/1352816031000075224.

Taleb, H. 2009. «Control Charts Applications for Multivariate Attribute Processes». *Computers & Industrial Engineering* 56 (1): 399-410. https://doi.org/10.1016/j.cie.2008.06.015.

Taleb, Hassen, Mohamed Limam, y Kaoru Hirota. 2006. «Multivariate Fuzzy Multinomial Control Charts». *Quality Technology & Quantitative Management* 3 (4): 437-53. https://doi.org/10.1080/16843703.2006.11673125.

Tang, A, A Mukherjee, y X Wang. 2023. «Distribution-free Phase-II monitoring of high-dimensional industrial processes via origin and modified interpoint distance based algorithms». *Computers & Industrial Engineering* 179 (Article 109161). https://doi.org/10.1016/j.cie.2023.109161.

Thurstone, L. 1947. *Multiple-Factor Analysis: A Development and Expansion of the Vectors of Mind*. Editado por University of Chicago Press: Chicago, IL, USA.

Tracy, N, J Young, y R Mason. 1992. «Multivariate control charts for individual observations». *Journal of Quality Technology* 28 (4): 88-95. https://doi.org/https://doi.org/10.1080/00224065.1992.12015232.

Tran, P.H., A. Ahmadi Nadi, T.H. Nguyen, K.D., Tran, y K.P. Tran. 2022. «Application of Machine Learning in Statistical Process Control Charts: A Survey and Perspective». En *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*, editado por Kim Phuc Tran, 7-42. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-83819-5_2.

Tuerhong, Gulanbaier, y Seoung Bum Kim. 2014. «Gower distance-based multivariate control charts for a mixture of continuous and categorical variables». *Expert Systems with Applications* 41 (4 PART 2): 1701-7. https://doi.org/10.1016/j.eswa.2013.08.068.

Vicente-Villardón, J. 2010. «MULTBIPLOT: A Package for Multivariate Analysis Using Biplots». Salamanca, España: Departamento de Estadística, Universidad de Salamanca.

Virtanen, P, R Gommers, T.E Oliphant, M Haberland, T. Reddy, D Cournapeau, E Burovski, P Peterson, W. Weckesser, y J Bright. 2020. «SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python». *Nat. Methods* 17: 261-72.

Xue, L, y P Qiu. 2020. «A nonparametric CUSUM chart for monitoring multivariate serially correlated processes». *Journal of Quality Technology* 53 (4): 396-409. https://doi.org/10.1080/00224065.2020.1778430.

Yue, J, y L Liu. 2017. «Multivariate nonparametric control chart with variable sampling interval». *Applied Mathematical Modelling* 52 (N): 307-904. https://doi.org/10.1016/j.apm.2017.08.005.

Zhao, C. 2023. «A novel parallel classification network for classifying three-dimensional surface with point cloud data». *J Intell Manuf* 34: 515-27. https://doi.org/10.1007/s10845-021-01802-2.

Apéndices

Apéndice A. Sección de la Base de datos *CMSG*

Work_relationship_and_profesional_profile	Satisfaction_with_knowledge_and_skills_acquired	Satisfaction_Curricular_Mesh	Satisfaction_Learning_Strategies	Satisfaction_Formative_Research	Satisfaction_application_of_Investigation_Linking	Satisfaction_Diffusion_of_research_results	Type_of_graduate_studying	Interest_by_specialty	Strategy_to_get_a_job	Employment_situation	Hierarchical_labor_level	Seniority_at_work	Monthly_income	Cohort	grad_year
Low Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Pediatrics	Personal References	Unemployed	None	None	None	Before 2009	2020
High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Master's Degree	Higher Medical Education	Press_Radio_TV	Independent Professional	Other	Less than 6 months	901 to 1200	Before 2009	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	UTMACH Job Board	Public Employee	Other	Less than 6 months	901 to 1200	Before 2009	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	Personal References	Independent Professional	Private MD	6 months to 1 year	901 to 1200	Before 2009	2017 - 2018
Very High Relationship	Moderately	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	None	Community Health	Personal References	Private Employee	Private MD	Less than 6 months	1201 to 1500	Before 2009	2019
Moderate Relationship	Moderately	Low Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	None	Surgery	Socioemployment Network	Private Employee	Private MD	1 to 2 years	376 to 600	Before 2009	2020
Very High Relationship	High	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Surgery	Personal References	Private Employee	Resident MD	Less than 6 months	901 to 1200	2009	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	Unsatisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	Less than 6 months	601 to 900	2009	2017 - 2018
Moderate Relationship	High Moderate	Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Master's Degree	Higher Medical Education	Personal References	Independent Professional	Private MD	1 to 2 years	601 to 900	2009	2020
High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Pediatrics	Personal References	Public Employee	Rural MD	6 months to 1 year	601 to 900	2009	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2009	2017 - 2018
Moderate Relationship	High Moderate	Satisfactory	Satisfactory	Low Satisfactory	Satisfactory	Satisfactory	None	Pediatrics	Personal References	Unemployed	None	None	None	2009	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Master's Degree	Surgery	Personal References	Private Employee	Other	1 to 2 years	601 to 900	2009	2019
Very High Relationship	High	Satisfactory	Satisfactory	Very Satisfactory	Satisfactory	Satisfactory	None	Higher Medical Education	UTMACH Job Board	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2009	2017 - 2018
High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	None	Higher Medical Education	Personal References	Unemployed	None	None	None	2009	2019
Very High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Surgery	Press_Radio_TV	Unemployed	None	None	None	2009	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2009	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2009	2017 - 2018
Very High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Master's Degree	Surgery	Personal References	Private Employee	Administrative	6 months to 1 year	1201 to 1500	2009	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Master's Degree	Gynecology and Obstetrics	Personal References	Private Employee	Resident MD	6 months to 1 year	601 to 900	2009	2017 - 2018
High Relationship	Moderately	Low Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	None	Surgery	Personal References	Unemployed	None	None	None	2009	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Clinical	Personal References	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	Socioemployment Network	Private Employee	Resident MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Pediatrics	Socioemployment Network	Public Employee	Resident MD	6 months to 1 year	1201 to 1500	2010	2019
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Pediatrics	Personal References	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Specialization	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	Moderately	Low Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Low Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
High Relationship	Moderately	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Community Health	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High Moderate	Very Satisfactory	Satisfactory	Satisfactory	Very Satisfactory	Satisfactory	None	Gynecology and Obstetrics	Personal References	Private Employee	Resident MD	Less than 6 months	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Specialization	Surgery	Personal References	Unemployed	None	None	None	2010	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Low Satisfactory	Satisfactory	Satisfactory	None	Gynecology and Obstetrics	Personal References	Private Employee	Resident MD	Less than 6 months	601 to 900	2010	2017 - 2018
Very High Relationship	High	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Clinical	Socioemployment Network	Public Employee	Administrative	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	Moderately	Low Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	None	Gynecology and Obstetrics	Socioemployment Network	Public Employee	Administrative	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	None	Higher Medical Education	Personal References	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High Moderate	Very Satisfactory	Satisfactory	Satisfactory	Very Satisfactory	Satisfactory	None	Surgery	Social Media	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Surgery	Personal References	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Moderate Relationship	Moderately	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Master's Degree	Higher Medical Education	Personal References	Unemployed	None	None	None	2010	2017 - 2018
High Relationship	High	Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Clinical	Personal References	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	Personal References	Private Employee	Resident MD	Less than 6 months	376 to 600	2010	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	Master's Degree	Surgery	Personal References	Private Employee	Private MD	1 to 2 years	2101 to 2400	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Pediatrics	Socioemployment Network	Unemployed	None	None	None	2010	2017 - 2018
Very High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Administrative	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Satisfactory	Very Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Community Health	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Very Satisfactory	Very Satisfactory	Satisfactory	Satisfactory	None	Pediatrics	UTMACH Job Board	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Satisfactory	Low Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High Moderate	Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
High Relationship	Moderately	Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	Low Satisfactory	None	Community Health	Social Media	Unemployed	None	None	None	2010	2017 - 2018
Moderate Relationship	Moderately	Low Satisfactory	Low Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	None	Pediatrics	Social Media	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
High Relationship	High Moderate	Low Satisfactory	Satisfactory	Low Satisfactory	Unsatisfctory	Unsatisfctory	None	Surgery	Personal References	Private Employee	Administrative	6 months to 1 year	1201 to 1500	2010	2017 - 2018
High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Clinical	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	Personal References	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High Moderate	Very Satisfactory	Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	Personal References	Unemployed	None	None	None	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Pediatrics	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Pediatrics	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Administrative	6 months to 1 year	901 to 1200	2010	2017 - 2018
Very High Relationship	High	Satisfactory	Very Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Community Health	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Higher Medical Education	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Low Satisfactory	Low Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	601 to 900	2010	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Community Health	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2010	2017 - 2018
Moderate Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Pediatrics	Personal References	Unemployed	None	None	None	2011	2017 - 2018
High Relationship	High Moderate	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	None	Clinical	Personal References	Unemployed	None	None	None	2011	2017 - 2018
High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Community Health	Socioemployment Network	Unemployed	None	None	None	2012	2017 - 2018
Very High Relationship	High	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Pediatrics	Personal References	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2012	2021
High Relationship	High Moderate	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	Very Satisfactory	None	Surgery	Socioemployment Network	Public Employee	Rural MD	6 months to 1 year	901 to 1200	2012	2021