Universidade Vigo

Trabajo Fin de Máster

Técnicas para el control estadístico de calidad a partir de datos multivariantes

Nekane Sabas Pirón

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Técnicas de control estatístico da calidade baseado en datos multivariados
Título en español: Técnicas para el control estadístico de calidad a partir de datos multivariantes
English title: Techniques for statistical quality control based on multivariate data
Modalidad: Modalidad A
Autor/a: Nekane Sabas Pirón, Universidade de Santiago de Compostela
Director/a: Salvador Naya Fernández, Universidad de Coruña; ,
Tutor/a:
Breve resumen del trabajo: En el presente proyecto se revisan diversas alternativas para el control estadístico de procesos en los sectores de la industria y servicios desde la perspectiva del análisis multivariante. Se describirán las nuevas herramientas utilizadas en el nuevo paradigma de datos correspondiente a la Industria 4.0, es decir, gráficos de control especialmente diseñados para ser aplicados a bases de datos compuestas por gran cantidad de variables monitorizadas continuamente frente al tiempo. También se propondrá la programación en R y la aplicación de dichas técnicas a una base de datos real referentes al sistema de enfriamiento de un hotel sudamericano. Se presentarán diferentes metodologías para la representación de los gráficos de control aplicados a datos multivariantes, incluyendo procedimientos válidos aún cuando los datos multivariantes no cumplen las especificaciones de normalidad e independencia. Finalmente, se tratará de implementar dichos métodos a datos reales, todo ello acompañado de un análisis exploratorio multivariante y aplicaciones de métodos de partición que favorezcan el análisis, con el objetivo de obtener altos porcentajes de detección de anomalías en el proceso sin elevar el porcentaje de falsas alarmas.

Don Salvador Naya Fernández, Departamento de Matemáticas de la Universidad de Coruña, informa que el Trabajo Fin de Máster titulado

Técnicas para el control estadístico de calidad a partir de datos multivariantes

fue realizado bajo su dirección por doña Nekane Sabas Pirón para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 15 de julio de 2020.

El director:

Don Salvador Naya Fernández

El director:

76353731X SALVADOR NAYA (R: Q6550005J)

Firmado digitalmente por 76353731X SALVADOR NAYA (R: Q6550005J)
Nombre de reconocimiento (DN): 2.5.4.13=Ref:AEAT/AEAT0287/PUESTO 1/36826/10012020133537, serialNumber=IDCES-76353731X, givenName=SALVADOR, sn=NAYA FERNANDEZ, cn=76353731X SALVADOR NAYA (R: Q6550005J), 2.5.4.97=VATES-Q6550005J, o=UNIVERSIDADE DA CORUÑA, c=ES
Fecha: 2020.07.13 16:19:03 +02'00'

El tutor:

El tutor:

La autora:

Doña Nekane Sabas Pirón

NEKANE SABAS

Firmado digitalmente por NEKANE SABAS
Fecha: 2020.07.13 16:51:55 +02'00'

Agradecimientos

Agradezco a mi familia su esfuerzo durante todos estos años de estudio y su apoyo incondicional en todas las decisiones tomadas, en especial a Idoia y Andoni, que son los que me han acompañado de la mano en todo mi recorrido académico y profesional.

Me gustaría agradecer la ayuda proporcionada a dos grandes profesionales de los cuales he aprendido mucho y a los que admiro enormemente. En primer lugar a Salvador Naya Fernández, mi tutor del presente Proyecto de Fin de Máster, por guiarme en todo este camino, junto a Javier Tarrío Saavedra, por su colaboración activa y dedicación. Por otro lado agradecer a Sonia Zaragoza Fernández, por su paciencia y tiempo dedicado a este proyecto.

Quiero dar las gracias también a todos los profesores que he tenido a lo largo del Máster, por los conocimientos adquiridos y su poder de alimentar la admiración y el respeto que siento por esta ciencia.

Finamente, dedicar el trabajo a aquellos que de alguna manera no están, pero siguen siendo el motor para hacer grandes cosas.

Índice general

Índice de figuras

Índice de cuadros

Resumen

Resumen en español

En la actualidad, el uso intensivo de la adquisición de datos, la sensórica y la utilización de ordenadores en el monitoreo de procesos ha conducido a la necesidad del análisis de procesos industriales con dos o mas variables correlacionadas, en los cuales el control estadístico de procesos y los análisis de capacidad deben ser llevados a cabo usando metodologías multivariantes. En el presente proyecto se revisan diversas alternativas para el control estadístico de procesos en los sectores de la industria y servicios desde la perspectiva del análisis multivariante. Se describirán las nuevas herramientas utilizadas en el nuevo paradigma de datos correspondiente a la Industria 4.0, es decir, gráficos de control especialmente diseñados para ser aplicados a bases de datos compuestas por gran cantidad de variables monitorizadas continuamente frente al tiempo. También se propondrá la programación en R y la aplicación de dichas técnicas a una base de datos real referentes al sistema de enfriamiento de un hotel en Latinoamérica, que representa un caso de estudio real.

Se presentarán diferentes metodologías de gráficos de control aplicados a datos multivariantes, tanto paramétricas como no paramétricas, tanto aquellas tradicionales, que tienen una larga historia de aplicación como técnicas más recientes. Finalmente, se evaluará la aplicación de dichos métodos al caso de estudio real de un gran hotel el Caribe, proporcionado por la empresa Fridama SA, incluyendo también un estudio exploratorio multivariante y aplicaciones de métodos de clasificación no supervisada que proporcionen información relevante acerca del funcionamiento de las instalaciones de climatización (HVAC) del hotel, siempre con el objetivo último de controlar las instalaciones HVAC y proporcionar un sistema de detección de anomalías que favorezca el mantenimiento predictivo, previniendo en la medida de lo posible la identificación de falsas alarmas.

Palabras clave: Control estadístico de calidad, gráficos de control multivariante, eficiencia energética, calidad 4.0.

Resumo en galego

Na actualidade, o uso intensivo de adquisición de datos, sensores e o uso de ordenadores no control de procesos levou á necesidade da análise de procesos industriais con dúas ou máis variables correlacionadas, nas que o control estatístico do proceso e as análises de capacidade deben realizarse empregando metodoloxías multivariables. Este proxecto revisa diversas alternativas para o control estatístico de procesos no sector e servizos desde a perspectiva da análise multivariante. Describiranse as novas ferramentas utilizadas no novo paradigma de datos correspondente a Industry 4.0, é dicir, gráficos de control especialmente deseñados para ser aplicados a bases de datos formadas por un gran número de variables controladas continuamente ao longo do tempo. Tamén se proporá a programación en R e a aplicación destas técnicas a unha base de datos real referida ao sistema de refrixeración dun hotel en América Latina, o que representa un estudo de caso real.

Presentaranse diferentes metodoloxías para a representación de gráficos de control aplicados a datos multivariados, incluídos procedementos válidos incluso cando os datos multivariables non cumpran as

especificacións de normalidade e independencia. Finalmente, avaliarase a aplicación destes métodos ao estudo real de casos dun gran hotel no Caribe, proporcionado pola empresa Fridama SA, incluíndo tamén un estudo exploratorio multivariante e aplicacións de métodos de clasificación non supervisados ??que proporcionan información relevante sobre o funcionamento de as instalacións de aire acondicionado (HVAC) do hotel, sempre co obxectivo último de controlar as instalacións de climatización e proporcionar un sistema de detección de anomalías que favoreza o mantemento preditivo, evitando, na medida do posible, a identificación de falsas alarmas.

Palabras chave: Control de calidade estatística, gráficos de control multivariante, eficiencia enerxética, calidade 4.0.

English abstract

Currently, the intensive use of data acquisition, sensors and the use of computers in process monitoring has led to the need for the analysis of industrial processes with two or more correlated variables, in which statistical process control and the capacity analyzes must be carried out using multivariate methodologies. This project reviews various alternatives for the statistical control of processes in the industry and services sectors from the perspective of multivariate analysis. The new tools used in the new data paradigm corresponding to Industry 4.0 will be described, that is, control charts specially designed to be applied to databases made up of a large number of variables continuously monitored over time. Programming in R and the application of these techniques to a real database referring to the cooling system of a Latin American hotel will also be proposed.

Different methodologies for the representation of control charts applied to multivariate data will be presented, including valid procedures even when the multivariate data does not meet the specifications for normality and independence. Finally, the application of these methods to the real case study of a large hotel in the Caribbean, provided by the company Fridama SA, will be evaluated, including also a multivariate exploratory study and applications of unsupervised classification methods that provide relevant information about the operation of the hotel's air conditioning (HVAC) facilities, always with the ultimate goal of controlling the HVAC facilities and providing an anomaly detection system that favors predictive maintenance, preventing as far as possible the identification of false alarms.

Key words: Statistical Quality Control, multivariate control chart, energy efficiency, quality 4.0.

Capítulo 1

Introducción

Debido al desarrollo de las tecnologías y los sensores de los últimos años, las empresas y negocios en general se encuentran con gigabytes y gigabytes de datos disponibles, y por supuesto, la gran parte de estos datos son datos multivariantes en su naturaleza, que deben de ser analizados y controlados para obtener la información más efectiva posible y fomentar así la mejora continua.

El objetivo de este trabajo es la revisión de las diversas alternativas para el control estadístico de procesos en los sectores de la industria y servicios desde la perspectiva del análisis multivariante. Se describirán las nuevas herramientas utilizadas en el nuevo paradigma de datos correspondiente a la Industria 4.0, es decir, gráficos de control especialmente diseñados para ser aplicados a bases de datos compuestas por gran cantidad de variables monitorizadas continuamente frente al tiempo. También se propondrá la programación en R y la aplicación de dichas técnicas a una base de datos real referentes al sistema de enfriamiento de un hotel sudamericano.

1.1. Antecesores en el Control de la Calidad

Hace aproximadamente una centuria e impulsado por el Dr. Walter A. Shewhart, se inició un nuevo campo de aplicación de la estadística. Este físico, ingeniero y estadístico norteamericano es considerado el padre del Control Estadístico de la Calidad. Sus teorías sobre el control de procesos y de forma especial los gráficos de control, constituyen la base para lo que hoy se conoce como Control Estadístico de Procesos.

El trabajo de Shewhart se centraba en la importancia de reducir la variación en un proceso de manufactura. Según este gurú del Control de la Calidad es fundamental comprender que el continuo proceso de ajuste para remitir las no-conformidades lleva a incrementar la variación y degradar la calidad. Shewhart enmarca el Control de Calidad en términos de variación por Causas Normales o Aleatorias y Causas Especiales o Asignables.

Shewhart no sólo creó las bases fundamentales para los gráficos de control, sino que introdujo el concepto de un "Estado bajo control" estadístico por medio de experimentos diseñados cuidadosamente". El Dr. Shewhart descubrió que la variación observada en datos de manufactura no siempre se comportaba igual que los datos en la naturaleza (movimiento Browniano de partículas).

Como conclusión, se tiene que aunque todo proceso muestra variación, algunos procesos muestran variación controlada que es natural al proceso, mientras que otros muestran variación sin control que no está presente en el sistema causal de proceso todas las veces. Este fenómeno del campo de la Estadística aplicada a la calidad introdujo las gráficas de control como herramienta para distinguir entre estos dos tipos de causas: asignables y no asignables o aleatorias.

En base a todas estas ideas, Shewhart fue autor de varios libros y publicaciones que fueron el punto de partida para la continuación del desarrollo de la teoría del Control de Calidad. Precisamente a raíz

de dichas publicaciones, fueron muchos los científicos que se pronunciaron al respecto y desarrollaron más en profundidad las ideas iniciales de Shewhart. Se deben nombrar a los llamados cinco grandes de la calidad, William Eduards Deming, Joseph M.Juran, Armand V. Feigenbaum, Kaoru Ishikawa y Phipip B. Crosby.

- William Eduards Deming

  Deming desarrolló el Control Estadístico de la Calidad, demostrando en el año 1940, que los controles estadísticos podrían ser utilizados tanto en operaciones de oficina como en las industriales. Este estadístico estadounidense popularizó una de las frases más mediáticas del Big Data, "en Dios confiamos, el resto que traiga datos".

  Durante la Segunda Guerra Mundial, Deming enseñó a los técnicos e ingenieros americanos ciertas herramientas estadísticas para la mejora de la calidad de los materiales de guerra. Sin embargo, su trabajo fue ignorado. Pocos años después, en 1950, se traslada a Japón acompañando a la misión americana para reflotar la economía japonesa de posguerra. En ese preciso momento, la industria y la economía de dicho país se encontraba en crisis. Fue entonces cuando volvió a intentar introducir las herramientas estadísticas para la mejora de la calidad a algunos administradores, ingenieros y científicos japoneses con el objetivo de producir productos y servicios con mejor calidad. En esta ocasión, aplicaron los conceptos propuestos por Deming, consiguiendo así que los japoneses mejorasen, dando un giro a su economía y productividad por completo, convirtiéndose así en los líderes del mercado mundial.

- Kaoru Ishikawa

  En 1949, Ishikawa, se vinculó a la UCIJ (Unión de Científicos e Ingenieros Japoneses), empezando a estudiar los métodos estadísticos y el control de la calidad. Este Japonés desarrollo lo que hoy en día se conocen como las 7 herramientas de Ishikawa:

  1. La Gráfica de Pareto.
  2. El diagrama de causa-efecto (Ishikawa).
  3. La estratificación.
  4. La hoja de verificación.
  5. El histograma.
  6. El diagrama de dispersión.
  7. Gráficos de Control.

- Joseph M.Juran

  En 1954, Juran visitó Japón, que se encontraba en proceso de implementar las ideas de Deming sobre el Control Estadístico de la Calidad. Jurán fue el impulsor de la idea de que el Control de la Calidad ha de ser un instrumento de la alta dirección. Con el objeto de hacer llegar dicha idea, comenzó a impartir ciertos seminarios a gerentes de empresas y altos cargos, consiguiendo hacer calar estas herramientas en el proceso de implantación del Control de la Calidad que Japón estaba llevando a cabo en esa época. La facilidad en cómo la cultura la calidad en Japón prosperara de manera tan rápida hay que atribuirla a la mentalidad de este pueblo, basado en los principios de seguir al líder y en la propia filosofía Zhen.

- Armand V. Feigenbaum

  Feigenbaum fue el fundador del concepto de Control Total de la Calidad (CTC), que define un sistema eficaz para integrar los esfuerzos en materia de desarrollo y mantenimiento de calidad, realizados por los diversos grupos de la organización, de modo que sea posible producir bienes y servicios a los niveles más económicos y que sean compatibles con la plena satisfacción de los clientes.

Con el objetivo de que la calidad no se convirtiera en tarea de todos pero de nadie al mismo tiempo, sugirió qe el Control de la Calidad debía estar respaldado por la gerencia y debería existir un grupo de personas dedicadas a ello. Es decir, Feigenbaum apostaba por un grupo de personas cuya especialización fuera la calidad de los productos y cuya única área de operaciones fuera el control de la calidad. De esta filosofía, nacen los hoy conocidos como Departamentos de Control de la Calidad.

- Phipip B. Crosby

  Por otro lado, Crosby desarrolló una teoría basándose fundamentalmente en que lo que realmente cuesta dinero son las cosas que no tienen calidad, es decir, el coste que supone no hacer las cosas bien a la primera. En base a esta idea, realizó su famosa tesis de prevención. Crosby comparte la idea de Ishikawa de que la calidad es la oportunidad y obligación de los dirigentes.

Muchas otras personas han surgido con concepciones e ideas particulares derivadas de su experiencia, pero a la vez todos coinciden en un conjunto de ideas que son básicas para que la calidad tenga un carácter total, es decir, que dicha filosofía sea impulsada por los líderes de la organización, que debe estar orientada al consumidor y que ha de ser un proceso de mejora continua, requiriendo tanto de la formación constante de los empleados como de la constante medición del proceso productivo.

1.2. Evolución Histórica de la Calidad

A lo largo de la historia, la calidad ha ido evolucionando de manera progresiva, donde se pueden observar cinco etapas muy diferenciadas:

- 1º Etapa. Desde la revolución industrial hasta 1930:

  Junto con la Revolución Industrial, también se dio la transformación del trabajo manual al trabajo mecanizado. Antes de esta revolución el trabajo era prácticamente artesanal, por lo tanto, el trabajador tenía la responsabilidad sobre la producción completa del producto.

  En base a esta forma de trabajar, a principios de 1900 surge el puesto de supervisor, que en muchas ocasiones era el propio propietario, el cual se encargaba de la supervisión de la calidad del trabajo. Uno de los grandes avances en relación al Control de la Calidad, vino con la Primera Guerra Mundial, ya que con este evento, los sistemas de fabricación fueron siendo más complejos, lo que condujo a que se creasen los primeros desempeños como inspectores de calidad, que eran personas dedicadas a tiempo completo a tareas organizativas de inspección separadas de las de producción. El objetivo principal de estos inspectores era fundamentalmente la detección de aquellos productos defectuosos para así, evitar que llegasen al punto de venta.

- 2º Etapa. 1930-1949:

  Con la Segunda Guerra Mundial, se obtuvieron grandes aportes a la tecnología, lo que conllevó a que la producción en masa se incrementase considerablemente. Dicha producción en masa fue la que requirió de Control Estadístico de la Calidad. Como se trataba de producción en masa con gran número de unidades, se comenzó a implementar la inspección por muestreo, dejando atrás la inspección al 100 %.

  En este momento surgen las normas MIL-STD, que hace referencia a los estándares militares establecidos en esta época en Estados Unidos, y eran empleadas para alcanzar los objetivos de normalización del Departamento de Defensa. Dicha normalización fueron beneficiosas en el logro de la interoperabilidad, los productos, lo que garantiza cumplan con ciertos requisitos, en común, fiabilidad, coste total de propiedad, la compatibilidad con la logística sistemas y objetivos relacionados con la defensa similares. En un inicio estas normas fueron ligadas a la industria militar, pero posteriormente serán incorporadas a las normas civiles conocidas como normas ISO.

El interés principal de esta época se caracteriza por el control que garantiza no sólo conocer y seleccionar los desperfectos o fallas de productos, sino también la toma de acciones correctivas sobre los procesos tecnológicos. En cuanto al desempeño de los inspectores de calidad también se dieron cambios importantes, ya que ahora no sólo debían realizar el control de calidad del producto final, sino que los controles que debían realizar estarían distribuidos a lo largo de todo el proceso productivo.

- 3º Etapa. 1950-1979:

  Al inicio de esta etapa, el objetivo principal del control de la calidad era que las piezas defectuosas no llegaran al cliente. Sin embargo, se inicia un proceso de cambio, en el que se busca la identificación de los requisitos de satisfacción y expectativas del consumidor. Gracias a este giro en el enfoque, se comienzan a coordinar todas las áreas organizativas en función del objetivo final, la calidad. Sin embargo, el sentimiento de vender todo lo que se producía todavía se encontraba muy arraigado.

  En esta tercera etapa, comienzan a aparecer programas y sistemas de calidad para las nuevas áreas de calidad de las empresas, donde además de la propia medición, se incorpora la planificación de la calidad en un sentido más amplio.

- 4º Etapa. Década del 80:

  En esta cuarta etapa, la Dirección Estratégica de la Calidad se centra en la elaboración de una estrategia encaminada al perfeccionamiento continuo de la calidad, involucrando a toda la empresa, y creada en función de las necesidades y expectativas de los clientes.

  La responsabilidad de la calidad recae sobre la dirección de la empresa, sin embargo, empiezan a ser conscientes de que es fundamental la participación de todos los trabajadores. En esta etapa, la calidad se ve como una oportunidad competitiva.

- 5º Etapa. 1990-2010:

  Fundamentalmente, la principal innovación de esta quinta y última etapa es que la distinción entre producto y servicio comienza a desaparecer, predominando la importancia del valor total que recibe el cliente. Esta nueva etapa es conocida como Servicio de Calidad Total.

  En los años 90, el cliente empieza a estar dispuesto sólo a pagar por lo que significa valor para él. Es por eso que la calidad es apreciada por el cliente desde dos puntos de vista, calidad perceptible y calidad actual. La primera es la clave para que la gente compre, mientras que la segunda es la responsable de lograr la lealtad del cliente con la marca y con la organización.

  La Industria 4.0 es un concepto que fue desarrollado desde el 2010 por el gobierno alemán para describir una visión de la fabricación con todos sus procesos interconectados mediante Internet de las cosas. Esta filosofía sería la que daría lugar en los próximos años a la calidad 4.0.

En la siguiente Tabla 1.1 se muestra la evolución que sufre la Calidad:

1.3. Filosofía Seis Sigma

La Metodología Seis Sigma promueve el uso de herramientas y métodos estadísticos para la mejora de procesos. Esta mejora de procesos está directamente relacionada con la reducción de la variabilidad.

El movimiento Seis Sigma y sus aplicaciones contribuyeron a crear una gran oferta para mejorar la calidad y por lo tanto la rentabilidad de muchas corporaciones multinacionales en el mundo.

La meta de Seis Sigma es llegar a un máximo de 3.4 defectos por millón de eventos u oportunidades (DPMO), entendiéndose como defecto cualquier evento en que un producto o servicio no logra cumplir los requisitos del cliente.

Cuadro 1.1: Evolución de la Calidad.

Periodo	Actividad	Esencia
1920	Inspección de la calidad	Separación de las unidades buenas de las malas
1950	Control de la calidad	Detección y prevención de los defectos en el proceso de fabricación
1970	Aseguramiento de la calidad	Incorporación del control de calidad en todas las actividades de la organización
1980	Gestión de la calidad	Integrar los esfuerzos de todos hacia el logro de la calidad
1990	Gestión total de la calidad	Extensión del logro de la calidad a todas las actividades que realiza la organización

1.3.1. Ciclo DMAIC

Para llevar a cabo la implementación de la Metodología Seis Sigma en un proceso, se debe seguir el modelo de cinco etapas denominado DMAIC, constituido por las etapas de Definición, Medición, Análisis, Mejora y Control. En la siguiente Tabla (1.2) se encuentran desarrollados los objetivos de cada una de las etapas, así como las herramientas o tareas que se suelen emplear para desarrollarlas adecuadamente:

1.3.2. Industria 4.0

En los últimos años se han dado movimientos y avances importantes en campos como la conectividad, la movilidad, el análisis de datos y también en la sistematización de procesos. Ante este nuevo escenario, nace la Industria 4.0, y con ella, la denominada Calidad 4.0.

En sí el concepto de la industria 4.0 surgió en Alemania, con idea de dar nombre a la idea de interconectar todas las partes de una empresa para dar lugar a una automatización efectiva y una empresa más inteligente. En otras palabras, se puede decir que consiste en la digitalización de la industria y todos los servicios relacionados con la empresa.

Con este nuevo concepto se quiere unificar el mundo virtual y el real, es decir, se utilizan las nuevas tecnologías en todas las partes de la empresa, incluyendo los procesos productivos. De esta forma, las instalaciones son capaces de autogestionarse de forma más autónoma adaptándose a los requisitos del mercado. La industria 4.0 tiene por pilares ciertas tecnologías que abalan este objetivo.

Dichos pilares tecnológicos sobre los que se crearán las fábricas del futuro son el Big Data y análisis de los datos, los robots autónomos, el empleo de simulaciones, sistemas para la integración vertical y horizontal, el IIoT o Internet de las cosas, la ciberseguridad, el Cloud Computing, la Fabricación Aditiva, Realidad Aumentada, Comunicación 5G, el Gemelo Digital, Edge computing y Blockchain. A modo de resumen se muestra la Figura 1.1.

Dada todas estas innovaciones tecnológicas, como ya se mencionaba previamente, se introduce también el concepto de Calidad 4.0, que muestra ciertos cambios respecto a la calidad tradicional aplicada en las industrias hasta el momento. No se trata de que la calidad 4.0 sustituya a los métodos tradicionales de calidad, sino que se debe entender como la oportunidad de mejorar lo ya existente

Tabla 1.2: Etapas de la metodología Seis Sigma.

Etapa	Objetivos	Tareas
Definir	▪ Identificar el problema. ▪ Establecer los objetivos. ▪ Establecer un mapa de procesos.	▪ Diagrama de procesos. ▪ Diagrama de Ishikawa. ▪ Diagrama de Pareto.
Medir	▪ Identificar las características clave. ▪ Desarrollar el plan de recolección de datos. ▪ Validar el sistema de medición. ▪ Toma de datos.	▪ Estudios R&R. ▪ Contrastes de bondad de ajuste.
Analizar	▪ Construir una hipótesis. ▪ Identificar posibles causas.	▪ Análisis exploratorio de datos. ▪ Inferencia Estadística. ▪ Estudios de Capacidad.
Mejorar	▪ Contrastar la hipótesis con datos experimentales. ▪ Establecer las posibles soluciones. ▪ Desarrollar el plan de implementación.	▪ Diseño de Experimentos. ▪ ANOVA. ▪ Regresión.
Controlar	▪ Analizar los datos. ▪ Establecer conclusiones. ▪ Comunicar los resultados. ▪ Preparar un plan de control y monitorización.	▪ Gráficos de Control.

mediante la digitalización de procesos y la interacción de las personas. Esto se encuentra directamente relacionado con el número de no conformidades, el análisis de datos, la interacción con los clientes, la gestión de los productos y servicios y la mejora continua. Todo ello permite un mayor rendimiento en dicha conexión integral entre las personas, las máquinas y los datos de una forma novedosa y más eficiente.

Junto con la revolución de la industria 4.0, se ha despertado cierta intranquilidad debido al miedo provocado a que dicha tecnología pueda sustituir el trabajo de las personas. Sin embargo, el objetivo de la calidad 4.0 no es gestionar el sistema de calidad a través de robots o sensores de medición prescin-

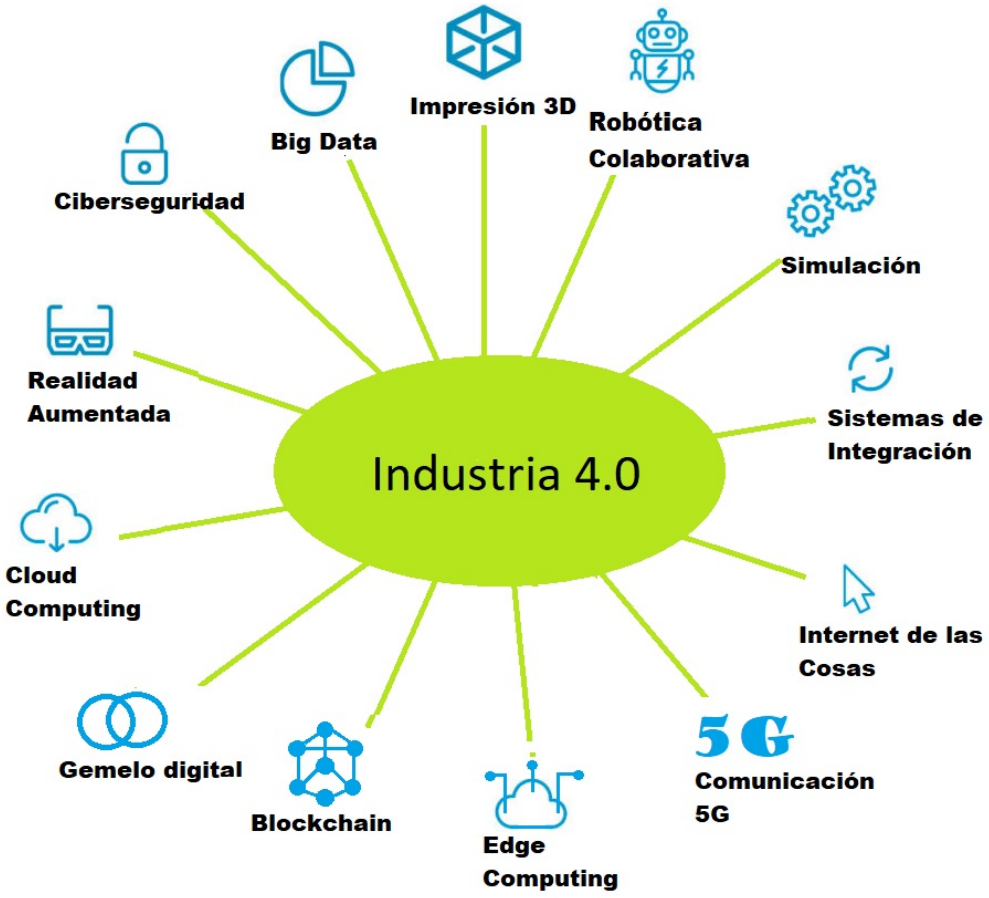


Figura 1.1: Los 13 pilares de la industria 4.0.

diendo del factor humano, sino todo lo contrario. Consiste en poner al alcance de los responsables del control de calidad una serie de instrumentos que permitan automatizar, analizar y obtener resultados de una manera más sencilla y eficiente.

A la hora de proceder a la implantación de un sistema de calidad 4.0 en una organización, se deben tener en cuenta primeramente ciertos puntos. Por un lado, se deberá analizar la situación real del problema, valorando también los beneficios que podría aportar el cambio en el desempeño empresarial. Una vez realizada esta primera valoración, resultará beneficioso procedimentar y establecer prioridades entre las iniciativas de calidad a aplicar, para no caer en el error de querer implementar todas las herramientas al mismo tiempo. Para finalmente ya proceder a la implementación, incluyendo la formación y la reeducación del personal, se debe conocer que es posible que aparezcan ciertos obstáculos como la falta de recursos y otras limitaciones anexas al propio sector industrial de la empresa.

Resulta evidente pensar que la industria 4.0, así como la calidad 4.0, conllevan una transformación que se fundamenta en el uso de tecnologías que mejoran la conectividad entre los distintos sistemas de la empresa, de manera que cualquier elemento (sensor, máquina, MES, ERP, etc.) pueda comunicarse y compartir información con el resto de componentes que integren el ecosistema de producción. Una de las consecuencias es la aparición y disponibilidad de grandes cantidades de datos, lo que requiere de conocimientos para la explotación de su explotación y conversión para la obtención de información valiosa para la toma de decisiones. Para ello, se requiere de técnicas estadísticas y modelos complejos

de aprendizaje automático o Machine Learning. Los datos proporcionados pueden utilizarse en el desarrollo de distintas aplicaciones que permitan la monitorización y control de procesos en tiempo real, la eficiencia energética, la optimización de recursos de producción, la seguridad y el mantenimiento predictivo de equipos de fabricación.

Capítulo 2

Graficos de Control

Los Gráficos de Control son la herramienta más empleada en el Control de Calidad, y en la metodología Seis Sigma se corresponde con la última fase del ciclo DMAIC, la fase de control. En esta fase, se deben monitorizar y vigilar los logros obtenidos en el resto de etapas anteriores, además de documentar las nuevas condiciones o especificaciones del proceso de estudio. Precisamente por ello, en esta etapa se requiere de herramientas basadas en la detección de errores.

2.0.1. Introducción a los Gráficos de Control

El objetivo de la mejora de la calidad no es sólo ofrecer productos de buena calidad, sino también mejorar la productividad y la satisfacción de los clientes. Una forma de mejorar la productividad es a través de la reducción de los defectos y por supuesto las revisiones, es decir, reducir las inspecciones, pero controlando al mismo tiempo los procesos en curso antes de que se generen productos o servicios defectuosos.

Al fin y al cabo, las variaciones de las características medidas son desviaciones de los objetivos prefijados. No obstante, dichas variaciones aparecen independientemente de que el proceso está o no bajo control, pudiendo diferenciar las causas de variación comunes o aleatorias y especiales o asignables. El interés está en poder predecir el nivel de calidad de los productos o servicios, y para ello, es necesario que el proceso sea estable, es decir, que las variaciones dadas en el proceso provengan de causas comunes o aleatorias.

Los gráficos de control representan la herramienta más importante en el análisis de las variaciones de los procesos de producción o servicios. Es un gráfico de dos dimensiones cuyo eje de ordenadas representa la variable monitorizada y los valores se representan secuencialmente con respecto al tiempo. Esos valores de la variable de interés pueden ser individuales o promedios referidos a grupos de cierto tamaño, lo que se denominan muestras racionales. En cuanto al eje de abscisas, este muestra los números de identificación asignados a la o las muestras. Una vez definidos los ejes, los valores de la variable se representan con puntos unidos mediante líneas rectas, favoreciendo así la identificación de patrones indicadores de cambios significativos en el rendimiento del proceso.

Los gráficos de control, a su vez, pueden ser vistos como un contraste de hipótesis. Al fin y al cabo, una hipótesis estadística es cualquier conjetura sobre una o varias características de interés de un modelo de probabilidad. En los gráficos de control, la hipótesis nula (H_0) es que el proceso está bajo control, y en consecuencia, la hipótesis alternativa (H_1) será que el proceso ha salido del control. Si se quiere verificar que el proceso se encuentra bajo control, se debe verificar para cada una de las

submuestras el correspondiente contraste.

$$\begin{aligned} H_0: &\quad \textit{El proceso está bajo control.} \\ H_1: &\quad \textit{El proceso salió de control.} \end{aligned} \tag{2.1}$$

No obstante, existe una probabilidad de que se generen falsas alarmas, es decir, que el proceso se encuentre bajo control y el gráfico sin embargo, indique lo contrario. Al igual que puede ocurrir la situación contraria, que en la realidad el proceso está fuera de control pero el gráfico no sea capaz de detectarlo. A estos errores se denominan respectivamente Error Tipo I y Error Tipo II, y pueden ser cuantificados a través de las probabilidades $\alpha = P(Error\ Tipo\ I)$ y $\beta = P(Error\ Tipo\ II)$.

De manera análoga, y partiendo del enfoque de los gráficos de control como un contraste donde la hipótesis nula indica que el proceso se encuentra bajo control, es posible designar error de tipo I o Riesgo del Vendedor o Productor a la decisión de rechazar la hipótesis nula cuando ésta es cierta. Y por el contrario, designar error de tipo II o Riesgo del Comprador o Consumidor a la decisión de no rechazar la hipótesis nula cuando es falsa. Las cuatro distintas situaciones posibles se pueden ver resumidas en la siguiente Tabla 2.1.

Tabla 2.1: Riesgos del Vendedor y el Comprador.

	H_0 es cierta	H_0 es falsa
No se rechaza H_0	Decisión correcta	Riesgo del Comprador o Consumidor
Se rechaza H_0	Riesgo del Vendedor o Productor	Decisión correcta

Muestreo de aceptación

Por norma general, si se quiere llevar a cabo un control de calidad de un proceso productivo, no se observan todas y cada una de las piezas o unidades, sino que se toma una muestra del lote para inspeccionar cualquiera que sea la característica de calidad del producto. Una vez inspeccionado, se debe sentenciar el lote, es decir aceptar o rechazarlo en función de la información de la muestra tomada.

El muestreo por aceptación suele resultar muy útil cuando la inspección es destructiva o el costo del 100 % es prohibitivo. Es decir, en general se emplea en aquellas situaciones en las que la inspección al 100 % incrementa la tasa de unidades defectuosas o incluso cuando existen riesgos potenciales de confiabilidad en el producto, y es necesario monitorearlo de manera continua.

Las unidades seleccionadas del lote para la inspección, deben ser seleccionadas al azar, con el objetivo de que éstas sean representativas de todas las unidades del lote, es decir, de la población. Para llevar a cabo el muestreo aleatorio simple propuesto anteriormente basta con asignar un número a cada unidad del lote y extraer n números al azar del rango N, siendo n el número de piezas a inspeccionar y N el número total de lotes producidos. Entióndase lote como unidad o conjunto de unidades productivas del proceso a analizar.

En el diseño de gráficos de control es necesario especificar tanto el tamaño muestral como la frecuencia del muestreo. Muestras grandes siempre hacen más fácil la detección de los pequeños cambios o variaciones en el proceso. Sin embargo, en la práctica se opta por subgrupos racionales, que implican la selección de subgrupos o muestras con el objetivo de que si existen causas atribuibles, la posibilidad de detectar diferencias entre subgrupos sea máxima, mientras que la misma posibilidad dentro de un subgrupo sea la mínima. En definitiva, normalmente se emplean submuestras pequeñas de tamaño 4 o 5 elementos, tomadas a intervalos preferiblemente cortos.

Como se puede observar, la decisión de aceptar o rechazar un lote se basa en la información recopilada en la muestra aleatoria, que muestra una información parcial respecto de toda la población, por lo que es posible cometer un cierto error en la decisión. En este contexto, se dice que existe cierto riesgo.

Como se ha indicado en el apartado previo, debido al riesgo existente a la hora de tomar la decisión y sentenciar el lote a analizar, existen dos tipos de error. Por un lado, el productor desea evitar el rechazo de un lote bueno, y por otro, el comprador desea evitar la aceptación de un lote malo, como se puede apreciar en la Tabla anterior, 2.1.

Se tienen por tanto los dos siguientes tipos de riesgos distintos:

- Riesgo del Vendedor o Productor:

  Este es el riesgo correspondiente al rechazo de un lote bueno. Generalmente es preferible aceptar lotes de este nivel, de calidad aceptable.

  - Nivel de Calidad Aceptable (AQL).

    Se trata de la definición numérica de un lote bueno, asociada con el Riesgo del Vendedor. El AQL es el máximo porcentaje de defectuosos que es permitido como un proceso promedio satisfactorio, estableciendo el Riesgo del Vendedor como la probabilidad de que un lote aceptable (es decir, un lote de calidad AQL o mejor) sea rechazado.

- Riesgo del Comprador o Consumidor:

  Este es el riesgo asociado a aceptar un lote de mala calidad, y obviamente, rara vez se desea aceptar lotes que tengan este nivel pobre de calidad.

  - Nivel de Calidad Rechazable (RQL).

    Esta es la definición numérica de un lote pobre, en relación al Riesgo del Comprador. El RQL representa la menor calidad, en porcentaje de defectuosos que se puede tolerar en un lote, definiendo el Riesgo del Consumidor como la probabilidad de que un lote de calidad rechazable sea aceptado por el plan de muestreo.

Lo ideal en un control de calidad es obtener gráficos de control con mínimos valores de α y β, algo muy complejo de lograr de manera simultánea. Una medida de rendimiento de los gráficos de control muy empleada es el ARL (Average Run Length), que es el número promedio de muestras monitorizadas hasta que el gráfico presenta una señal, o un punto fuera de los límites de control.

Se necesitaría un valor de ARL muy grande en el caso en el que el proceso se encuentra bajo control, indicando que el gráfico es menos susceptible a errores del Tipo I, ya que:

$$ARL = \frac{1}{\alpha} \tag{2.2}$$

Por otro lado, se requieren valores pequeños de ARL cuando el proceso se encuentra fuera de control, indicando que el gráfico es rápido en la detección de anomalías, de acuerdo a:

$$ARL = \frac{1}{1-\beta} \tag{2.3}$$

En este contexto, resulta muy recomendable graficar el riesgo β en función de la magnitud del cambio que se pretende estudiar, normalmente expresado en unidades de desviación estándar. A estas curvas se les denomina Curva Característica de Operación, o curva OC (Operating Characteristics). Esta herramienta, junto a algunas otras se pueden contemplar en el **??**.

Construcción por fases

La mayoría de las veces en las que se quiere obtener un proceso bajo control, se emplea el control estadístico de la calidad por Fases. Es decir, existe una primera fase, denominada fase I, que trata de realizar un análisis retrospectivo, en el que se quiere chequear que efectivamente el proceso se encuentra bajo control desde que se recogió la primera de las muestras. En esta primera fase, se requiere de un análisis exhaustivo, ya que esta fase será la que dictamine si el proceso está o no bajo control en el futuro. En definitiva, en la fase I, se estiman los límites de control utilizando una muestra preliminar, para lo que será necesario que todas las muestras se encuentren entre los límites de control, eliminando aquellas que salgan de control para la estimación de los límites de calidad que posteriormente serán empleados para la fase II de monitoreo.

En la segunda fase, fase II, se emplean gráficos de control para verificar que el proceso continua bajo control según lo que en la primera fase se estableció como controlado. En esta fase, las muestras tomadas posteriormente se representan en un gráfico con los límites de control previos, monitorizando la variabilidad del proceso en función de la media y la covarianza obtenida en la primera de las fases de control.

Cuando las observaciones individuales de la variable de estudio están dentro de los límites de control, se dice que el proceso está estadísticamente bajo control. Nótese que los límites de control son completamente diferentes de los límites de especificación (aquellos aceptados por el cliente o fijados por los ingenieros, que representan la consigna o target). Los límites de control se calculan como un intervalo de confianza. Se suelen tomar aquellos que distan de la media en tres desviaciones típicas ($\mu \pm 3\sigma$).

2.1. Gráficos de Control Univariantes

Los inicios del uso de los gráficos de control estadístico de la calidad se remontan al trabajo realizado por Walter A. Shewhart en 1920. Este pionero del control estadístico de procesos desarrolla toda su teoría suponiendo una distribución normal, donde el 99,73 % de las observaciones se encuentran entre $\pm 3\sigma$.

Como se comenta en apartados anterior, el procedimiento lógico del control de la calidad consiste en lograr que el proceso está bajo control estadístico, eliminando las causas atribuibles de variación y disminuyendo las variaciones aleatorias de modo que las mediciones de la variable de calidad produzcan valores dentro del intercalo especificado como aceptable. Esta condición se ve representada en la Figura 2.1.

Un gráfico de control es una herramienta gráfica que permite monitorizar una característica de la calidad en función del tiempo respecto a una línea central y un límite superior e inferior. De esta forma, cuando un elemento de la muestra o más quedan fuera de dichos límites, indica la presencia de causas especiales, es decir, causas no aleatorias. En consecuencia, esta causa asignable debe ser detectada y por supuesto eliminada. Por el contrario, no existen causas especiales, se dice que el proceso se encuentra bajo control.

2.1.1. Proceso "Fuera de Control"

Los gráficos de control son una herramienta sencilla, que permiten identificar una señal fuera de control. Si el proceso no tiene causas especiales que afecten su variabilidad, entonces los estadísticos de control caerán dentro de los límites de control de una forma aleatoria, o lo que es lo mismo, sin patrones evidentes.

Las causas pueden afectar la localización del proceso (promedio, mediana), la variación (rango, desviación estándar) o incluso ambos. El objetivo del análisis de una gráfica de control es identificar alguna evidencia de que la variabilidad o localización del proceso no están operando en un nivel cons-

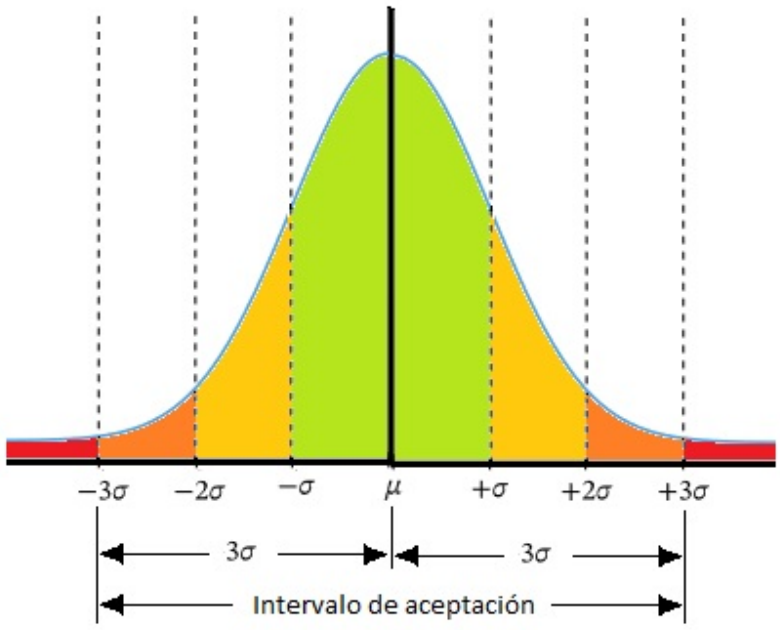


Figura 2.1: Representación del intervalo de aceptación ($\pm 3\sigma$).

tante. En otras palabras, que una de las dos características o ambas están fuera de control estadístico. Obviamente, una vez esto es identificado, resulta fundamental la toma de acciones.

Normalmente, la media o promedio será usado para el estadístico de control de localización y el rango para el estadístico de control de variación, aunque las conclusiones establecidas para estos estadísticos de control también aplican por igual a otros posibles estadísticos de control. Los estadísticos de variación y localización son analizados por separado, y es dice que un proceso no es estable (no se encuentra en control estadístico), a menos que ambas gráficas no cuenten con condiciones fuera de control (indicaciones de causas especiales).

Nótese que no todas las causas especiales son negativas, existen algunas causas especiales que pueden resultar en un mejoramiento positivo del proceso en términos de algún decrecimiento de la variación del rango. En este caso, estas causas especiales debieran ser evaluadas para una posible institucionalizacion dentro del proceso.

Como se mencionaba previamente, en cada una de las fases del control estadístico, al tomar en cuenta la variabilidad natural del proceso, proporciona elementos de juicio al trabajador para tomar decisiones sólo en casos en los que aparece una causa de variación ajena al sistema, y deja por tanto a la Dirección o Gerencia la oportunidad y la responsabilidad al mismo tiempo de innovar para efectuar cambios que reduzcan la variabilidad natural, fomentando así la mejora continua, en función de los requerimientos del consumidor final.

Las señales “Fuera de Control” se dan cuando la señal presenta cierto tipo de tendencia o patrones de comportamiento anormal en los puntos. En caso de que esto ocurra, se deben investigar las causas de dicho comportamiento, las causas especiales que la están creando, y corregirse para volver a tener el proceso “Bajo Control”.

A continuación se describen las señales propuestas por la empresa General Electric e impulsadas por **Duffy y Spatz (1959)**, de que el proceso se encuentre “Fuera de Control”:

- Punto(s) fuera de un Límite de Control:

Esta situación ocurre cuando uno o más puntos por arriba o por debajo de los límites de control. Un punto fuera de los límites de control, puede deberse a varios motivos, como los que se mencionan a continuación:

- El límite de control o punto graficado ha sido calculado o graficado erróneamente.
- La variabilidad de pieza a pieza, o la dispersión de la distribución ha incrementado, ya sea en ese punto o en parte de una tendencia.
- El sistema de medición ha cambiado.
- El sistema de medición requiere una discriminación diferente.

- Patrones o Tendencias Dentro de los Límites de Control:

  - Réplicas:

    Es la forma adquirida por los puntos cuando estos forman un grupo, bien sea por arriba o por abajo de la línea central. Y se llama longitud de la racha al número de puntos que están arriba o abajo de la línea central. Si el tamaño de la racha es de 7 puntos consecutivos o más, el proceso se debe considerar anormal.

    Una racha por encima del promedio de rango significa la existencia de una dispersión mayor, o incluso un cambio en el sistema de medición.

    Sin embargo, una racha por debajo del promedio de rango significa una menor dispersión, lo cual es bueno. De hecho, se deberá investigar e incluir la mejora. A su vez, también puede denotar un cambio en el sistema de medición. En ambos casos, la observación de rachas debe hacer saltar la alarma, motivando la búsqueda del motivo de su aparición en el gráfico de control.

  - Tendencia ascendente o descendente:

    Ocurre cuando los puntos van en secuencia ascendente o descendente. En estos casos, no existe un criterio para decidir si la tendencia es anormal o no, pero si dicha tendencia continúa, los puntos caerán fuera de los límites de control o asumirán la forma de una racha.

- Patrones no aleatorios obvios:

  - Adhesión a los límites de control:

    Cuando 2 o más puntos consecutivos caen dentro del tercio cercano a las líneas límites, se considera que el proceso es anormal. Con el objetivo de identificar esta tendencia, resulta de gran ayuda dividir el espacio comprendido entre la línea central y las líneas de control en tres partes iguales.

  - Adhesión a la línea central:

    Si los puntos se concentran en el centro, más de 2/3 partes, se juzga el proceso como anormal. Esto es considerado para rachas largas, cuando se tienen más de 25 subgrupos.

    Para determinar adhesión a la línea central, conviene como en el caso anterior, dividir el espacio entre los límites en 6 partes iguales, y observar si los puntos caen dentro de los dos sectores cercanos a la línea central. Si todos los puntos caen dentro existe la adhesión.

  - Periodicidad:

    Se dice que el proceso muestra periodicidad, si los puntos se mueven más o menos a intervalos iguales hacia arriba y hacia abajo.

A pesar de que resulta inteligente investigar todas las señales como posibles evidencias de causas especiales, se debe conocer que estas pudieron ser causadas por el propio sistema, lo que indicaría que no existe problema alguno en el proceso. En caso de que no se encuentre evidencia clara de la causa

especial, cualquier acción correctiva serviría para incrementar la variabilidad total del proceso, lo cual no es deseable, y precisamente es lo que se quiere evitar.

Asimismo, en la realidad no es posible la obtención de un proceso productivo perfecto. La meta de los gráficos de control es la obtención de un escenario razonable, donde el proceso resulte estable y económicamente aceptable. Un proceso controlado no es aquel proceso que muestra una gráfica donde nunca se sale fuera de control, ya que de ser así, el investigador se debería plantear seriamente si la operación debiera ser graficada o no, incluso se debería dudar de s las mediciones se están tomando correctamente.

A continuación, a modo de resumen, en la Tabla 2.2, se muestran las diferentes señales de "Fuera de Control", ya descritas anteriormente.

Tabla 2.2: Criterios Típicos para detectar un proceso Fuera de Control.

	Criterios típicos para Causas Especiales
1	1 punto más allá de 3 desviaciones estándar (σ) de la línea central
2	7 puntos consecutivos en el mismo lado de la línea central
3	6 puntos consecutivos, todos ellos crecientes o decrecientes
4	14 puntos consecutivos, alternando arriba y abajo
5	2 de 3 puntos $> 2\sigma$ de la línea central (mismo lado)
6	4 de 5 puntos $> 1\sigma$ de la línea central (mismo lado)
7	15 puntos consecutivos dentro de 1σ de la línea central (ambos lados)
8	8 puntos consecutivos $> 1\sigma$ de la línea central (ambos lados)

En caso de que alguna de estas situaciones se da en los gráficos de control del proceso que se está analizando, como se comentó anteriormente, se debe proceder a la identificación de causas comunes y por supuesto a su control o eliminación. A continuación se muestran tres estrategias muy empleadas en la práctica para la reducción de las causas comunes de variación:

- Estratificación:

  Se trata de examinar las diferentes características respecto la salida del proceso (día de la semana donde se dio la variación más alta, que parte es la que más variación muestra etc.), y típicamente, las gráficas de Pareto son muy útiles cuando se estratifican los datos.

- Disgregación:

  En este caso se debe proceder a la división del proceso en sus componentes y estudiar la variación en cada paso del proceso. Se relaciona muy seguido con estudios de capacidad y de rendimiento. Los Diagramas de Flujo, los histogramas, y las gráficas de pareto pueden ser de gran utilidad en este proceso de desintegrar los datos.

- Experimentación:

  Finalmente, no puede faltar la estrategia de experimentar, cambiando algunos factores en diferentes niveles y analizando los resultados y los efectos. Sin embargo, este método puede resultar

costoso y lo común es que se intente únicamente después de haber hecho la estratificación y/o la disgregación, es decir como última opción, aunque esto dependerá del propio proceso.

Para finalizar con los conceptos básicos de los gráficos de control univariantes, y retomando el concepto de error en la clasificación de los lotes seleccionados para analizar, en un principio, podría resultar lógico establecer valores muy bajos para α y β, con el objetivo de obtener un ARL grande para el proceso bajo control y un ARL pequeño para situaciones de fuera de control. El precio de llevar a cabo dicha idea sería recurrir a tamaños de muestra demasiado grandes, lo que habitualmente no es factible. Un forma de lograr la minimización de α y β a la vez es a través de optimización, **Garcia-Diaz y Aparisi (2003)**.

Sin embargo, en líneas generales, en muchos estudios de control de calidad se opta por establecer $\alpha = 0{,}0027$, que corresponde a un $ARL \approx 370$ para el proceso bajo control, indicando que serían necesarias tomar un promedio de 370 muestras para detectar una falsa alarma cuando el proceso está bajo control. Muchos otros autores también propusieron diversos valores para controlar estos falsos positivos, como **Roberts (1959)**que empleaba métodos de Monte Carlo, **Brook y Evans (1972)**, que fueron los primeros en emplear cadenas de Markov para calcular el ARL etc.

2.1.2. Gráficos de Control Paramétricos

Uno de los gráficos más empleados es el gráfico $\bar{X}$, que se basa en los intervalos de confianza para la media, que con una probabilidad de $1-\alpha$, el parámetro media pertenecerá a dicho intervalo, $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n)}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n)}}$. La idea de estos gráficos de control es estimar los límites de control con los datos de muestras seleccionadas de una determinada forma, que normalmente se basa en las ideas iniciales de Shewhart.

Definiendo $\bar{R} = \sum_{k=1}^{m} \frac{R_k}{m}$, con $R_k = max(X_k) - min(X_k)$ y fijando A_2 como una constante seleccionada en base al tamaño muestral, es posible definir el valor de la línea central y los límites del gráfico de control en cuestión.

$$CL = \bar{\bar{X}} \qquad UCL = \bar{\bar{X}} + A_2\bar{R} \qquad LCL = \bar{\bar{X}} - A_2\bar{R} \tag{2.4}$$

El gráfico $\bar{X}$ puede ser calculado tanto con R como con S, la desviación estándar.

$$CL = \bar{\bar{X}} \qquad UCL = \bar{\bar{X}} + A_2\bar{S} \qquad LCL = \bar{\bar{X}} - A_2\bar{S} \tag{2.5}$$

Sin embargo, habitualmente, se emplean ambos de manera conjunta con el objeto de monitorizar la dispersión del proceso. Obteniendo para el gráfico R:

$$CL = \bar{R} \qquad UCL = D_4\bar{R} \qquad LCL = D_3\bar{R} \tag{2.6}$$

y para el gráfico S:

$$CL = \bar{S} \qquad UCL = B_4\bar{S} \qquad LCL = B_3\bar{S} \tag{2.7}$$

En las expresiones anteriores, D_4, D_3, B_4, B_3 son constantes tabuladas en función del tamaño muestral.

2.1.3. Gráficos de Control No Paramétricos

Los gráficos de control no paramétricos que se presentan a continuación son los gráficos desarrollados por **Regina y Liu (1995)**, y son los gráficos de control r, Q, S y S^*.

Cuando los datos a tratar no cumplen el supuesto de que la variable es univariante y se distribuye bajo una distribución normal, emplear los gráficos como el gráfico de control para la media $\bar{X}$ y el gráfico CUSUM no es adecuado. En este contexto, Regina Liu propone tres tipos de gráficos de control: r, Q y S. Estos gráficos de control no paramétricos se basan en el estadístico $r_n(.)$, el cual viene definido por:

$$r_n(x) = \frac{2}{k} min(\#(x_i > x), \#x_i > x)) + \frac{\#(x_i = x)}{k} \tag{2.8}$$

donde # representa el número de observaciones. El estadístico describe directamente el rango e la variable, como una medida de centralidad, de modo que la mediana, como observación más central es la más representativa.

Gráficos de Control r

Este gráfico es similar al gráfico $\bar{X}$ para mediciones individuales. En primero lugar, para construir este tipo de gráfico no paramétrico, se requiere del cálculo de cada uno de los r_i, junto con los respectivos parámetros asociados al gráfico:

$$\begin{aligned} \textit{Línea central} &= 0.5 \\ LC &= \alpha \end{aligned} \tag{2.9}$$

Una vez construido el gráfico de control, se dirá que el proceso se encuentra fuera de control si algún $r_n(x)$ está por debajo del valor del LC. Cabe resaltar que se denota α como la proporción de alarma y que en este tipo de gráficos no existe un LCS. El estadístico de contraste $r_n(x)$ indica, como se ha dicho previamente, el rango respecto a una probabilidad, la propia cantidad de datos que están menos centrales que x.

El gráfico de control r se puede comprender como un contraste de hipótesis, donde las hipótesis nula y alternativa son:

$$\begin{aligned} H_0: &\quad \textit{La nueva observación tiene la misma distribución que la distribución de referencia.} \\ H_1: &\quad \textit{Existe un cambio en la ubicación o dispersión de la distribución de la nueva observación respecto a la distribución de referencia.} \end{aligned} \tag{2.10}$$

Gracias al cálculo de los rangos, es posible identificar cambios de localización y escala. Por lo tanto, si la dispersión de las nuevas observaciones es menor, pero la localización no se ve alterada, los rangos obtenidos serán mayores y no se detectarían anomalías en el gráfico r. Para ver la importancia de las hipótesis que deben cumplir los datos se realizarán varias comparaciones de curvas OC para distribuciones normales y exponenciales, que dejan de manifiesto la importancia del uso adecuado de los gráficos de control.

De manera rápida, es posible plantear una comparación entre el gráfico de control $\bar{X}$ de Shewhart y el gráfico de control r de Liu. Para ello, se simulan dos muestras, por un lado una de ellas con distribución normal $N(\mu, \sigma)$ y por otro, un conjunto de observaciones que siguen una distribución exponencial $E(\lambda)$. Ambas muestras se emplean para construir las curvas características de operación asociadas al gráfico de de control $\bar{X}$ de Shewhart y al gráfico de control r de Liu, con α = 0,05 y muestras de n=5 observaciones.

En el primer caso, para el gráfico de $\bar{X}$ de Shewhart, en el eje de abscisas muestra cambios en la media de las observaciones a medida que este valor aumenta $x\sigma$ unidades, es decir, que el punto (x,y) indica que si la calidad sigue una distribución $N(\mu + x\sigma, \sigma)$ en lugar de una distribución $N(\mu, \sigma)$, la probabilidad de estar dentro de los límites de control es precisamente y.

Nótese que la curva característica de operación para el gráfico de control para la media se ha obtenido según su valor exacto, mientras que en el caso del gráfico de control r, es simulada, de manera que salvo el pequeño error asociado a que no es un cálculo exacto.

Por un lado, como muestra la Figura 2.2 se refleja cómo cuando las observaciones a estudiar siguen una distribución normal, el gráfico de curva OC para ambas metodologías parece idéntica. No obstante, el gráfico r es más eficiente ya que permite observar cambios en las medias de las observaciones mientras que el gráfico X de Shewhart, no.

En la Figura 2.3 se representan las curvas características OC cuando las observaciones siguen una distribución exponencial. En este caso, es posible observar las curvas características de operación cuando las observaciones siguen una distribución exponencial. Se puede observar claramente que la

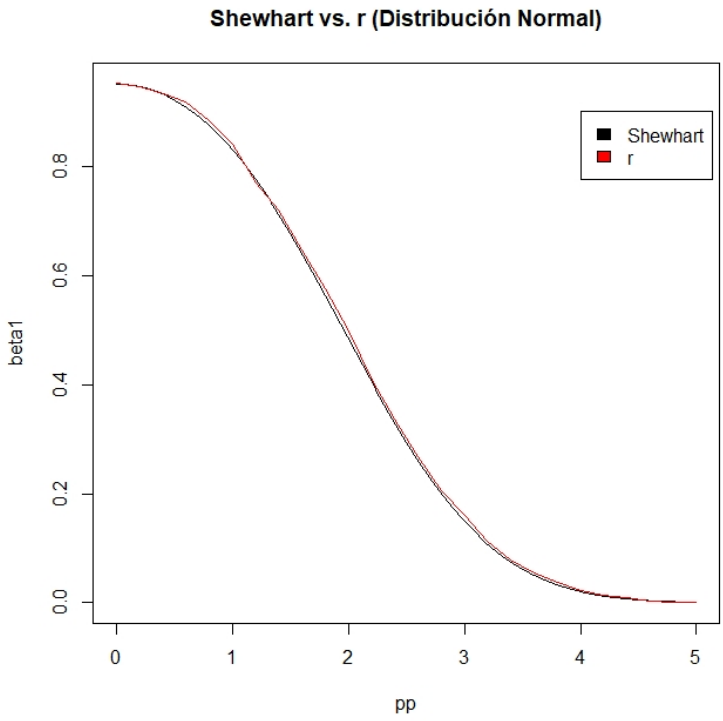


Figura 2.2: Comparación de Shewhart vs. r con datos normales.

curva para el gráfico de control X no responde correctamente ante una distribución no normal, y sin embargo, la curva para el gráfico r de Liu funciona mejor que la anterior ante la distribución exponencial. Cabe hacer una pequeña excepción, ya que presenta una pequeña anomalía, debido a que la cola de la derecha de la distribución exponencial es más larga que la cola izquierda, llevando a la curva a que muestre un valor de la probabilidad superior en x=0.2 que en x=0.

Gráficos de Control Q

El gráfico de control Q se puede considerar como el gráfico de control no paramétrico análogo al gráfico de control paramétrico $\bar{X}$. Al igual que en el caso anterior, para proceder a la construcción de este gráfico es necesario la obtención de los estadísticos $r_n(x)$ a partir de los valores de las medias muestrales y posteriormente, se establecen los valores de los parámetros necesarios para la construcción del gráfico. Sin embargo, se deben diferenciar tres situaciones posibles, para las que los parámetros se calcularán de manera diferente.

- En primer lugar, si el tamaño muestral es mayor o igual a 5 y los parámetros asociados a la distribución de los datos son conocidos, se calcularán los parámetros gráficos como:

$$\begin{aligned} &\textit{Línea central} = 0.5 \\ &LC = 0.5 - z_\alpha (12n)^{1/2} \end{aligned} \tag{2.11}$$

- Si por el contrario, el tamaño muestral es mayor o igual a 5 pero en este caso, los parámetros

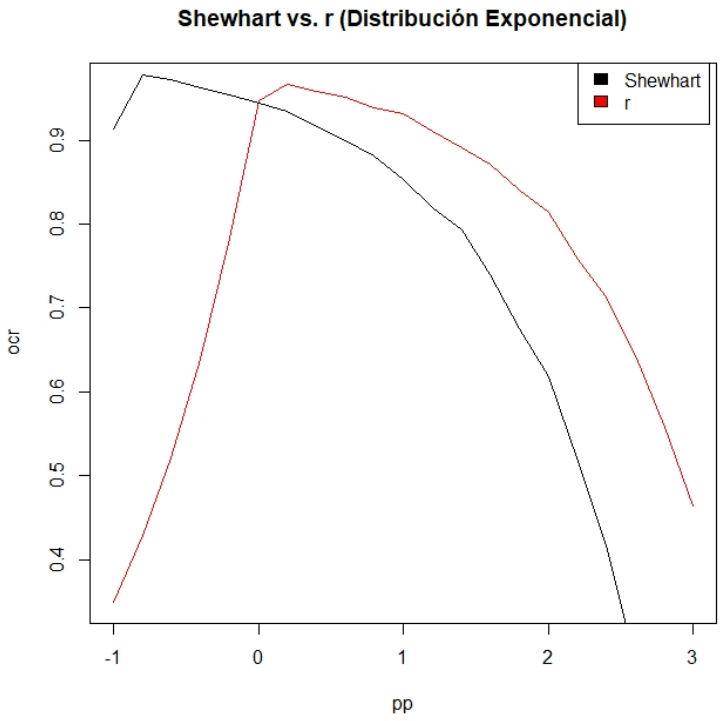


Figura 2.3: Comparación de Shewhart vs. r con datos exponenciales.

asociados a la distribución de los datos son desconocidos, los parámetros gráficos serán:

$$\begin{aligned} &\textit{Línea central} = 0.5 \\ &LC = 0.5 - z_\alpha \sqrt{\frac{\frac{1}{n}+\frac{1}{k}}{2}} \end{aligned} \tag{2.12}$$

- Finalmente, si se da el caso en el que el tamaño muestral es inferior a 5 y además el valor de α es pequeño, los parámetros gráficos serán:

$$\begin{aligned} &\textit{Línea central} = 0.5 \\ &LC = \frac{(n!\alpha)^{1/n}}{n} \end{aligned} \tag{2.13}$$

Como para el resto de gráficos de control, si algún $r_n(x)$ está por debajo del valor del LC, se declara que el proceso está bajo control.

De manera rápida es posible plantear una comparación entre el gráfico de control $\bar{X}$ de Shewhart y el gráfico de control Q de Liu. Para ello, se simulan dos muestras, por un lado una de ellas con distribución normal $N(\mu, \sigma)$ y por otro, un conjunto de observaciones que siguen una distribución exponencial $E(\lambda)$. Ambas muestras se emplean para construir las curvas características de operación asociadas al gráfico de de control $\bar{X}$ de Shewhart y al gráfico de control Q de Liu, con $\alpha = 0{,}05$ y muestras de n=5 observaciones.

Por un lado, como muestra la Figura 2.4 es posible observar las curvas características de operación cuando las observaciones se distribuyen bajo una normal. Cabe decir que no se hallan grandes

diferencias entre el gráfico de control Q y el gráfico para la media paramétrico, pero sin embargo, el segundo es más efectivo ya que detecta variaciones más pequeñas en la media de las observaciones de una forma más rápida, obteniéndose una curva OC más próxima a la ideal.

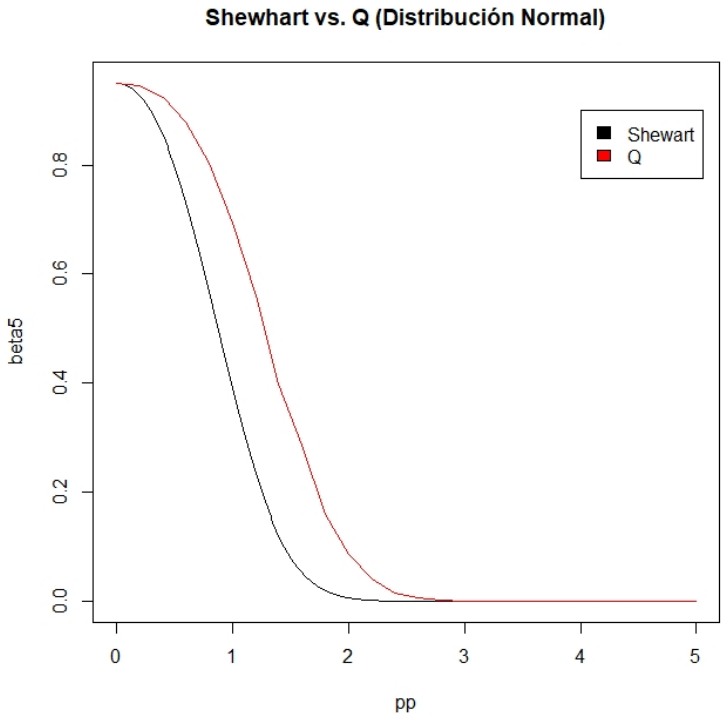


Figura 2.4: Comparación de Shewhart vs. Q con datos normales.

En la Figura 2.5 se representan las curvas características OC cuando las observaciones siguen una distribución exponencial. En este caso, es posible observar inmediatamente la diferencia existente entre ambos métodos, ya que la curva para el gráfico de control de Shewhart no responde bien ante una distribución que no sea normal mientras que la curva para el gráfico Q de Liu funciona mejor que la anterior ante la distribución exponencial. No obstante, existe una pequeña anomalía, ya que se ve que el valor de la probabilidad es mayor cuando x=0.2 que cuando x=0, pero este suceso viene motivado simplemente porque la cola a la derecha de la exponencial es más larga que la cola de la izquierda.

Gráficos de Control S y S^*

En cuanto al gráfico de control S y S^*, es un gráfico de control no paramétrico basado en el gráfico de control paramétrico CUSUM de sumas acumuladas. El primer paso consiste en, una vez más, hallar el estadístico $r_n(x)$ para cada una de las mediciones de la muestra y se calcula posteriormente el valor del S_n como se muestra a continuación:

$$S_n = \sum_{i=1}^{n} \left[r(X_i) - \frac{1}{n} \right] \tag{2.14}$$

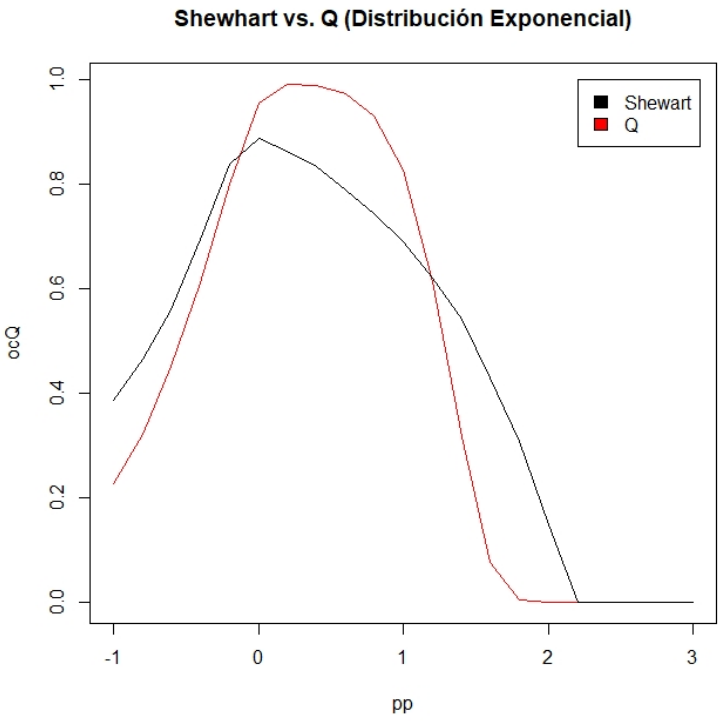


Figura 2.5: Comparación de Shewhart vs. Q con datos exponenciales.

junto con los valores de la línea central y el LC:

$$\begin{aligned} &\textit{Línea central} = 0 \\ &LC = -\left(z_\alpha \left(\tfrac{n}{12}\right)^{1/2}\right) \end{aligned} \tag{2.15}$$

En situaciones donde el tamaño muestral es muy grande, es preferible emplear el gráfico de control S^*, y de manera similar al caso anterior del gráfico S, se deben calcular los $S_n^*(G)$:

$$S_n^* = \frac{S_n}{\sqrt{\frac{n}{12}}} \tag{2.16}$$

En este tipo de gráficos, como el tamaño muestral es muy grande, los cálculos para los parámetros gráficos se ven simplificados:

$$\begin{aligned} &\textit{Línea central} = 0 \\ &LC = -z_\alpha \end{aligned} \tag{2.17}$$

2.2. Gráficos de control multivariantes

La palabra multivariante no solo significa muchas variables, sino que dichas variables puede que están relacionadas. Los métodos estadísticos por su lado se refieren al conjunto de técnicas y procedimientos de análisis de los datos, así como interpretación, representación y toma de decisiones en

base a los mismos. La fundación de los métodos estadísticos multivariantes fue desarrollada de manera gradual, empezando a principios del siglo XX.

La gran mayoría de los métodos estadísticos que se emplean en el control ingenieril de procesos industriales son métodos estadísticos univariantes.

2.2.1. Distribución normal multivariante

El análisis estadístico multivariante se suele desarrollar en un escenario donde las distribuciones de las muestras es normal multivariante. Esta aproximación viene motivada por el teorema central del límite. Recordando la función de densidad para variables aleatorias unidimensionales con distribución normal de media μ y desviación típica σ:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left[\frac{(x-\mu)}{\sigma}\right]^2}{2}} \tag{2.18}$$

donde $-\infty < x < \infty$. Transformando $\left[\frac{(x-\mu)}{\sigma}\right]^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$, para el caso multivariante, donde $p \geq 2$, se puede obtener la generalización $(x-\mu)'(\Sigma)^{-1}(x-\mu)$ conocida como la distancia de Mahalanobis, en la cual μ es el vector de valores esperados

$$\mu' = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_p] \tag{2.19}$$

y Σ es la matriz de varianzas-covarianzas de dimensión $p \times p$,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \tag{2.20}$$

es posible obtener la expresión correspondiente a la densidad normal multivariante

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{(x-\mu)'(\Sigma)^{-1}(x-\mu)}{2}} \tag{2.21}$$

donde $-\infty < x_i < \infty$.

2.2.2. Estructura de los datos

En el caso de datos multivariantes, se tienen k muestras de tamaño n con p variables o características de la calidad. El elemento x_{ijk}, por lo tanto, indicará la i-ésima observación de la j-ésima variable en la k-ésima muestra. En la práctica, los parámetros media μ y desviación típica σ son desconocidos, por lo que deben ser estimadas por $\bar{\bar{x}}$ y S respectivamente ($\bar{\bar{x}}_j = \frac{\sum_{k=1}^{m} x_{\bar{j}k}}{m}$). De manera que los datos muestran una estructura como la que se muestra a continuación:

$$\begin{array}{cc} & \begin{array}{cccccccccc} & & 1 & & & & 2 & & & \cdots & & p & & \\ \end{array} \\ \begin{array}{c} \\ Muestra \\ (k) \\ \\ \end{array} & \begin{array}{cccccccccc} 1 & x_{111} & x_{211} & \cdots & x_{n11} & x_{121} & x_{221} & \cdots & x_{n21} & \cdots & x_{1p1} & x_{2p1} & \cdots & x_{np1} \\ 2 & x_{112} & x_{212} & \cdots & x_{n12} & x_{122} & x_{222} & \cdots & x_{n22} & \cdots & x_{1p2} & x_{2p2} & \cdots & x_{np2} \\ \vdots & & \vdots & & & \vdots & & & \ddots & & \vdots & & & \\ m & x_{11m} & x_{21m} & \cdots & x_{n1m} & x_{12m} & x_{22m} & \cdots & x_{n2m} & \cdots & x_{1pm} & x_{2pm} & \cdots & x_{npm} \end{array} \end{array} \tag{2.22}$$

Para la estimación de la media de cada una de las variables en cada una de las muestras se empleará la expresión

$$\bar{x}_{jk} = \frac{\sum_{i=1}^{n} x_{ijk}}{n} \tag{2.23}$$

Para la estimación de la desviación típica, se empleará la cuasi-varianza muestral obteniendo la matriz

$$S = \begin{pmatrix} \bar{S}_1^2 & \bar{S}_{12}^2 & \cdots & \bar{S}_{1p}^2 \\ \bar{S}_{12}^2 & \bar{S}_2^2 & \cdots & \bar{S}_{2p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{S}_{1p}^2 & \bar{S}_{2p}^2 & \cdots & \bar{S}_p^2 \end{pmatrix} \tag{2.24}$$

donde los elementos diagonales son las varianzas asociadas a las p características y los elementos que quedan fuera e la diagonal son las estimaciones de las covarianzas. Para su construcción se emplearán las siguientes expresiones:

$$\bar{S}_j^2 = \frac{\sum_{k=1}^{m} S_{jk}^2}{m} \quad con \quad \bar{S}_{jk}^2 = \frac{\sum_{i=1}^{n} (x_{ijk} - \bar{x}_{jk})^2}{n-1} \tag{2.25}$$

y

$$\bar{S}_{jl} = \frac{\sum_{k=1}^{m} S_{jlk}}{m} \quad con \quad S_{jlk} = \frac{\sum_{i=1}^{n} (x_{ijk} - \bar{x}_{jk})(x_{ijk} - \bar{x}_{lk})}{n-1} \tag{2.26}$$

2.2.3. Gráficos de control para datos multivariantes paramétricos

Gráfico de control χ^2

La función de densidad normal multivariante viene descrita por un elipsoide, centrado en el vector de medias y con los ejes en la dirección de los autovectores de la matriz de varianzas-covarianzas, estableciendo μ como el punto de origen y longitud $\pm c\sqrt{\lambda_j} e_j$, siendo

$$(x-\mu)'\Sigma^{-1}(x-\mu) = c^2. \tag{2.27}$$

Si x sigue una distribución normal multivariante $N_p(\mu, \Sigma)$, $(x-\mu)'\Sigma^{-1}(x-\mu)$ sigue una distribución $\chi^2_{\alpha,p}$. Por lo tanto, se cumplirá que

$$(x-\mu)'\Sigma^{-1}(x-\mu) \leq \chi^2_{\alpha,p}. \tag{2.28}$$

En otras palabras, este método consiste en dibujar el elipsoide, pudiendo interpretar dicho área como la región de confianza y tomando los puntos de la muestra que queden fuera de esta como evidencias de causas especiales. No obstante, este método resulta complejo cuando el número de parámetros es superior a $p > 2$.

Gráfico de control T^2 de Hotelling

El método gráfico de control de la calidad T^2 fue desarrollado por **Hotelling (1947)**, sus primeras aplicaciones se dieron en la segunda guerra mundial. El método fue desarrollado en 1947 y en la actualidad sigue siendo uno de los métodos más aplicados en el control de calidad de datos multivariantes. Se trata de una metodología para datos multivariantes análoga al control gráfico de Shewhart $\bar{X}$.

Partiendo de que en la práctica la media μ y la varianza Σ son desconocidas y por tanto, han de ser estimadas mediante $\bar{x}$ y S respectivamente y la generalización de la normalidad multivariante del estadístico t:

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}, \tag{2.29}$$

aplicando

$$t^2 = \frac{(\bar{x} - \mu)^2}{\frac{S^2}{n}} = n(\bar{x} - \mu)(S^2)^{-1}(\bar{x} - \mu) \tag{2.30}$$

se obtiene la siguiente generalización:

$$T^2 = (\bar{X} - \bar{\bar{X}})'(S)^{-1}(\bar{X} - \bar{\bar{X}}) \tag{2.31}$$

donde $\bar{X}$ hace referencia al vector de medias y S es la matriz de varianzas-covarianzas estimadas. Nótese también que T^2 mide al fin y al cabo la distancia, concretamente la distancia de Mahalanobis, entre el vector de medias esperado (parámetro del proceso) y el vector de medias observado (medias aritméticas de la muestra) tomando en cuenta su matriz de varianzas-covarianzas.

El estadístico T^2 sigue una distribución F con grados de libertad p y $(mn - m - p + 1)$. Además, se supone que la estimación de la matriz de varianza-covarianzas S es no singular y que la matriz de datos X es completa, es decir, que no existen datos faltantes [**Ferrer et al. (2007)**].

Sin embargo, pueden darse ocasiones en las que resulta complicado o incluso imposible registrar todos los datos del proceso, y que la estructura de correlación entre las variables es muy fuerte, lo que hace que los supuestos anteriores no se cumplan. Ante estos escenarios, se suelen aplicar los métodos de proyección sobre estructuras latentes, ya que gracias a la aplicación de un PCA (Principal Components Analysis), es posible la reducción de la dimensión de los datos en unas pocas variables latentes ortogonales entre si, donde quedaría resuelto el problema de datos faltantes [**MacGregor y Kourti (1995)**].

El criterio consiste en comparar por lo tanto la distancia de Mahalanobis descrita previamente con los límites de control establecidos, si la distancia resulta mayor que el límite de control establecido, se concluirá que el proceso podría estar fuera de control. Entonces, para un control por fases, el limite superior UCL para la Fase I del control de calidad sería:

$$UCL = \frac{p(m-1)(n-1)}{(mn - m - p + 1)} F_{\alpha,p,mn-m-p+1} \tag{2.32}$$

Mientras que para la fase II el límite correspondiente se establece como:

$$UCL = \frac{p(m+1)(n-1)}{(mn - m - p + 1)} F_{\alpha,p,mn-m-p+1}. \tag{2.33}$$

De acuerdo a **Lowry y Montgomery (1995)**, la aplicación de este método requiere del cumplimiento de ciertos requisitos. Por un lado el número de variables o características de la calidad han de ser entre 2 a 10, y por otro, se deben tomar como mínimo 20 muestras, de tamaño 2-10 cada una. Obviamente estos requerimientos pueden ser difíciles de cumplir dependiendo de la propia naturaleza del proceso.

Descomposición En los métodos gráficos empleados para el control de la calidad, cuando uno de los datos se encuentra fuera de los límites de control, se dice que existen evidencias de que el proceso ha sufrido un efecto no aleatorio, es decir una causa especial. En el caso del control de calidad multivariante, el proceso solo depende de una variable por lo que se le atribuirá todo el peso a dicha variable, pero sin embargo, cuando existen diferentes variables, es de vital importancia poder identificar cuál es la fuente de esa causa especial que lleva a que el proceso se encuentre fuera de control.

Por lo tanto, para el caso de datos multivariantes, se emplean métodos de descomposición, cuyo objetivo es la identificación de la variable responsable de varianza atribuible a causas especiales. No obstante, a lo largo de la historia se han propuesto distintos métodos para dicha identificación, como los límites de Bonferroni propuestos por **Alt (1985)**, **Murphy (1987)**, **Doganaksoy et al. (1991)**y **Wierda (1994)**. Para la descomposición, el método propuesto por **Mason et al. (1996)** resulta el más empleado y se conoce también como la descomposición MYT.

La metodología que se debe seguir es sencilla. Partiendo de que la Fase I se encuentra bajo control, se emplea el vector de medias y la matriz de varianzas-covarianzas de dicha fase para obtener el

estadístico T^2 correspondiente.
En este punto, comienza la descomposición MYT, que consiste en:

1. Calcular los valores del estadístico T^2 correspondientes a cada una de las variables de manera independiente.
$$T_j^2 = \frac{n(x_j - \bar{x_j})^2}{S_j^2} \tag{2.34}$$
donde $\bar{x_j}$ es la media y S_j^2 es la varianza de la j-ésima variable.
2. Realizar la comparación correspondiente para los elementos de la Fase II de acuerdo a la expresión:
$$UCL = \frac{p(m+1)(m-1)}{m(m-p)} F_{\alpha,p,m-p} \tag{2.35}$$
3. Excluir aquellas variables que cumplan que $T_j^2 > UCL$.
4. Construir el estadístico T^2 para las combinaciones de las variables que permanecen sin excluir.
5. Excluir aquellas variables cuyo estadístico T^2 excede el limite.
6. Realizar este procedimiento hasta llegar a la última de las combinaciones posibles, que incluirá todas y cada una de las variables de calidad del estudio.

Caso particular En ocasiones, puede ocurrir que debido a la naturaleza del proceso, solo se pueda medir una observación por cada intervalo de tiempo. Esto implica que el proceso solamente conste de una observación por variable, es decir, $n = 1$. El estadístico T^2 de Hotelling muestra una pequeña modificación tal que
$$T^2 = (X - \bar{X})'(S)^{-1}(X - \bar{X}) \tag{2.36}$$
y como precisamente, n=1, el límite superior en la Fase I queda simplificado como se muestra a continuación:
$$UCL = \frac{(m-1)^2}{m} \beta_{\alpha,p/2,(m-p-1)/2} \tag{2.37}$$
donde β es la distribución beta con $p/2$ y $(m-p-1)/2$ grados de libertad. De manera análoga para el límite superior en la Fase II:
$$UCL = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha,p,m-p} \tag{2.38}$$

Debido a la falta de subgrupos en este caso particular donde existe una única observación, surgen dificultades para calcular la matriz de varianzas-covarianzas S, no obstante existen metodologías alternativas para calcular dicha matriz, como por ejemplo, la propuesta por **Sullivan y Woodall (1996)**:
$$S_{sw} = \frac{\sum_{k=1}^{m}(x_i - \bar{x})(x_i - \bar{x})'}{m-1} \tag{2.39}$$

Otro método de estimación de la matriz de varianzas-covarianzas en este caso es el de **Holmes et al. (1993)**, que plantean:
$$S_{hm} = \frac{\begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_m - x_{m-1} \end{bmatrix} \begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_m - x_{m-1} \end{bmatrix}'}{2(m-1)} \tag{2.40}$$

No obstante, en la práctica, los dos métodos presentan resultados muy similares.

Gráfico de control de la varianza generalizada

Al igual que en el caso univariante resultaba muy útil visualizar la media del proceso junto con la dispersión, en el caso multivariante, resulta muy útil monitorizar la variabilidad. Para poder monitorizar simultáneamente la variabilidad del proceso, existen diversos métodos, aunque el de la varianza generalizada es el más empleado en la actualidad. El término de varianza generalizada hace referencia al determinante de la matriz de varianzas-covarianzas. Su objetivo, por lo tanto es graficar los resultados empleando el determinante de la matriz de varianzas-covarianzas a lo largo de los límites inferior y superior naturales.

Si la matriz de covarianzas Σ es conocida, los parámetros necesarios para obtener la gráfica correspondiente a la variabilidad generalizada son los siguientes:

$$UCL = \left|\Sigma\right| (b_1 + 3b_2^{1/2}) \tag{2.41}$$

$$CL = b_1 \left|\Sigma\right| \tag{2.42}$$

$$LCL = max \begin{cases} |\Sigma| (b_1 - 3b_2^{1/2}) \\ 0 \end{cases} \tag{2.43}$$

donde

$$b_1 = \frac{1}{(n-1)^{2p}} \prod_{j=1}^{p} (n-j) \tag{2.44}$$

$$b_2 = \frac{1}{(n-1)^{2p}} \prod_{j=1}^{p} (n-j) \left\{ \prod_{i=1}^{p} (n-i+2) - \prod_{i=1}^{p} (n-i) \right\} \tag{2.45}$$

Nótese que el parámetro n debe ser superior al número de variables p.

En ocasiones la matriz Σ suele ser estimada con S, matriz definida-positiva, basándose en la relación $|S| = b_1 |\Sigma|$. De acuerdo a dicha relación, los parámetros gráficos quedan de la siguiente manera:

$$UCL = \frac{|S|}{b_1} (b_1 + 3b_2^{1/2}) \tag{2.46}$$

$$CL = |S| \tag{2.47}$$

$$LCL = max \begin{cases} \frac{|S|}{b_1} (b_1 - 3b_2^{1/2}) \\ 0 \end{cases} \tag{2.48}$$

Gráfico de control de media móvil exponencialmente ponderada multivariante

Este método, conocido también como MEWMA (Multivariate Exponentially Weighted Moving Average Control Chart), no es más que la extensión natural multivariante de los gráficos de control de media móvil exponencialmente ponderada univariante, propuesto por **Roberts (1959)**. Fue introducida por **Lowry et al. (1995)** y es un método muy sensible a la detección de cambios no aleatorios del proceso, y se basa en el principio de la media ponderada de los vectores previamente ya observados.

En este caso bajo el supuesto de $n = 1$, el estadístico es

$$T^2 = Z_i' \Sigma_{Z_i}^{-1} Z_i > h \tag{2.49}$$

donde

$$Z_i = \lambda X_i + (1 - \lambda) X_{i-1}. \tag{2.50}$$

Por lo tanto, si $Z_0 = 0$, resulta que λ es una matriz diagonal $p \times p$ de la constante de suavización con $0 < \lambda_i \leq 1$, aunque en realidad en la práctica no existe razón alguna para emplear diferentes valores de λ. De hecho, el valor más usual para λ es 0.1. En caso de que $n > 1$, bastaría con emplear $\bar{X}_i$ en lugar de X_i en las expresiones anteriores.

Lowry et al. (1995) propone dos alternativas para el cálculo de Σ_Z, la matriz de varianzas-covarianzas:

$$\Sigma_{Z_i} = \frac{\lambda \left[1 - (1 - \lambda)^{2i}\right]}{2 - \lambda} (\Sigma) \tag{2.51}$$

Y para el cálculo de la matriz de varianzas-covarianzas asintótica:

$$\Sigma_{Z_i} = \frac{\lambda}{2 - \lambda} (\Sigma) \tag{2.52}$$

La primera opción, el cálculo exacto de la matriz Σ, muestra un mejor comportamiento en la práctica. Los gráficos de control de media móvil exponencialmente ponderada multivariante puede verse como los gráficos de control T^2 de Hotelling cuando el parámetro $\lambda = 1$.

Gráfico de control suma acumulativa

El gráfico de control MCUSUM (Multivariate Cumulative Sum Control Chart) es la extensión al caso multivariante del método CUSUM propuesto por **Page (1993)**. Está enfocado en la mejora de la sensibilidad del gráfico de control T^2 para la detección de pequeñas variaciones del proceso, basándose en el principio de acumulación de la información de las observaciones de la muestra.

Para la construcción de este tipo de gráficos existen cuatro métodos.

- La primera de las metodologías fue introducida por **Woodall y Ncube (1985)**, y trata de monitorizar de manera individual el vector de medias empleando gráficos univariantes CUSUM. De manera análoga al procedimiento CUSUM, son gráficos bilaterales en los que los estadísticos son:

$$S_{i,j}^- = min \left\{ \begin{array}{c} 0 \\ S_{i-1,j}^- + \frac{\bar{X}_{i,j} - \mu_{0,j}}{\sigma_{0,j}/\sqrt{n}} + k_j^- \end{array} \right\} \tag{2.53}$$

$$S_{i,j}^+ = min \left\{ \begin{array}{c} 0 \\ S_{i-1,j}^+ + \frac{\bar{X}_{i,j} - \mu_{0,j}}{\sigma_{0,j}/\sqrt{n}} + k_j^+ \end{array} \right\} \tag{2.54}$$

  donde $\mu_{0,j}$ es el j-ésimo elemento del vector de medias μ, $\sigma_{0,j}$ es el j-ésimo elemento de la diagonal de la matriz Σ y k es una constante. Nótese que cuando $i = 1$, $S_{i,j}^- = 0$ y $S_{i,j}^+ = 0$. Concluyendo que en este caso, los límites son

$$UCL = h_j^+ \qquad LCL = h_j^+. \tag{2.55}$$

- **Healy (1987)** propuso un procedimiento para detectar pequeñas variaciones del proceso basándose en la combinación lineal de las variables.

$$S_i = max \left\{ \begin{array}{c} 0 \\ S_{i-1} + a' \quad \bar{X}_i - k \end{array} \right\} \tag{2.56}$$

$$a' = \frac{(\mu_1 - \mu_0)'(\frac{\Sigma_0}{n})^{-1}}{\left[(\mu_1 - \mu_0)'(\frac{\Sigma_0}{n})^{-1}(\mu_1 - \mu_0)\right]^{1/2}} \tag{2.57}$$

$$k = 0.5\frac{(\mu_1 - \mu_0)'(\frac{\Sigma_0}{n})^{-1}(\mu_1 - \mu_0)}{\left[(\mu_1 - \mu_0)'(\frac{\Sigma_0}{n})^{-1}(\mu_1 - \mu_0)\right]^{1/2}} \tag{2.58}$$

de manera que, $UCL = h$.

- **Crosier (1988)** presentó dos métodos, el primero de ellos procede con el siguiente estadístico

$$T_i^2 = \left[S_i'\left(\frac{\Sigma}{n}\right)S_i\right]^{1/2} > h \tag{2.59}$$

donde

$$S_i = \begin{cases} 0 & si \quad C_i \leq k \\ (S_{i-1} + \bar{X}_i - \mu_0)(1 - \frac{k}{C_i} & si \quad C_i > k) \end{cases} \tag{2.60}$$

donde $S_0 = 0$, k¿0 y

$$C_i = \left[(S_{i-1} + \bar{X}_i - \mu_0)'(\frac{\Sigma_0}{n})^{-1}(S_{i-1} + \bar{X}_i - \mu_0)\right]^{1/2}, \tag{2.61}$$

y por lo tanto, el límite es $UCL = h$.

- Finalmente, **Pignatiello y Runger (1990)** propusieron otros dos gráficos MCUSUM, siendo el método siguiente el que mejor resultados ofrece. El estadístico de contraste correspondiente es

$$T_i^2 = max\begin{cases} 0 \\ \left[S_i'(\frac{\Sigma}{n})^{-1}S_i\right]^{1/2} - kn_i \end{cases} \tag{2.62}$$

donde

$$S_i = \sum_{j=i-n_i+1}^{i}(\bar{X}_i - \mu_0) \tag{2.63}$$

$$n_i = \begin{cases} n_{i-1} + 1 & si \quad T_{i-1}^2 > 0 \\ 1 & - \end{cases} \tag{2.64}$$

Por lo que el límite superior del gráfico de control será $UCL = h$.

Método de proyección por PCA

El análisis de componentes principales (PCA) es una técnica multivariante enfocada en la transformación ortogonal de una base de datos correladas, es decir, la obtención de componentes principales tiene por objetivo la reducción de la dimensionalidad estableciendo dichas componentes como combinación lineal de las variables originales.

El PCA se basa principalmente en dar una aproximación a la matriz de datos $X = [x_1, x_2, \cdots, x_m]$ en términos de producto de dos pequeñas matrices denotadas por $T = [t_1, t_2, \cdots, t_m]$, puntuaciones o scores, y $P = [p_1, p_2, \cdots, p_m]$, pesos o loadings. Estas dos matrices son capaces de capturar el patrón

esencial de los datos de X siempre y cuando exista una alta correlación. La matriz T, por su lado, muestra el patrón de los individuos de la matriz X en las nuevas coordenadas, mientras que por otro lado, la matriz P capta los pesos de cada una de las variables en las componentes principales. Dichas matrices, T, P y la propia matriz de datos X vienen relacionadas de la siguiente manera:

$$T = X'P \tag{2.65}$$

es decir, las componentes principales serán las nuevas variables definidas por:

$$t_j = X'p_j \qquad j = 1, 2, \cdots, m. \tag{2.66}$$

De manera que la primera componente principal t_1 viene calculada empleando aquel vector p que maximiza la varianza de la matriz T, bajo la condición de $p'p = 1$, que parte del cálculo de autovalores y autovectores de la matriz de varianza-covarianzas de X. Siendo $S = var(X)$ la matriz de varianza-covarianzas de X, y como $S \geq 0$ y simétrica, la diagonalización de S viene dada por:

$$S = P\Lambda P' \tag{2.67}$$

donde $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$ es una matriz diagonal de misma dimensión que S y recoge los autovalores de S y P es ortogonal, por lo que $PP' = P'P = I$ cuyas columnas son los autovectores de S,$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$.

Se supone que la matriz X es centrada, es decir que las variables fueron restadas por su propia media, lo que conlleva a que para las k primeras componentes principales, la matriz X puede descomponerse de la siguiente forma:

$$X = \sum_{j=1}^{k} t_j p_j' + E = TP' + E \tag{2.68}$$

donde E es la matriz que contiene la variabilidad no explicada por el modelo PCA, es decir, la parte de varianza correspondiente a las $t_{m-k} = (t_{k+1}, \cdots, t_m)$ componentes principales descartadas del modelo.

En resumen, el algoritmo del PCA consiste en la descomposición de la matriz X en valores singulares (SVD) mediante el NIPALS (Non-Linear Iterative Partial Least Squares) **Wold et al. (1978)**.

- Descomposición de la matriz X en valores singulares (SVD)

  Este método empleado para el cálculo de scores y loadings (puntuaciones y pesos) del modeo PCA consiste en aplicar la siguiente expresión:

  $$SVD(X) = USV' \tag{2.69}$$

  donde V es la matriz de los autovectores de X, es decir, la matriz de los pesos o loadings, y S es la matriz diagonal que contiene las raíces cuadradas de los autovalores de la matriz X. La matriz U, es aquella matriz que tiene por columnas los scores o puntuaciones.

- Algoritmo NIPALS

  Se trata de un método sencillo y alternativo para el cálculo de los scores en el análisis de componentes principales. Consiste en un procedimiento iterativo que realiza la búsqueda de aquellos loadings que maximizan la varianza de los scores correspondientes, siguiendo los siguientes pasos:

  1. $t_inicial = x_j$
  2. $p' = \frac{t'X}{t't}$
  3. Normalizar mediante: $p = \frac{p}{\|p\|}$
  4. $t = \frac{Xp}{p'p}$
  5. Se debe asumir convergencia si el valor de t no cambia significativamente respecto al valor calculado en la iteración anterior. Si se obtiene convergencia, se debe parar, mientras que en caso contrario, se debe volver al paso número 2.

Propiedades de las componentes principales

- Las componentes principales tienen varianza decreciente

$$\begin{aligned} var(t_1) &= var(Xp_1) = p_1' S p_1 = \lambda_1 p_1' p_1 = \lambda_1 \\ var(t_2) &= var(Xp_2) = p_2' S p_2 = \lambda_2 p_2' p_2 = \lambda_2 \\ &\vdots \\ var(t_m) &= var(Xp_m) = p_m' S p_m = \lambda_m p_m' p_m = \lambda_m \end{aligned} \tag{2.70}$$

  donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$.

- La correlación entre las componentes principales es nula, ya que P es una matriz ortogonal:

$$cov(t_i, t_j) = cov(Xp_i, Xp_j) = p_i' S p_j = \lambda_j p_i' p_j = 0, \quad para \quad i \neq j \tag{2.71}$$

- Las covarianzas entre cada componente principal y las variables originales X_i son tal que:

$$cov(t_j, [x_1, x_2, \cdots, x_m]) = \lambda_j t_j', \quad j = 1, 2, \cdots, p. \tag{2.72}$$

  y a su vez,

$$cov(T, X) = \frac{1}{n} T' X' X = T' S = T'(T \Lambda T') = \Lambda T'. \tag{2.73}$$

  Nótese que las filas de dicha matriz están formadas por las covarianzas entre t_j y las variables originales $x_1, x_2, \cdots, x_m$.

Para definir en estas nuevas coordenadas la observación i-ésima, es decir, la fila $x_i' = (x_{i1}', x_{i2}', \cdots, x_{im}')$:

$$t_i' = x_i' P = (x_i' p_1, x_i' p_2, \cdots, x_i' p_m) \tag{2.74}$$

La variación total de X se define como la traza de la matriz estimada de varianza-covarianzas, es decir, $tr(S) = \sum_{i=1}^{m} \lambda_i$. Como consecuencia, la variación total de $T = XP$ es igual a la variación total de X para las m componentes principales seleccionadas, que vendrá dado por:

$$tr(var(T)) = tr(\frac{1}{n}) = tr(T'ST) = tr(T'T\Lambda T'T) = \sum_{i=1}^{m} \lambda_i \tag{2.75}$$

porque $S = T \Lambda T'$, con T siendo una matriz ortogonal.

De manera que, el porcentaje de variabilidad explicada puede darse directamente por:

$$VE(\,\%) = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} * 100 = \frac{\sum_{i=1}^{m} \lambda_i}{tr(S)} * 100, \quad k < m. \tag{2.76}$$

Cuando dicho cociente es cercano al 100 %, las variables $t_1, t_2, \cdots, t_k$ pueden reemplazar a las originales $x_1, x_2, \cdots, x_m$ sin gran pérdida de información con respecto a la variabilidad total.

Si x es un vector referente a una variable de calidad p-dimensional con autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, la combinación lineal elegida puede ser:

$$\begin{aligned} c_1 &= e_{11} x_1 + e_{12} x_2 + \cdots + e_{1p} x_p \\ c_2 &= e_{21} x_1 + e_{22} x_2 + \cdots + e_{2p} x_p \\ &\vdots \\ c_1 &= e_{p1} x_1 + e_{p2} x_2 + \cdots + e_{pp} x_p \end{aligned} \tag{2.77}$$

donde e_{ij}es el elemento j-ésimo del i-ésimo autovector y c_j los ejes del nuevo sistema de coordenadas, definido como una rotación del original en dirección de la máxima variabilidad. Y es esa precisamente ese el criterio de selección de las componentes principales, aquellas componentes que maximizan la variabilidad. Es por ello, por lo que resulta útil conocer el porcentaje de variabilidad explicada por cada una de las componentes

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \tag{2.78}$$

PCA basado en la matriz de correlaciones Es usual aplicar el análisis de componentes principales a las variables estandarizadas. Esto equivale a trabajar con la matriz de correlaciones, en lugar de la matriz de covarianzas.

Se denomina matriz de correlaciones a la matriz cuadrada y simétrica que está compuesta por unos en su diagonal, mientras que fuera de ella se encuentran los coeficientes de correlación entre las variables:

$$\begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix} \tag{2.79}$$

La matriz de correlaciones es una matriz semidefinida positiva. De manera que si D es la matriz diagonal de orden k construida de manera que en la diagonal principal se encuentran las desviaciones típicas de las variables, la matriz de correlaciones R está relacionada con la matriz de varianzas-covarianzas, S, mediante la siguiente relación:

$$R = D^{-1}SD^{-1} \tag{2.80}$$

A su vez, esto directamente implica que se cumpla que $S = DRD$.

Entonces, es posible aplicar la matriz de correlaciones en el análisis de componentes principales, ya que para el correcto uso del análisis de componentes principales basado en la matriz de varianzas-covarianzas se requiere que los datos sean homogéneos, lo que puede comprobarse aplicando un test de homogeneidad, como puede ser el test de homogeneidad no paramétrico de Levene.

Algunos procedimientos estadísticos comunes asumen que las varianzas de las poblaciones de las que se extraen diferentes muestras son iguales. Esto no siempre ocurre cuando las escalas de medida o unidades de la muestra no son iguales. Este test tiene como hipótesis nula que las varianzas poblacionales son iguales, o de manera equivalente:

$$H_0: \quad \textit{Datos homogéneos} \tag{2.81}$$

$$H_1: \quad \textit{Datos no homogéneos} \tag{2.82}$$

Esta hipótesis nula está directamente relacionada con la homocedasticidad de la muestra. El estadístico de contraste de Levene, W es:

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i(Z_{i.} - Z_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i} N_i(Z_{ij} - Z_{i.})^2} \tag{2.83}$$

Cada Z_{ij} se calcula en base a la mediana de cada uno de las observaciones. Y al igual que cualquier otro contraste de hipótesis, si el p-valor resultante es inferior a un cierto nivel de significación prefijado α, se dice que existen evidencias significativas para rechazar la hipótesis nula de homogeneidad.

Selección del número de componentes El PCA trata de reducir la dimensión de los datos sin gran pérdida de información siempre y cuando los datos se encuentren correlacionados. En otras palabras, es un método empleado para quedarse con pocas variables latentes obtenidas gracias a la estructura de correlación de los datos y conservando gran parte de la variabilidad.

La matriz X, es una matriz de dimensión $n \times m$, por X. Obviamente, cuando interesa la reducción de la dimensión, se busca reducir la matriz X en una matriz T de scores, con dimensión $n \times k$, con $k << m$, que recoge la mayor información de X. Lo que implica que lógicamente las $m-k$ componentes excluidas del modelo no presentan información relevante y constituyen la matriz de los residuos E. Para lograr una selección de componentes principales equilibrada, existen diversos métodos, pero todos buscan un porcentaje de variabilidad explicado elevado sin excederse en el número de componentes principales.

Baillo y Grane (2007) presentan 3 criterios de selección de componentes principales:

- Porcentaje de Variabilidad Explicada:

  Es el método más sencillo, y consiste en establecer un valor para el porcentaje de variabilidad explicada, $VE(\,\%) = 90\,\%$, y considerar aquellas primeras componentes principales que consigan superar dicho porcentaje prefijado.

- Criterio de Kaiser:

  Este criterio incluye en la selección aquellas componentes principales cuyos autovalores sean superiores a $\bar{\lambda} = \frac{tr(S)}{p}$, o por el contrario, superiores a 1 si se han calculado las componentes a partir de la matriz de correlación R.

- Modificación de Jollife:

  Parte del criterio definido previamente, el criterio de Kaiser, que tiene la pecularidad de que cuando $m \leq 20$, tiende a incluir pocas componentes. Por lo que la modificación de Jollife trata de excluir aquellas componentes cuyos autovalores sean menores que $0{,}7\bar{\lambda} = 0{,}7\frac{tr(S)}{p}$, o menores que 0.7 en caso de que se hayan calculado las componentes principales a partir de la matriz de correlación R.

Existen muchos otros autores que a lo largo del tiempo han propuesto criterios de selección del número de componentes principales como **Otoole y Abdi (1993)**, **Jolliffe et al. (2002)**, **Jackson et al. (1991)** y **Peres-Neto et al. (2005)**.

Inferencia o Validación del modelo PCA Los conceptos y los criterios introducidos respecto al modelo PCA, tratan de obtener conclusiones en base a las observaciones de la muestra, es decir, no se aplican a la población en estudio, lo que se conoce como modelo de efectos fijos.

Considerando la matriz X_k como la matriz representativa de las k componentes principales seleccionadas, se puede evaluar la similitud entre dicha matriz y X, la matriz original de datos. El procedimiento más sencillo y más empleado para realizar esta comparativa es el método de la suma de los cuadrados residuales, RSS:

$$RSS_k = \|X - X_k\| = tr(E'E) = I - \sum_{i=1}^{k} \lambda_i \tag{2.84}$$

Es base a dicha expresión, cuanto menor sea el valor de RSS, mejor ajuste presentará el modelo PCA. En el momento en el que se pretende estimar el valor de una nueva observación de la población, primeramente se debe garantizar que el modelo PCA no varía muco con la introducción de esas nuevas observaciones. Es importante que para el nuevo cálculo de la capacidad generalizada del modelo, no se deben emplear los métodos estándar, sino que para un cálculo eficiente, se deben emplear procedimientos basados en técnicas de remuestreo, como bootstrap y validación cruzada.

Para estas técnicas se separan los datos en dos grupos, denominados muestra de aprendizaje y muestra de pruebas (learning set y test set respectivamente). Primeramente se establece un conjunto de pruebas, normalmente se emplea la metodología "one-leave-out" **Quenouille (1956)**, dejando cada observación fuera del conjunto y las demás constituyen el conjunto de aprendizaje **Efron (1982)**.

El conjunto de aprendizaje es empleado después para estimar la observación dejada fuera que constituye el conjunto de pruebas. Empleando este procedimiento, se tiene por objetivo garantizar que el modelo es robusto cuando se introducen nuevas observaciones en el modelo. Los valores predichos conformarán la matriz $\hat{X}_k$.

Entones, podrá medirse la calidad general del modelo PCA con efectos aleatorios empleando las k componentes principales a través de la diferencia entre X y $\hat{X}_k$. En el PCA de efectos fijos se empleaba RSS para la evaluación de la calidad del modelo, y en efectos aleatorios se emplearán PRSS (Predicted Residual Sum of Squares).

$$PRSS_k = \left\| X - \hat{X}_k \right\| \tag{2.85}$$

Y como en el caso anterior de efectos fijos, cuanto menor sea el valor de PRSS, mejor ajuste tendrá el modelo PCA analizado. Se debe tener en cuenta que es posible observarse un decrecimiento de la calidad de predicción de un modelo ajustado al incrementar el número de componentes principales seleccionadas, ya que este hecho llevaría a una sobreestimación, es decir, que la información del conjunto de entrenamiento no es útil para ajustar el conjunto test **Abdi et al. (2010)**. En muchos casos, para determinar el número óptimo de componentes que mejor representan al modelo se emplea el siguiente estadístico:

$$Q_k^2 = 1 - \frac{PRSS_k}{RSS_{k-1}} \tag{2.86}$$

Este estadístico de bondad de ajuste evalúa en simultáneo el número óptimo de componentes y la robustez del modelo. En líneas generales, se suelen mantener en el modelo aquellas componentes que presenten un valor de $Q_k^2 \geq 0{,}0975$ (1-0.952).

Aplicación de PCA a Gráficos de Control A continuación, se presenta la aplicación del PCA en el análisis estadístico de la calidad mediante gráficos de control multivariantes, ya que los gráficos univariantes son incapaces de modelar la estructura de correlación existente en este tipo de datos.

Un método muy empleado en el control de procesos es adaptar el SPC en el entorno de variables latentes. Partiendo del gráfico T^2 de Hotelling para el control de k variables latentes de un proceso:

$$T_k^2 = \sum_{j=1}^{k} \frac{t_j^2}{s_{t_j}^2} \tag{2.87}$$

El estadístico T^2 de Hotelling para este caso, es la suma de orden k de los cocientes correspondientes al cuadrado de cada una de las variables y sus respectivas varianzas. Dichas varianzas son la diagonal de la matriz de varianza-covarianza Σ y son iguales a los autovalores λ_j de la matriz. Por lo que es posible reescribir la expresión anterior como:

$$T_k^2 = \sum_{j=1}^{k} \frac{t_j^2}{\lambda_j} \tag{2.88}$$

El objetivo de que en el denominador de la expresión aparezca la varianza correspondiente a cada una de las componentes principales no es más que cada componente principal tenga el mismo peso en el cálculo del estadístico T^2. Además, no siempre la matriz de datos originales X está estructurada correctamente de acuerdo a que las k primeras componentes principales expliquen su mayor parte de variabilidad, por lo que al dividir los t_j^2 por sus pequeñas varianzas, evitan que pequeñas desviaciones de estas t_j^2 que no tienen apenas efecto sobre X, den lugar a una señal fuera de control en T^2.

De manera que si ocurre un evento especial, totalmente nuevo que no estaba presente en los datos de referencia empleados para desarrollar el PCA, entonces nuevas componentes aparecerán y la

nueva observación quedaría fuera de alcance. Debido a esto, **Kresta y MacGregor (1995)** propone el gráfico del error cuadrático de predicción (SPE), muy útil para este tipo de situaciones. Para la construcción de este gráfico, la expresión empleada es la siguiente:

$$SPE_x = \sum_{j=1}^{m} (x_{nuevo,i} - \hat{x}_{nuevo,i})^2 \tag{2.89}$$

De manera que mientras el proceso está bajo control, el valor correspondiente al SPE_x será pequeño, y en ese caso, SPE será representativo del ruido, lo que se conoce también como la variabilidad no explicada por el modelo. Sin embargo, un valor alto de SPE_x, indica que el modelo de proyección no es válido para dicha observación, y en ese caso, la monitorización debe ser llevada a cabo empleado gráficos T^2 y SPE.

Jackson et al. (1991) propuso tres aplicaciones de las componentes principales en los gráficos de control: El gráfico de control de Hotelling aplicado a las puntuaciones de las componentes principales, el gráfico de control aplicado a los residuos y el gráfico de control univariante para cada puntuación. No obstante, el primero de los métodos es el más empleado en control estadístico de la calidad. El motivo del uso de este método es sencillo y muy intuitivo ya que es posible reducir un análisis donde existen 6 o 7 variables o características de la calidad a 2 o 3 variables, lo que puede llevar al caso sencillo donde es posible controlar el proceso a través de un elipsoide en dos o tres dimensiones.

Es decir, este método trata de realizar el análisis de componentes principales sobre las variables o características de la calidad originales. Posteriormente, se procede a la reducción de la dimensionalidad, tomando un número inferior de variables en función de la variabilidad que puede ser explicada. Finalmente, se aplicará el método gráfico de χ^2, suponiendo que los parámetros son conocidos y se cuenta con una base de datos suficientemente grande, o se empleará un gráfico de control T^2, suponiendo que los parámetros son desconocidos.

PCA para procesos dinámicos Hasta el momento, el modelo de componentes principales se aplica a procesos en los que las variables de estudio están correlacionadas y las observaciones son independientes. Sin embargo, es muy natural que los procesos industriales muestren una dependencia temporal, además de dicha estructura de correlación entre las variables. Bajo estas situaciones, los algoritmos empleados en el cálculo de la matriz T de los scores o puntuaciones y la correspondiente P de los loadings o pesos no son capaces de captar toda la información relevante necesaria en la matriz X, **Vanhatalo y Kulachi (2015)**. El impacto negativo de la autocorrelación viene reflejado en la varianza explicada por las componentes principales determinadas.

Por lo que cuando los datos muestran estructura de correlación y autocorrelación, se dice que los datos contienen informaciones dinámicas, y en estos casos, el PCA revela una aproximación estática lineal, en lugar de revelar las relaciones exactas entre las variables. Si se pretende modelar un proceso dinámico, se debe emplear un modelo Dynamic PCA (DPCA), que fue propuesto por **Ku et al. (1995)**. Este método DPCA consiste en incluir en la matriz de los datos X las variables decaladas en el tiempo (lag), ajustando un PCA posteriormente sobre ella.

- Definición del DPCA

  Se partirá de la matriz X de dimensión $t \times m$ que recoge dichas variables correlacionadas y autocorrelacionadas del proceso dinámico del proceso en estudio. Se denominará $\acute{X} = [X_0, X_1, X_2, \cdots, X_s]$ a la matriz extendida, con dimensión $t \times [m(s+1)]$, que recoge las variables de X junto con las variables decaladas de orden $0, 1, 2, \cdots, s$. Se debe considerar como $X_0 = X$, X_1 como la matriz que recoge las variables de X decaladas con lag=1, y sucesivamente, por lo que en general, X_s recoge las variables de X decaladas lag=s.

  Las componentes principales se calcularán de acuerdo a:

$$PC_{j,t} = P_j' \acute{X} \tag{2.90}$$

donde $PC_{j,t}$ es la componente principal j-ésima y P_j' su correspondiente vector de loadings o pesos.

Considerando un caso sencillo, en el que X solamente tiene solo dos variables correlacionadas y autocorrelacionadas. Por ejemplo con s=1 (lag=1), la matriz $\acute{X}$ quedaría $\acute{X} = [X_0, X_1]$ y por lo tanto las componentes principales serían calculadas como se muestra a continuación:

$$\begin{aligned} PC_{1,t} &= p_{11}x_{1,t} + p_{12}x_{2,t} + p_{13}x_{1,t-1} + p_{14}x_{2,t-1} \\ PC_{2,t} &= p_{21}x_{1,t} + p_{22}x_{2,t} + p_{23}x_{1,t-1} + p_{24}x_{2,t-1} \\ PC_{3,t} &= p_{31}x_{1,t} + p_{32}x_{2,t} + p_{33}x_{1,t-1} + p_{34}x_{2,t-1} \\ PC_{4,t} &= p_{41}x_{1,t} + p_{42}x_{2,t} + p_{43}x_{1,t-1} + p_{44}x_{2,t-1} \end{aligned} \tag{2.91}$$

siendo p_{ij} los elementos de la matriz P asociado al loading i-ésimo y la variable j-ésima. Como se puede observar una vez visto el procedimiento, a diferencia del PCA, las componentes principales en el método DPCA son combinaciones lineales de las variables originales y sus respectivos retardos. Las variables $x_{1,t-1}$ y $x_{2,t-1}$ son las variables que recogen dicha información adicional que el PCA no es capaz de captar.

- Lags y el número de componentes DPCA

  La decisión a la hora de determinar el número de lags necesarios para formar la matriz $\acute{X}$ fue discutida por muchos autores. A continuación se muestran tres de los criterios más empleados en la práctica para determinar dicho número de lags.

  Por un lado, se encuentra el método KSG-95 fue propuesto por **Ku et al. (1995)** y consiste en identificar el número de lags analizando la relación lineal entre las componentes principales de la matriz extendida mediante análisis de los gráficos de autocorrelaciones y correlaciones cruzadas de los scores o puntuaciones.

  Posteriormente, **Rato y Reis (2013)** propusieron el método RR-13, basado en identificar los lags de acuerdo a la minimización en simultáneo del principal valor singular y el principal ratio del valor singular de la matriz extendida $\acute{X}$.

  Cinco años más tarde, **Vanhatalo y Kulachi (2017)** propusieron un método que consiste en identificar el número máximo de lags analizando los autovalores de las matrices de correlaciones simples y parciales calculadas en base a la matriz $\acute{X}$. Si el autovalor entre X y X_s es próximo a cero, indicará que la variable correspondiente al lag=s no presenta información relevante al modelo, por lo que sencillamente dicho lag puede ser descartado.

A su vez, la selección de las componentes principales del modelo DPCA puede llevarse a cabo empleando cualquiera de los métodos de selección planteados para los modelos PCA.

Partial Least-Square Regression (PLS)

Partial Least-Square (PLS) es un método que se aplica a procesos multi-input, multi-output (MIMO) y se basa en el ajuste de un modelo de regresión entre las variables explicativas X y respuestas Y (con variables latentes).

El método fue introducido por **Wold et al. (1978)** a finales de los años sesenta, motivado por ciertos estudios econométricos. A día de hoy, se considera la generalización del modelo de regresión lineal múltiple (MRL), solucionando los problemas derivados de la multicolinealidad.

La calidad predictiva de los modelos de regresión suele depender del número de variables explicativas en el modelo. Y es precisamente en este punto donde el PLS muestra una ventaja frente a otros modelos de regresión.

Metodología PLS La metodología PLS trata de proyectar los datos correlacionados en estructuras de variables latentes, pero con una diferencia frente al PCA, ya que en PLS las variables latentes son construidas en dirección de máxima covarianza entre la matriz explicativa X y respuestas Y.

Teniendo en cuenta que la matriz X es de dimensión $n \times m$ y la matriz de variables respuestas Y es de dimensión $n \times p$, el modelo PLS aplica una relación externa, que modela por un lado X e Y, y por otro lado, de manera interna.

Definiendo t_j como la variable latente de la matriz X, w_j su correspondiente peso, u_j como la variable latente de la matriz Y y su respectivo pero q_j, $j = 1, 2, \cdots, k$:

$$t_j = Xw_j \tag{2.92}$$

$$u_j = Yq_j \tag{2.93}$$

Como esta metodología busca la dirección de mayor covarianza entre las componentes de X con las de Y, los vectores w y q tienen norma unitaria y son calculados de modo que se maximiza la covarianza entre t_j y u_j. Dicho de otra forma, maximizar $cov(t, u)$ para $\|w\| = \|q\| = 1$.

La relación interna entre los scores o puntuaciones de t_j Y u_j viene dada por:

$$u_j = b_k t_k \tag{2.94}$$

donde $b_k = \frac{u_k' t_k}{t_k' t_k}$ y se encuentra bajo las mismas reglas que los coeficientes de regresión de los modelos MRL o PCR.

En cuanto a las relaciones externas en las matrices X e Y:

$$X = T_k W_k' + E \tag{2.95}$$

$$Y = U_k Q_k' + F \tag{2.96}$$

donde T, W, E son los scores loadings y residuos respectivamente de los bloques de X, y U, Q, F son los scores, loadings y residuos de Y.

La forma de proceder a la construcción de los scores del modelo PLS, **Altaf et al. (2013)** propuesto el algoritmo NIPALS, el cual se presenta a continuación:

1. Considere las matrices $E_{j+1} = E_j$, $F_{j+1} = F_j$, con $j = 1, 2, \cdots, k$, $E_1 = X$, $F_1 = Y$, tomar u_j igual a una columna de Y.
2. $w_j' = \frac{u_j' E_j}{u_j' u_j}$
3. Normalizar w_j de acuerdo a $w_j = \frac{w_j}{\|w_j\|}$
4. $t_j = \frac{E_i e_j}{w_j' w_j}$
5. Empleando la matriz F_j para calcular $q_j' = \frac{t_j' F_j}{t_j' t_j}$
6. Normalizar q_j de acuerdo a $q_j = \frac{q_j}{\|q_j\|}$
7. Calcular u_{nuevo} como $u_{nuevo} = \frac{F_j q_j}{q_j' q_j}$
8. Verificar que hay convergencia (se deberá asumir que existe convergencia siempre que u_{nuevo} no cambie significativamente respecto al valor de u_j en el paso anterior). Si se obtiene convergencia, seguir al paso 9. En caso contrario, se deberá volver al paso 2 y sustituir u_j por u_{nuevo}.
9. El loading p_j de las variables explicativas viene dado por $p_j' = \frac{t_j' E_{j-1}}{t_j' t_j}$

10. Empleando las matrices U y T de los scores determinados, se calcula el coeficiente b_j de regresión interna para las variables latentes.
    $b_j = \frac{t_j' u_j}{t_j' t_j}$
11. Finalmente, las matrices de los residuos de las matrices X e Y serán:

$$\begin{aligned} E_{j+1} &= E_j - t_j' p_j \\ F_{j+1} &= F_j - t_j b_j p_j' \end{aligned} \tag{2.97}$$

Selección del número de componentes Para poder garantizar la captura de la variabilidad esencial de los datos originales es necesaria la selección óptima del número de componentes principales del modelo PLS. Ya que si se selecciona un número excesivo de variables latentes, puede conllevar a que el ajuste sea excesivo debido a la inclusión de variables latentes que explican el ruido del proceso. Mientras que por otro lado, si el número de variables latentes es muy reducido, no se lograra modelar el comportamiento principal del proceso completamente, lo que desembocará en predicciones pobres.

Wold et al. (1987) presenta un método de selección de variables latentes basado en la validación cruzada, los pasos a seguir son los siguientes:

1. La base de datos se divide en s conjuntos, donde $s-1$ conjuntos son empleados para el entrenamiento y el resto para testar el modelo, lo que se conoce como grupo de pruebas.
2. Una variable latente se determina a partir del conjunto de entrenamiento, y posteriormente se aplica al grupo de pruebas. Se calcula el PRESS (Predicted Error Sum of Squares), repitiendo el proceso para cada no de los s subgrupos excluidos. Finalmente, se calcula el PRESS total.
3. Se repite el procedimiento para las variables latentes restantes, y se calculan los correspondientes PREESS totales.

El estadístico R de Wold es calculado como:

$$R = \frac{PRESS(k+1)}{PRESS(k)} \geq \alpha \tag{2.98}$$

donde k es el número de variables latentes y $\alpha = 0{,}95$ es el umbral.

Este estadístico evalúa la contribución de cada variable latente al modelo PLS. Recordando que $PRESS_k = \left\| X - \hat{X}_k \right\|$ y $PRESS_{k+1} = \left\| X - \hat{X}_{k+1} \right\|$. Si k es el número óptimo de variables, $PRESS_{k+1}$ no será significativamente inferior que $PRESS_k$ ($R \geq 0{,}95$)de manera que la nueva variable latente no aportaría información significativa al modelo.

PCA para procesos dinámicos Muchos han sido los estudios realizados tratando de adaptar el modelo PLS a sistemas dinámicos. El modelo PLS originalmente fue desarrollado con el objetivo de modelar procesos con datos de entrada X y salida Y altamente correlacionados. El modelo, tiene como propósito también la reducción de la dimensión de X e Y, formando variables latentes, donde se ajusta un modelo de regresión. Cuando existen relaciones dinámicas entre X e Y además de la alta correlación entre variables, se tiene que el PLS resulta inadecuado ya que deja una gran cantidad de covarianza sin modelar.

Existen diversos métodos propuestos por distintos autores, los más relevantes se muestran a continuación:

- El método propuesto por **Wold et al. (1987)** consiste en incluir los decalajes de las variables en la matriz de variables explicativas y ajustar posteriormente el modelo PLS sobre la nueva matriz construida. Ante esta propuesta, **Qin y McAvoy (1996)** propusieron un enfoque similar, que consiste en extender la matriz de entrada X incluyendo un cierto número de variables explicativas y respuestas decaladas.

- El método de **Kaspar y Ray (1993)** está desarrollado en base al diseño de filtros mediante un conocimiento dinámico previo de modo que se elimina la estructura dinámica en la matriz X de entrada. A continuación, se procedería a ajustar un modelo dinámico con X e Y, tratando precisamente de modelizar la relación dinámica de los scores. Por su lado, **Lakshminarayanan et al. (1997)** llevaron a cabo una modificación partiendo de este procedimiento en el que ajustan la relación interna mediante un modelo dinámica entre los scores de X e Y.

- El método desarrollado por **Dong y Qin (2015)** es el método más reciente. Según ellos, las metodologías presentadas hasta el momento presentan una inconsistencia entre el modelo dinámico ajustado a los scores, que modela la relación interna entre las variables, y el modelo externo estático, ya que los scores extraídos de manera estática son forzados a tener una relación dinámica en el modelo interno.

  Estos dos autores proponen como alternativa el modelo dynamic-inner PLS (DiPLS). Esta metodología es la que mejor resultados obtiene y el procedimiento se desarrolla a continuación:

  Sea x_k e y_k las variables explicativa y respuesta en el instante k ($k = 1, 2, \cdots, N+1$), con u_k y t_k sus respectivos scores. El modelo verifica que:

$$u_k = \beta_0 t_k + \beta_1 t_{k-1} + \cdots + \beta_s t_{k-s} + r_k \tag{2.99}$$

  y $u_k = y_k' q$ y $t_k = x_k' w$, donde w es el peso de x_k, q es el peso de y_k, y r_k es el ruido.

  Para cada factor, la relación interna será definida por:

$$\hat{u}_k = x_k' w\beta_0 + x_{k-1} w\beta_1 + \cdots + x_{k-s}' w\beta_s = \left[x_k' x_{k-1}' ... x_{k-s}'\right](\beta * w) \tag{2.100}$$

  con $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_s)'$.

  El modelo externo consistente con el modelo interno anterior presentado debe maximizar a covarianza entre u_k y $\hat{u_k}$ como se muestra a continuación:

$$cov(u_k, \hat{u_k}) = \frac{1}{N} \sum_{k=s}^{N+s} q' y_k \left[x_k' x_{k-1}' ... x_{k-s}'\right](\beta * w) \tag{2.101}$$

  Es importante definir X e Y como:

$$X = [x_0 \quad x_1 \quad \cdots \quad x_{s+N}]' \tag{2.102}$$

$$Y = [y_0 \quad y_1 \quad \cdots \quad y_{s+N}]' \tag{2.103}$$

  y las matrices Y_s y Z_s:

$$Y_s = [y_s \quad y_{s+1} \quad \cdots \quad y_{s+N}]' \tag{2.104}$$

$$Z_s = [X_s \quad X_{s-1} \quad \cdots \quad X_0]' \tag{2.105}$$

  y el modelo es ajustado bajo el criterio de maximización:

$$max[q' Y_s' Z_s (\beta * w)] \tag{2.106}$$

  con $\|w\| = 1$, $\|q\| = 1$ y $\|\beta\| = 1$, siendo s el orden dinámico del modelo.

  Para proceder a la resolución de este problema de optimización, se requiere la aplicación de multiplicadores de Lagrange, y los vectores q, w y β que satisfacen dicha optimización serán determinadas mediante un proceso iterativo que inicia por seleccionar los vectores unitarios q, w y β, y a continuación son calculados como:

$$q = Y_s' Z_s (\beta * w); \quad q := \frac{q}{\|q\|} \tag{2.107}$$

$$w = (\beta * I)' Z_s' Y_s q; \quad w := \frac{w}{\|w\|} \tag{2.108}$$

$$\beta = (I * w)' Z_s' Y_s q; \quad \beta := \frac{\beta}{\|\beta\|} \tag{2.109}$$

y los scores t y u de las variables explicativas y respuestas respectivamente son representados por:

$$t = [t_0 \quad t_1 \quad ... \quad t_{s+N}] = X_w \tag{2.110}$$

$$u = [u_0 \quad u_1 \quad ... \quad u_{s+N}] = Y_q \tag{2.111}$$

y la relación dinámica del modelo es definida por:

$$u_s = \alpha_0 t_s + \alpha_1 t_{s-1} + \cdots + \alpha_s t_0 + r_s \tag{2.112}$$

donde r_s es el residuo del modelo. Si se pretende aplicar un modelo DiPLS para la estructura autorregresiva de las variables respuestas, el modelo se representa como se muestra a continuación:

$$u_s = \phi_0 u_{s-1} + \cdots + \phi_1 u_{s-f} + \alpha_0 t_s + \alpha_1 t_{s-1} + \cdots + \alpha_s t_0 + r_s, \tag{2.113}$$

como resultado del siguiente problema de maximización:

$$max \quad q'(\gamma_0 Y_s' + \gamma_1 u_{s-1} + \cdots + \gamma_f Y_{s-f}) * (\beta_0 X_s + \beta_1 X_{s-1} + \cdots + \beta_S X_0) w \tag{2.114}$$

con las restricciones correspondientes: $\|w\| = 1$, $\|q\| = 1$, $\|\beta\| = 1$ y $\|\gamma\| = 1$.

2.2.4. Gráficos de control para datos multivariantes no paramétricos

A continuación, se va a desarrollar la implementación de los gráficos de control r, Q, S y S^* aplicados al contexto multivariante. Estos métodos no paramétricos se desarrollan a partir del concepto de profundidad de los datos, empleando métodos de clasificación por rangos para determinar si una observación es diferente de aquellas que conforman el conjunto original. Estos gráficos de control son propuestos por **Regina y Liu (1995)**, y son similares a los ya planteados en el caso univariante.

En primer lugar, cabe introducir el concepto de profundidad, ya que los gráficos propuestos por Liu se basan precisamente en el concepto de profundidad de los datos, con el único requerimiento de contar con una distribución de referencia, que describe una distribución k-dimensional, con $k \geq 1$. A partir de la distribución de probabilidad en k-dimensión, una función de profundidad asigna a cada punto de $\mathbb{R}^k$ su grado de centralidad respecto de la distribución de referencia.

La profundidad de los datos está basada en el hecho de que cualquier densidad de probabilidades diferencia datos centrales de los periféricos, de forma que una función de profundidad asigna a cada y en $\mathbb{R}^k$ un valor no negativo, que puede interpretarse como su localización en el conjunto de datos. Por lo tanto, las profundidades grandes corresponden al centro de la distribución. Este método de clasificación por rangos es empleado para determinar si una observación es diferente de aquellas que conforman el conjunto original. Las funciones de profundidad deben satisfacer ciertas propiedades como que deben ser invariantes afín, mostrar monotonicidad, maximalidad al centro y desvanecimiento al infinito.

Existen diferentes funciones de profundidad, como la profundidad de Mahalanobis, la profundidad simplicial y la profundidad de Tukey. En este caso, Regina Liu emplea la profundidad de Mahalanobis debido a su simplicidad, tanto en el cálculo como en la interpretación de la misma, y obviamente esta medida de profundidad parte de la distancia de Mahalanobis.

$$MD = \frac{1}{\left[1 + (y - \mu)' \sum^{-1} (y - \mu)\right]} \tag{2.115}$$

Donde μ es el vector de medias y $\sum^{-1}$ responde a la inversa de la matriz de varianzas-covarianzas de la distribución de referencia. En un contexto en el que no se conocen los parámetros de la distribución de referencia, el cálculo de la profundidad de Mahalanobis es la siguiente:

$$MD = \frac{1}{\left[1 + (y - \bar{Y})' S^{-1} (y - \bar{Y})\right]} \tag{2.116}$$

Donde $\bar{Y}$ es el vector de medias muestrales de los datos $Y_1, Y_2, \cdots, Y_m$ y S es la matriz de varianzas-covarianzas estimadas de la muestra de referencia.

La clasificación por rangos en el caso multivariante indica el grado de centralidad de una observación con respecto de las demás. En primer lugar se obtienen las profundidades de cada una de las observaciones, y posteriormente se procede a ponderar entre la cantidad total de observaciones el número de veces que dicho valor es menor o igual a todos los demás valores de profundidad, es decir, análogamente al caso univariante, se trata de un conteo. El estadístico de clasificación por rangos en este caso, $r(.)$ por tanto tiene la siguiente expresión:

$$r(x_i) = \frac{\#\left\{y_j | D(y_j \leq D(x_i, j = 1, 2, \cdots, n))\right\}}{n} \tag{2.117}$$

Gráfico de control multivariante r

Para proceder a la construcción de la gráfica multivariante r se deben seguir los siguientes pasos:

1. Calcular el vector de medias, la matriz de varianzas-covarianzas y la profundidad de cada uno de los datos, $D(y_i)$ con $i = 1, 2, \cdots, m$.

2. Obtener los estadísticos de orden de las $D(y_i)$ con $i = 1, 2, \cdots, m$, que se denotarán por $y_{[1]}, y_{[2]}, \cdots, y_{[m]}$.

3. Sean $x_1, x_2, \cdots, x_m$ las nuevas observaciones con distribución continua, se deben de calcular sus respectivas profundidades.

4. Calcular el rango $r(.)$ para cada una de las observaciones.

5. Graficar los estadísticos de clasificación por rangos de cada x_i respecto al tiempo, con un límite de control UL=0.5 y un límite inferior de control tal que LC=α. Donde α es la proporción de alarma.

Liu demostró que éste estadístico se distribuye como una uniforme $U[0, 1]$.

Para proceder con la construcción del gráfico de control r no paramétrico, se deben calcular los r_i, y después los parámetros gráficos, que vienen dados a continuación:

$$\begin{aligned} &\textit{Línea central} = 0.5 \\ &LC = \alpha \end{aligned} \tag{2.118}$$

Como hasta ahora, el proceso se dice que está fuera de control en el caso en el que algún $r_n(x)$ se encuentre por debajo del limite LC.

Generalmente, cuando en un proceso industrial se quieren controlar diversas características de calidad, se opta por la construcción de gráficas Shewhart, como el gráfico de control multivariante $T^{"}$ de Hotelling, que es un método paramétrico que requiere de dos suposiciones esenciales, la normalidad de los datos y de la no correlación entre mediciones sucesivas. Pero este supuesto no siempre se cumple.

Gráfico de control multivariante Q

Este gráfico es el gráfico no paramétrico análogo al gráfico de control paramétrico $\bar{X}$. Para su construcción se debe comenzar por hallar la función de profundidad a través de la definición de Mahalanobis, y posteriormente calcular el estadístico de clasificación por rangos $r(.)$.

La distribución empírica de la variable X se define como se muestra a continuación:

$$\begin{aligned} Q &= \mathbb{P}\{D(Y) \leq D(X)\} \\ Q(r) &= \left(\tfrac{1}{n}\right) \sum_{i=1}^{n} r(X_i) \end{aligned} \tag{2.119}$$

Para el cálculo de los parámetros gráficos se deben distinguir tres distintas situaciones:

- En primer lugar, si el tamaño muestral es mayor o igual a 5 y los parámetros asociados a la distribución de los datos son conocidos, se calcularán los parámetros gráficos como:

$$\begin{aligned} &\textit{Línea central} = 0.5 \\ &LC = 0.5 - z_\alpha (12n)^{1/2} \end{aligned} \tag{2.120}$$

- Si por el contrario, el tamaño muestral es mayor o igual a 5 pero en este caso, los parámetros asociados a la distribución de los datos son desconocidos, los parámetros gráficos serán:

$$\begin{aligned} &\textit{Línea central} = 0.5 \\ &LC = 0.5 - z_\alpha \sqrt{\frac{\frac{1}{n} + \frac{1}{k}}{2}} \end{aligned} \tag{2.121}$$

- Finalmente, si se da el caso en el que el tamaño muestral es inferior a 5 y además el valor de α es pequeño, los parámetros gráficos serán:

$$\begin{aligned} &\textit{Línea central} = 0.5 \\ &LC = \frac{(n!\alpha)^{1/n}}{n} \end{aligned} \tag{2.122}$$

Como para el resto de gráficos de control, si algún $r_n(x)$ está por debajo del valor del LC, se declara que el proceso está bajo control.

Gráfico de control multivariante S y S^*

En cuanto al gráfico de control S y S^*, es un gráfico de control multivariante no paramétrico basado en el gráfico de control paramétrico CUSUM de sumas acumuladas. El primer paso consiste en, una vez más, hallar el estadístico $r_n(x)$ para cada una de las mediciones de la muestra y se calcula posteriormente el valor del S_n como se muestra a continuación:

$$S_n = \sum_{i=1}^{n} \left[r(X_i) - \frac{1}{2} \right] \tag{2.123}$$

junto con los valores de la línea central y el LC:

$$\begin{aligned} &\textit{Línea central} = 0 \\ &LC = -\left(z_\alpha \left(\tfrac{n}{12}\right)^{1/2} \right) \end{aligned} \tag{2.124}$$

En situaciones donde el tamaño muestral es muy grande, es preferible emplear el gráfico de control S^*, y de manera similar al caso anterior del gráfico S, se deben calcular los S_n^*:

$$S_n^* = \frac{S_n}{\sqrt{\frac{n}{12}}} \tag{2.125}$$

En este tipo de gráficos, como el tamaño muestral es muy grande, los cálculos para los parámetros gráficos se ven simplificados:

$$\begin{aligned} Línea\ central &= 0 \\ LC &= -z_\alpha \end{aligned} \tag{2.126}$$

Una vez más, cuando alguno de los puntos del gráfico de control se encuentre bajo la línea del límite LC, se dirá que el proceso está fuera de control estadístico y siguiendo con la filosofía del ciclo DMAIC, se deberá proceder a corregir y eliminar las causas que hayan podido provocar esta situaciones.

Comparativa de la metodología con datos simulados A continuación se va a realizar una comparación del comportamiento de los gráficos empleando datos normales y datos que no provienen de una distribución normal. Para esta comparación empleando el gráfico de control r, se crea una pequeña simulación de datos, generando 100 observaciones bivariantes con $\lambda = 0.5$ y otras 40 con $\lambda = 4$. Para esta pequeña prueba, se toman las últimas 40 observaciones de la primera muestra generada y las otras 40 de la segunda muestra, formando así una lista de 80 observaciones para dos variables del proceso.

En la Figura 2.6 se observa la gráfica de control r a partir de los datos simulados. Es notable la diferencia cuando se trata de la primera muestra de observaciones generadas o de la segunda, ver Figura 2.7, debido a la diferencia de medias existente entre ellas. Cuando las variables son generadas de una distribución no-normal, como en este caso, que sigue una distribución exponencial, el desempeño del método no paramétrico es mejor que el de la T^2 de Hotelling. Además, si se observan los ejemplos generados de distribuciones normales para ambas gráficas, el comportamiento de la gráfica T^2 de Hotelling es más eficiente a menos que el cambio en la media de las variables sea grande, pues en tal caso ambas gráficas parecen tener un desempeño bastante similar.

2.2.5. Capacidad

La capacidad de un proceso puede describirse como un campo en el control estadístico de la calidad enfocado en la determinación de la viabilidad del proceso a la hora de cumplir con las especificaciones. Normalmente, la capacidad de un proceso se expresa en forma de ratio o de índice entre tolerancias y el rendimiento del propio proceso. Se dice que un proceso es capaz cuando casi todas las muestras se encuentran dentro de los valores límite de las especificaciones. Los límites del estudio de capacidad vendrán definidos de acuerdo a las 3σ desde la media, debido al supuesto de normalidad.

Existen ciertos índices de capacidad muy empleados en la industria, como los que se muestran a continuación, donde USL y LSL indican el límite superior e inferior de las especificaciones respectivamente.

$$C_p = \frac{USL - LSL}{6\sigma} \tag{2.127}$$

$$C_{pk} = min\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right) \tag{2.128}$$

$$C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu + T)^2}} \tag{2.129}$$

$$T = \frac{USL - LSL}{2} \tag{2.130}$$

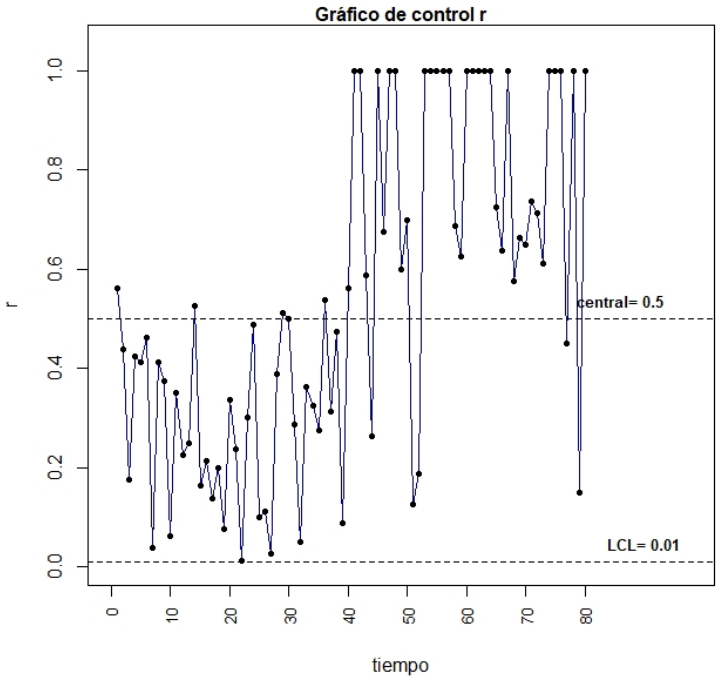


Figura 2.6: Gráfico de control r.

El último índice, es el objetivo y viene dado por el punto medio de las especificaciones. Para calcular el resto de los índices, es necesario conocer los parámetros desviación típica y/o media σ y μ, lo que en la práctica implica emplear la estimación correspondiente a cada una de ellas. Cuando los índices anteriores C_p, C_{pk}, C_{pm} son superiores a 1, es posible concluir de acuerdo a sus expresiones que el proceso es capaz.

Índices de capacidad en procesos multivariantes (MPCI)

Un índice de capacidad puede definirse como un ratio de las especificaciones de un proceso, ofreciendo información sobre la satisfacción de los requerimientos.

Los índices de capacidad, al fin y al cabo, tratan de realizar una comparación de la salida de un proceso y los límites de control establecidos. Las variables o características de calidad a estudiar pueden seguir una distribución normal o no, incluso pueden estar correladas entre ellas o por el contrario, independientes. En la práctica si los datos no son normales, se suelen aplicar transformaciones para obtenerla. Bajo este supuesto, si los datos son normales y las variables independientes entre sí, los índices de capacidad de procesos multivariantes usan una forma elíptica como región del proceso, y realizan la comparación con el área de especificación.

El cálculo de los índices de capacidad multivariantes implica ciertas limitaciones. Por un lado, se debe suponer que los datos siguen una distribución normal. Y por otro lado, el tamaño muestral requerido debe ser suficientemente grande como para realizar una correcta estimación de la matriz de varianzas-covarianzas.

A raíz de estas limitaciones, **Krzysztof Ciupke (2015)** propone un nuevo enfoque, redefiniendo el área del proceso, la cual tendrá una forma elíptica pero definida empleando modelos unilaterales de la forma de un polinomio de segundo grado.

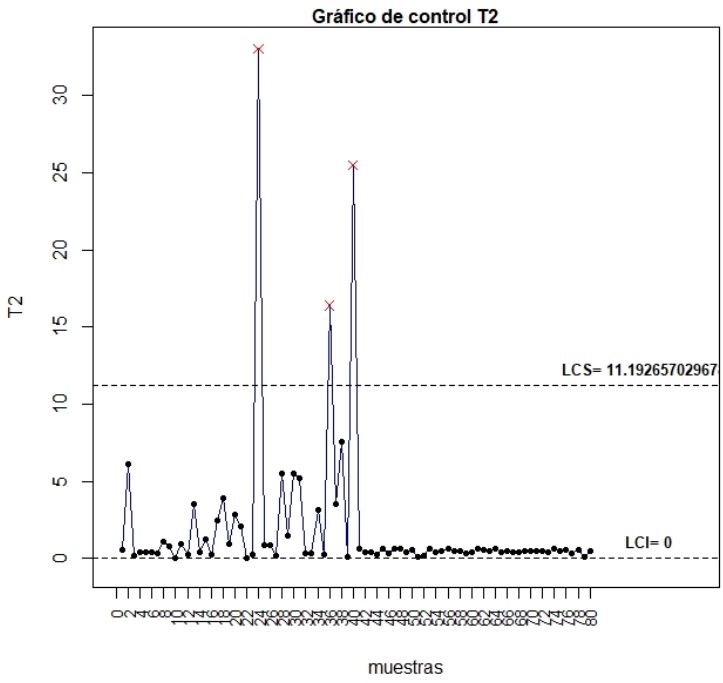


Figura 2.7: Gráfico de control T^2 Hotelling.

Vector de capacidad de procesos multivariantes

Con el propósito de obtener una extensión de los indices univariantes C_p,C_{pk} y C_{pm}, se define el vector de capacidad de procesos multivariantes. Este concepto fue introducido por **Shahriari et al. (2009)**, basándose en el trabajo de **Hubele et al. (1991)**. Este vector consta de las tres componentes y está desarrollado bajo la suposición de que proceso sigue una distribución normal multivariante.

$$[CpM, PV, LI] \tag{2.131}$$

- La primera de las componentes de dicho vector, CpM, representa el ratio entre las áreas o volúmenes entre las tolerancias y la región modificada del proceso.

$$CpM = \left[\frac{\prod_{i=1}^{p}(USL_i - LSL_i)}{\prod_{i=1}^{p}(UPL_i - LPL_i)}\right]^{1/p} \tag{2.132}$$

  siendo p el número de variables o características de la calidad a analizar en el proceso. Ambos áreas o volúmenes son rectángulos en procesos bivariantes y prismas rectangulares en el caso de tres variables. A continuación, se muestra en la Figura 2.8 la representación gráfica de las regiones mencionadas en el caso bivariante.

  Por un lado, el rectángulo exterior delimita el área de la tolerancia. Por otro lado, la región modificada del proceso se representa con el rectángulo azul, el cual circunscribe al elipsoide referente a la región del proceso. El elipsoide es un contorno de densidad de probabilidad centrado en la media del proceso, que es construido a partir de la descomposición espectral de la matriz de varianzas-covarianzas centrada en el vector de medias.

  Los límites correspondientes a la región del proceso, el límite inferior del proceso (LPL_i) y el límite superior del proceso (UPL_i), se calculan resolviendo el sistema de ecuaciones de la primera

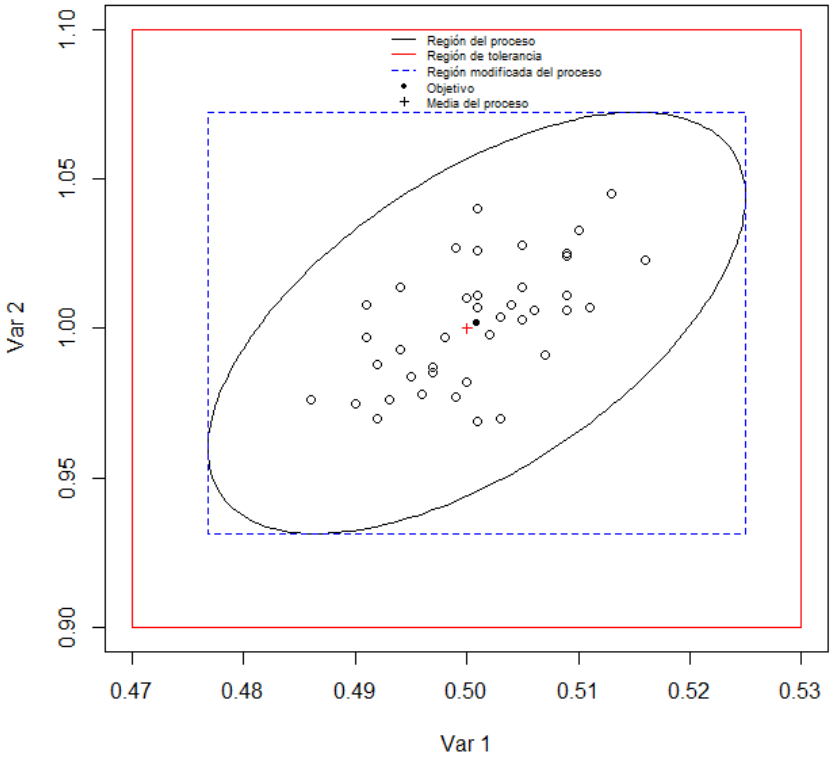


Figura 2.8: Representación gráfica de la región modificada del proceso.

derivada de la forma cuadrática de acuerdo a **Nickerson (1994)**.

$$(X_\mu)'(\Sigma)^{-1}(X_\mu) = \chi^2_{\alpha,p} \tag{2.133}$$

con una distribución χ^2 con p grados de libertad. Y por lo tanto, la solución de las ecuaciones resultantes para el cálculo de los límites son:

$$LPL_i = \mu_i + \sqrt{\frac{\chi^2_{\alpha,p} det(\Sigma_i^{-1})}{det(\Sigma^{-1})}} \qquad UPL_i = \mu_i + \sqrt{\frac{\chi^2_{\alpha,p} det(\Sigma_i^{-1})}{det(\Sigma^{-1})}} \tag{2.134}$$

donde $det(\Sigma_i^{-1})$ es el determinante de la matriz Σ resultante de eliminar la columna y la fila i-ésima.

Valores de CpM superiores a 1 indican que la región modificada del proceso es menor que la región de tolerancias.

- La segunda componente del vector, PV, es el vector de cercanía entre el objetivo y la media del proceso, expresada por la siguiente hipótesis:

$$PV = P\left(T^2 > \frac{p(m-1)}{m-p} F_{p,m-p}\right) \tag{2.135}$$

donde

$$T^2 = n(\bar{(}X) - \mu)'(S)^{-1}(\bar{(}X) - \mu) \tag{2.136}$$

y $F_{p,m-p}$ es una distribución F con grados de libertad p y (m-p).

PV toma valores entre 0 y 1, y valores próximos a 0 indican baja proximidad, es decir, que la media del proceso y el objetivo se encuentran muy distantes.

- La tercera componente del vector de capacidad de procesos es LI, el cual compara la localización de la región modificada del proceso y la tolerancia, mostrando cuando alguna de las partes de la región del proceso caen fuera de la región de tolerancia.

  Por lo tanto, LI=0 implica que al menos en una de las direcciones la región de tolerancia está siendo excedida, o dicho de otro modo:

$$LI = \begin{cases} 1 \textit{ si la región modificada está contenida en la región de tolerancia ingenieril} \\ 0 \textit{ en otro caso} \end{cases} \tag{2.137}$$

A modo de resumen, el vector de capacidad de procesos propuesto por **Shahriari et al. (2009)** ofrece una comparación de los volúmenes de las regiones, la cercanía entre centros y la extensión de dichas regiones

Índice multivariante de capacidad

Otro índice multivariante ampliamente aceptado es el MCpm, propuesto por **Taam et al. (1993)**. Este índice se define como la relación entre los volúmenes de los elipsoides de la región de tolerancia y la región de proceso dada por el elipsoide de control.

A diferencia del primer componente del vector de **Shahriari et al. (2009)** , que se calcula como la proporción de los rectángulos en caso bivariante o hipercubos para más dimensiones, el MCpm es la relación de los elipsoides. La región de tolerancia modificada es el elipsoide más grande construido en la región de tolerancia y con centro en el objetivo.

El índice se calcula como:

$$MCpm = \frac{vol.(R_1)}{vol.(R_2)}, \tag{2.138}$$

donde R_1 y R_2 son la región modificada de tolerancia y el elipsoide de confianza respectivamente. Este ratio puede ser estimado como

$$MCpm = \frac{C_p}{D} \tag{2.139}$$

con

$$C_p = \frac{vol.(\textit{región de tolerancia})}{vol.(\textit{región estimada del proceso})}. \tag{2.140}$$

El elemento del numerados es el hiperelipsoide que tiene el volumen determinado por

$$vol.(\textit{región de tolerancia}) = \frac{2\pi^{p/2} \prod_{j=1}^{p} l_j}{p\Gamma(p/2)} \tag{2.141}$$

donde l_j es la longitud de los semi-ejes. Por otro lado, en cuanto al denominador:

$$vol.(\textit{región estimada del proceso}) = |S|^{1/2} (Nk)^{p/2} \left[\Gamma(p/2+1)\right]^{-1} \tag{2.142}$$

donde K es el percentil de una distribución χ^2 y

$$D = \left[1 + \frac{m}{m-1}(\bar{X} - \mu)'(S)^{-1}(\bar{X} - \mu)\right]^{1/2}. \tag{2.143}$$

Por lo que finalmente,

$$MCpm = \frac{vol.(R_1)}{\left\{|S|^{1/2} (Nk)^{p/2} \left[\Gamma(p/2+1)\right]^{-1}\right\} * \left[1 + \frac{m}{m-1}(\bar{X} - \mu)'(S)^{-1}(\bar{X} - \mu)\right]^{1/2}} \tag{2.144}$$

- **Pan y Lee (2010)** recientemente apuntaron a que el índice de capacidad multivariante propuesto por **Taam et al. (1993)** sufre una sobreestimación cuando las variables o características de calidad no son independientes. Por lo tanto, se propone la siguiente corrección:

$$(X-T)'(A^*)^{-1}(X-T) = \chi^2_{p,1-\alpha} \tag{2.145}$$

donde cada elemento A^*_{ij} es obtenido de la siguiente manera:

$$A^*_{ij} = r_{ij}\left(\frac{USL_i - LSL_i}{2\sqrt{\chi^2_{p,1-\alpha}}}\right)\left(\frac{USL_j - LSL_j}{2\sqrt{\chi^2_{p,1-\alpha}}}\right) \tag{2.146}$$

el término r_{ij} es el coeficiente de correlación entre los elementos i y j. Finalmente, el índice propuesto para evitar la sobreestimación es:

$$NMCpm = \left(\frac{|A^*|}{|S|}\right)^{1/2} \tag{2.147}$$

Índice multivariante de capacidad basado en PCA

La aplicación de componentes principales viene fundamentado por las claras ventajas que ofrecen, la obtención de variables incorreladas y por supuesto, la reducción de la dimensionalidad.

Existen diversos métodos de obtención de índices multivariantes de capacidad basados en PCA, pero todos ellos parten de la descomposición espectral de la matriz de varianzas-covarianzas.

$$\Sigma = UDU' \tag{2.148}$$

donde la matriz U representa los autovectores y D la matriz diagonal de los autovalores $D = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$.

La componente principal i-ésima resulta sencillamente $PC_i = u_i'x$. Por lo que análogamente, las especificaciones ingenieriles resultan:

$$LSL_{PC_i} = u_i'LSL \qquad USL_{PC_i} = u_i'USL \qquad T_{PC_i} = u_i'T \tag{2.149}$$

Muchos autores han sido los que se han pronunciado acerca de la selección del número de componentes principales a tomar, no obstante, en líneas generales se suele tomar aquellas primeras componentes principales con las que se llega a explicar el 80 % de la variabilidad.

Existen diversos autores que han planteado diferentes estimaciones de los índices de capacidad basados en componentes principales.

- **Wang y Chen (1998)** plantean la extensión multivariante de los indices univariantes C_p,C_{pk} y C_{pm}.

$$MC_p = \left(\prod_{i=1}^{v} C_{p,PC_i}\right)^{1/v} \tag{2.150}$$

donde

$$C_{p,PC_i} = \frac{USL_{PC_i} - LSL_{PC_i}}{6\sigma_P C_i} \tag{2.151}$$

donde v hace referencia al número de componentes principales seleccionadas y $\sigma_{PC_i} = \sqrt{\lambda_i}$. De manera análoga, se obtienen los índices MC_pk y MC_pm sustituyendo en las expresiones los índices originales por los C_{pk,PC_i} y C_{pm,PC_i}:

$$C_{pk,PC_i} = min\left\{\frac{USL_{PC_i} - \mu}{3\sigma_{PC_i}}, \frac{\mu - USL_{PC_i}}{3\sigma_{PC_i}}\right\} \tag{2.152}$$

$$C_{pm,PC_i} = \frac{USL_{PC_i} - LSL_{PC_i}}{6\sqrt{\sigma^2_{PC_i} + (\mu - T)^2}} \tag{2.153}$$

En este método, todas las componentes principales reciben la misma importancia, incluso cuando es conocido que la primera de las componentes principales es la que mayor variabilidad explicada recoge, seguida de la segunda y sucesivamente.

- **Xekalaki y Perakis (2002)** propuso una modificación de los índices planteados por **Wang y Chen (1998)** aplicando una ponderación en función de la variabilidad explicada de cada componente principal.

$$MXPC_p = \frac{\sum_{i=1}^{v} \lambda_i C_{p,CP_i}}{\sum_{i=1}^{v} \lambda_i} \tag{2.154}$$

De manera análoga se ponderan los índices originales, obteniendo los correspondientes $MXPC_{pk}$ y $MXPC_{pm}$.

- Por su lado, **Wang et al. (2005)** propuso otra manera de emplear las componentes principales, empleando una ponderación de la media geométrica. Siguiendo esta idea, el índice MWC_p viene dada como se muestra a continuación:

$$MWC_p = \left(\prod_{i=1}^{v} C_{p,CP_i}^{\lambda_i} \right)^{1/\sum_{i=1}^{v} \lambda_i}. \tag{2.155}$$

Se ha de proceder de la misma manera para la obtención de los índices MWC_{pk} y MWC_{pm}.

Capítulo 3

Análisis exploratorio de datos para el control de calidad

En los Análisis Estadísticos de Control de Procesos, la normalidad e independencia de los datos es requerida, tanto en el caso univariante como en el multivariante. En el último de los casos, con el uso de subgrupos racionales, se hace uso del Teorema Central del límite, lo que permite en cierto modo asumir normalidad en los datos. No obstante, la ausencia de normalidad puede llevar a un análisis no satisfactorio y es por ello por lo que a continuación se analizarán diversos métodos de chequear la normalidad y la independencia de los datos, así como las posibles herramientas que permiten trabajar cuando alguna de estas dos características no se cumple.

3.1. Normalidad

En el caso univariante, existen tanto métodos gráficos, histogramas y QQ-Plots entre otros, como métodos más específicos como los test de normalidad, como pueden ser los conocidos test χ^2, Kolmogorov-Smirnov, D'Agostino, Shapiro-Wilks, Jarque-Bera etc. Sin embargo, se debe tener en cuenta que aunque los datos muestren normalidad en el análisis marginal, no implica que en el análisis conjunto los datos sean normales. En otras palabras, tener un vector X normal multivariante tal que $X = (x_1, x_2, \cdots, x_p)$, cada componente $x_i, con \quad i = 1, 2, \cdots, p$ tiene distribución normal $N(\mu_i, \sigma_i^2)$. No obstante, es importante conocer que no necesariamente la condición recíproca sea cierta, es decir, el hecho de que la densidad de cada uno de esos x_i sea normal univariante no implica necesariamente que el vector multivariante X sea normal.

En la literatura existen infinidad de métodos para chequear la normalidad multivariante, sin embargo en este apartado se estudiarán los cuatro test más potentes como son el test Shapiro-Wilk (1965), de Mardia (1970), el test de Henze y Zirkler (1990) y el test de Royston (1992).

- Test de Shapiro-Wilk (1965)

  El test propuesto en 1965 por Samuel Shapiro y Martin Wilk es uno de los más empleados tanto en el análisis univariante como en el multivariante. Para el caso multivariante, este test afirma que un vector aleatorio es normal multivariante si y solo si su estandarización lo es, es decir:

  $$X \in N_d(\mu, \Sigma) \Leftrightarrow z = \Sigma^{-1/2}(X - \mu) \in N_d(0, I_d) \tag{3.1}$$

  De manera que z tiene sus d componentes incorreladas y por lo tanto z es normal si y solo si sus componentes son normales. Esto se traduce en que bastará con comprobar si las d variables resultantes de la estandarización multivariante son normales. Por lo tanto, si $W_1, W_2, \cdots, W_d$ son los estadísticos de Shapiro-Wilk de cada componente de $z_1, z_2, \cdots, z_n$, entonces el estadístico

de Shapiro-Wilk multivariante es:

$$MVN = \frac{1}{d}\sum_{j=1}^{d} W_j \tag{3.2}$$

donde W_j son los estadísticos de Shapiro-Wilk univariantes.

- Test de Mardia (1970)

  Este test chequea grados de asimetría y curtosis, y es la generalización del caso univariante. Se debe recordar que los coeficientes de asimetría y curtosis provienen de los momentos de orden 3 y 4 con respecto a la media respectivamente. Por ello, **Mardia (1970)** plantea las expresiones para el cálculo de los coeficientes de asimetría (A_m) y curtosis (A_k) multivariante de la siguiente manera:

$$A_m = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^3 \tag{3.3}$$

$$A_k = \frac{1}{n}\sum_{i=1}^{n} r_{ii}^2 \tag{3.4}$$

  donde

$$r_{ij}^3 = \left[(x_i - \bar{x})'\Sigma^{-1}(x_j - \bar{x})\right]^3 \tag{3.5}$$

$$r_{ii}^2 = \left[(x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x})\right]^2 \tag{3.6}$$

  Mardia (1970,1974) propuso aproximaciones en la distribución de los coeficientes de asimetría (A_m) y curtosis (A_k). Para el coeficiente de asimetría, propuso la aproximación a una χ^2, aplicando:

$$A_m \frac{(p+1)(m+1)(m+3)}{6[(m+1)(p+1)-6]} \sim \chi^2_{\alpha,\frac{[p(p+1)(p+2)]}{6}} \tag{3.7}$$

  mientras que para el coeficiente de curtosis, empleaba una aproximación normal, siendo:

$$A_k \sim N(p(p+2), \frac{8p(p+2)}{m}) \tag{3.8}$$

- Test de Henze y Zirkler (1990)

  Henze y Zirkler (1990) proponen un test de normalidad multivariante basándose en la función característica empírica, y son muchos los estudios de simulación que sustentan el buen comportamiento de este test, **Thode (2002)**. El estadístico viene dado por:

$$T = \frac{1}{n^2}\sum_{k=1}^{p}\sum_{j=1}^{m} e^{-\frac{b^2}{2}|y_j - y_k|^2} + 2(1+2b^2)^{-m/2}\frac{1}{n}\sum_{j=1}^{m} e^{-\frac{b^2}{2(1+b^2)}y_j^2} \tag{3.9}$$

  donde

$$|y_j - y_k|^2 = (x_j - x_k)'S^{-1}(x_j - x_k) \tag{3.10}$$

$$y_j^2 = (x_j - \bar{x})'S^{-1}(x_j - \bar{x}) \tag{3.11}$$

$$b = \frac{1}{\sqrt{2}}\left[\frac{n(2m+1)}{4}\right]^{\frac{1}{(m+4)}} \tag{3.12}$$

El estadístico T propuesto por **Henze y Zirkler (1990)** tiene una distribución lognormal de media:

$$\bar{T} = 1 - (1 + 2b^2)^{-m/2} \left(1 + \frac{mb^2}{1 + 2b^2} + \frac{m(m+2)b^4}{2(1+2b^2)^2}\right) \tag{3.13}$$

y varianza:

$$\begin{aligned} Var(T) = 2(1 + 4b^2)^{-m/2} + 2(1 + 2b^2)^{-m} \left(1 + \frac{2mb^4}{(1+2b^2)^2} + \frac{3m(m+2)b^8}{4(1+2b^2)^4}\right) \\ -4w^{-m/2} \left(1 + \frac{3mb^4}{2w} + \frac{m(m+2)b^8}{2w}\right) \end{aligned} \tag{3.14}$$

con

$$w = (1 + b^2)(1 + 3b^2) \tag{3.15}$$

Por lo tanto, $T \sim L_{\mu,\sigma}$, con:

$$\mu = log \left[\left(\frac{\bar{T}}{Var(T) + \bar{T}}\right)^{1/2}\right] \tag{3.16}$$

$$\sigma = \left[log \left(\frac{Var(T) + \bar{T}^2}{\bar{T}^2}\right)\right]^{1/2} \tag{3.17}$$

- Test de Royston (1992)

  El test propuesto por **Royston (1992)** es la extensión al caso multivariante del test univariante de Shapiro Wilks. El estadístico de Royston es:

$$H = \frac{e \sum j = 1p R'_j}{p} \tag{3.18}$$

donde

$$R_j = \left\{\Phi^{-1}\left[\frac{\Phi^{-1}(-Z_j)}{2}\right]\right\}^2 \tag{3.19}$$

El cálculo de Z_j dependerá del número de observaciones. Por un lado, si $4 \geq n \geq 11$:

$$Z_j = \frac{log\left\{\gamma - [log(1 - W_j)] - \mu\right\}}{\sigma} \tag{3.20}$$

siendo W_j el valor del estadístico univariante del Test de Shapiro-Wilks. Por otro lado, si $12 \geq n \geq 2000$

$$Z_j = \frac{log(1 - W_j) + \gamma - \mu}{\sigma} \tag{3.21}$$

Para otros casos, el estadístico H viene dado por

$$e = \frac{m}{1 + (m - 1)\bar{c}} \tag{3.22}$$

con

$$\bar{c} = \frac{\sum j = 1p \sum k = 1m c_{ij}^5 - m}{m^2 - m} \tag{3.23}$$

$$c_{ij}^5 = r_{ij}^5 \left[1 - \frac{0{,}715(1 - r_{ij})^{0{,}715}}{v}\right] \tag{3.24}$$

siendo r_{ij} el coeficiente de correlación.

El estadístico H de Royston se aproxima a una distribución χ^2 con e grados de libertad.

3.1.1. Ausencia de normalidad

Cuando los datos no siguen una distribución normal, una de las alternativas puede ser aplicar cierta transformación a los datos, entendiendo transformación como la aplicación de una función matemática a los datos originales.

En un contexto multivariante existen dos enfoques diferentes para atacar el problema de la falta de la normalidad en los datos originales. Uno de los enfoques se centra en la normalidad marginal, mientras que el otro se enfoca directamente en la normalidad multivariante.

A continuación se presentan tres transformaciones, la transformación Box-Cox (BCT), la transformación Johnson (JT) y finalmente la transformación Box-Cox Multivariante (MBCT). Las dos primeras muestran un enfoque marginal a diferencia de la última, la transformación MBCT.

- Transformación Box-Cox (BCT)

  La familia de transformaciones Box-Cox fueron propuestos por **Box (1964)**, y se trata de aplicar la siguiente función:

$$y_i = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda} & cuando \quad \lambda \neq 0 \\ log(x_i) & cuando \quad \lambda = 0 \end{cases} \tag{3.25}$$

  donde sencillamente x_i es el dato original, λ es la potencia y y_i es el dato transformado.

  Una de las alternativas muy empleada en la práctica es la determinación del parámetro λ en función de aquel valor que maximiza el logaritmo de la función de verosimilitud.

  Esta transformación tiene la ventaja de que su aplicación es realmente sencilla, mientras que tiene la desventaja de que no permite valores negativos, aunque este último problema puede ser resuelto aíadiendo una constante a los datos originales.

- Transformación Johnson (JT)

  La familia de distribuciones Z, compuesta por la distribución ilimitada (SU), lognormal (SL) y la distribución limitada (SB), fue propuesta por **Johnson et al. (1949)**. Las distribuciones de Johnson son muy útiles para el contexto de la falta de normalidad de los datos y sus expresiones se muestran a continuación:

  - Distribución ilimitada (SU):

$$Z = \gamma + \eta sinh^{-1}\left(\frac{x-\varepsilon}{\lambda}\right) \tag{3.26}$$

  - Distribución lognormal (SL):

$$Z = \gamma + \eta lnh^{-1}\left(x-\varepsilon\right) \tag{3.27}$$

  - Distribución limitada (SB):

$$Z = \gamma + \eta ln\left(\frac{x-\varepsilon}{\lambda+\varepsilon-x}\right) \tag{3.28}$$

  donde se tienen las siguientes consideraciones:

$$\eta, \lambda > 0 \quad , \quad -\text{ínf} < \gamma < \infty \quad , \quad -\text{ínf} < \varepsilon < \infty \quad y \quad \varepsilon < x < \varepsilon + \lambda \tag{3.29}$$

  Chou et al. (1998) propuso una metodología sencilla, basada en la aplicación del método de distribución percentil para la transformación de datos no normales. El método trata de optimizar la transformación basándose en la estimación de parámetros propuesta por **Slifker y Shapiro (1980)**, buscando la transformación que mejor se ajusta a la distribución normal. La función óptima se selecciona en función de aquella transformación que obtiene los mejores resultados en test de Shapiro-Wilks.

- Transformación Box-Cox Multivariante (MBCT)

  Velilla (1993) propuso una extensión multivariante de la transformación de Box-Cox. Siendo $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_p]$ el vector de parámetros que tras aplicación de la siguiente transformación $X^{(\lambda)} = \left(X_1^{(\lambda_1)}, X_2^{(\lambda_2)}, \cdots, X_p^{(\lambda_p)}\right)$ sobre los datos originales obtiene una distribución normal multivariante con media $(\mu^{(\lambda)})$ y matriz de varianzas-covarianzas $(\Sigma^{(\lambda)})$. El parámetro λ es seleccionado en función de aquel valor que maximiza la función de máxima verosimilitud.

3.2. Independencia

La independencia en los datos es un requerimiento en el análisis estadístico de procesos, pero sin embargo, en la práctica no es muy usual chequear esta propiedad. De hecho, es lógico que en ciertos procesos y formas de medición empleadas, los datos están autocorrelados, debido a la dependencia temporal.

El primer chequeo para la independencia de los datos puede ser un gráfico sencillo que consiste en comparar las observaciones actuales y las ya procesadas, conocido también como gráfico de autocorrelaciones o correlograma, y fue introducido por **Box y Jenkins (1976)**.

La autocorrelación se calcula de la siguiente manera:

$$r_h = \frac{C_h}{C_0} \tag{3.30}$$

donde

$$-1 \leq r_h \leq 1 \tag{3.31}$$

con covarianza:

$$C_h = \frac{\sum_{t=1}^{m-h}(x_t - \bar{x})(x_{t+h} - \bar{x})}{m} \tag{3.32}$$

y varianza:

$$C_0 = \frac{\sum_{t=1}^{m}(x_t - \bar{x})^2}{m} \tag{3.33}$$

donde m es el tamaño de muestra y h es el retraso.

Es muy común que en la representación gráfica se grafiquen también los límites de control llamados en ocasiones bandas de confianza, calculadas como:

$$CL = \pm\frac{Z_{1-\alpha/2}}{\sqrt{m}}. \tag{3.34}$$

Cuando un r_h cae fuera de las bandas de confianza, se dice que existen evidencias de autocorrelación o dependencia.

3.2.1. Ausencia de independencia

Cuando se detecta dependencia temporal, el problema puede ser enfocado de dos maneras. La primera, consiste en emplear un gráfico de control específico como el propuesto por **Apley y Tsung (2002)** y **Kalagonda y Kulkarni (2004)** . La segunda opción, obviamente consiste en eliminar los efectos de la autocorrelación, que puede ser llevado a cabo a través de la descomposición en un modelo autorregresivo multivariante y analizar los residuos, que deberían presentar independencia.

Cuando se tienen datos multivariante autocorrelacionados,también es posible emplear otros métodos como estimadores robustos, residuos de series temporales o modelos en espacio de estados. Los más empleados suelen ser los estimadores robustos como los estimadores MCD (Minimum Covariance Determinant), estimadores MVE (Minimum Volume Ellipsoidal) Y medias truncadas (Trimmed Means).

El análisis de robustez que ataña a este contexto, se centra en el estudio del comportamiento del gráfico T^2 con presencia de outliers. El objetivo es principalmente calcular los límites de control y estimaciones de las medidas de posición y escala, para ajustar estas medidas, de forma que recojan de una manera más adecuada el comportamiento del proceso.

Para el método de estimadores MCD, la forma de proceder es que dada una muestra de tamaño n, se trata de obtener una estimación del parámetro de interés mediante una muestra de $h < n$ observaciones del conjunto de información, seleccionada de forma que minimice el determinante de la matriz de covarianzas, y utilizar esta muestra para obtener las estimaciones correspondientes. Por lo tanto, estos estimadores van a ser resistentes a outliers dado que no son utilizados en la determinación de la estimación correspondiente.

En el caso de los estimadores MVE, consiste en determinar el estimador de posición como el centro de un elipsoide de volumen mínimo que contenga al menos h puntos de las observaciones que tenemos, donde h toma el valor $E\left[\frac{n}{2}+1\right]$. En este método, el estimador de covarianza que se va a emplear viene determinado como el elipsoide multiplicado por un factor que haga que haga que el estimador sea consistente en distribuciones normales multivariantes. Este método sigue el siguiente algorítmo a la hora de calcular el estimador:

1. Tomar una muestra de tamaño $p+1$, cuyos índices se denotan por J, y calcular su media $\bar{x}_J$ y matriz de covarianzas S_J.
2. Calcular $m_J^2 = med\left(\left(x_i - \bar{x}_J\right) S_J^{-1} \left(x_i - \bar{x}_J\right)^t\right)$.
3. Determinar $det\left(m_J^2 S_J\right)^{1/2}$.
4. Repetir el proceso un número elevado de veces y elegir el J que minimice dicho determinante.

Se toma como estimador de la media x_J y como estimador de la matriz de covarianzas $(\chi^2_{p,0,5}) m_J^2 S_J$.

El tercer método de estimadores robustos es el de medias truncadas. Los elementos de la familia de medias truncadas utilizan para su definición un número real que denotaremos por $\alpha \in [0, 50)$. La media truncada a nivel α sería:

$$\hat{\mu}_\alpha(x) = \frac{1}{n-2a} \sum_{i=a+1}^{n-a} x_{[i]} \tag{3.35}$$

con $a = E\left[\frac{\alpha n}{100}\right]$ y $x_{[i]}$ la i-ésima observación de la muestra ordenada. Como se trata de un estimador para datos autocorrelados, se debe emplear también un estimador de correlación entre variables, S_{xy}^R, que se define como $S_{xy}^R = C(\beta_{xy}) \hat{\mu}\left(\left\{(X_i - \bar{X}_R)(Y_i - \bar{Y}_R)\right\}_i\right)$. Siendo $C(\beta_{xy})$ un coeficiente de consistencia cuyo valor depende del valor de truncamiento β y $\bar{X}_R$ e $\bar{Y}_R$ son las medias truncadas de las variables X e Y. Los coeficientes de consistencia se calculan como $C(\beta) = \frac{nC_{uv}(\beta)}{n-1}$, donde n es el tamaño muestral y $C_{uv} = \left[\hat{\mu}_\beta(X)\right]^{-1}$ donde $X \sim \chi_1^2$.

Capítulo 4

Caso de estudio real: control de las instalaciones de climatización de un hotel

Durante un gran periodo de tiempo, los métodos estadísticos multivariantes eran vistos como métodos demasiado complejos para su aplicación e interpretación en la industria. Sin embargo, gracias al reciente desarrollo y el avance realizado en las ciencias de la computación, la mayor parte de los cálculos correspondientes al análisis estadístico multivariante se puede lograr en un tiempo relativamente corto. Además, debido al avance en la visualización de datos y las técnicas de animación, presentar los resultados del análisis multivariante se vuelve cada vez más fácil e intuitivo. A este respecto, el software estadístico R proporciona un amplio conjunto de funciones que implementan tanto técnicas de control estadístico de la calidad univariantes, multivariantes y funcionales, como diferentes y variadas utilidades para la gestión de datos y la presentación de resultados, además de permitir la completa modificación y adaptación del código correspondiente según las necesidades del usuario. Dadas estas características, el software R, y en particular su paquetes qcc [**Scrucca L.**, **(2004)**] , qcr [**Miguel Flores**, **(2016)**] , qualityTools [**Thomas Roth**, **(2016)**] y MPCI [**Edgar S.F. y Michele S.**, **(2012)**] , entre otros, son las herramientas computacionales que se han empleado en el presente Trabajo Fin de Máster.

A continuación, se van a aplicar los distintos métodos analizados para el control de la calidad a un caso de estudio real. Específicamente, los datos corresponden a medidas tomadas en las instalaciones de una conocida cadena de hoteles en Latinoamérica. Estas medidas caracterizan al sistema de climatización (en este caso refrigeración) del que dispone este hotel en concreto. El intervalo temporal en el que fueron tomadas está comprendido entre el 1 de enero del 2019 al 31 de julio del mismo año. Las mediciones son horarias, es decir, existen 24 medidas u observaciones multivariantes por día.

4.1. Descripción de los datos

Se han tomado medidas correspondientes a una serie de variables críticas para la calidad del sistema de refrigeración de un hotel. Siguiendo las indicaciones de los encargados del control y mantenimiento de sus instalaciones, se han monitorizado variables relacionadas con la temperatura (miden el confort térmico y el funcionamiento de las instalaciones de climatización), el consumo (miden la eficiencia energética) y la ocupación (mide la actividad del hotel). Se estudian variables relacionadas básicamente con consumos y temperaturas. Tras un pequeño análisis inicial de las variables críticas para la calidad (CTQ) medidas, es posible eliminar ciertas variables que se obtenían como combinación lineal o relaciones sencillas de otras principales. Finalmente, las variables seleccionadas para proceder al estudio son las siguientes 18:

- **date**: Recoge las fechas en las que se producen las mediciones, desde el día 01/01/2019 hasta el 31/07/2019.
- **time**: Las mediciones se realizan cada hora, desde las 00:00 hasta las 23:00 del día.
- **consum_chillers**: Se trata de la medición de la potencia consumida por los dos chiller (intercambiadores de calor mediante los cuales se controla la temperatura en el hotel) con los que cuenta la instalación, en kW.
- **main_pot_activa**: Hace referencia a la medición del total de la potencia activa del sistema, en kW.
- **ocupacion_total**: Esta variable recoge la ocupación en cuanto a huéspedes del hotel, en porcentaje.
- **clima_on_total**: Mide el porcentaje de tiempo en el que el sistema se encuentra activo.
- **temp_clima_on_prom**: Temperatura en el interior del hotel, medida en ºC.
- **temp_ext**: Temperatura exterior, medida en ºC.
- **temp_ch1_in**: Temperatura de entrada en el chiller 1, en ºC.
- **temp_ch2_in**: Temperatura de entrada en el chiller 2, en ºC.
- **temp_ch1_out**: Temperatura de salida en el chiller 1, en ºC.
- **temp_ch2_out**: Temperatura de salida en el chiller 2, en ºC.
- **consum_ch1**: Potencia consumida por el chiller 1, en kW.
- **consum_ch2**: Potencia consumida por el chiller 2, en kW.
- **torre3_vent1**: Variable indicadora que indica que el ventilador 1 de la torre de refrigeración está en funcionamiento, torre3_vent1=1, o si por el contrario, está apagado, torre3_vent1=0.
- **torre3_vent2**: Variable indicadora que indica que el ventilador 2 de la torre de refrigeración está en funcionamiento, torre3_vent2=1, o si, por el contrario, está apagado, torre3_vent2=0.
- **pot_torre3_vent1**: Representa el consumo de potencia, en kW, del ventilador 1 de la torre de refrigeración.
- **pot_torre3_vent2**: Es el consumo de potencia, en kW, del ventilador 2 de la torre de refrigeración.

Nótese que la medida de consum_chillers es la suma de consum_ch1 y consum_ch2, y a pesar de haber eliminado algunas variables por estar relacionadas con otras más relevantes (según el criterio de los encargados de mantenimiento) como medidas de diferencia de temperaturas, estas se han mantenido para poder observar de dónde proviene el consumo total de los chiller, pudiendo diferenciar la potencia proveniente del chiller 1 y la del chiller 2.

Se puede realizar a priori el diagrama de Ishikawa, teniendo en cuenta las diversas variables identificadas como críticas para la calidad del sistema, para que las instalaciones HVAC del hotel funcionen adecuadamente, sean eficientes energéticamente y aporten el requerido control higrotérmico. El diagrama de Ishikawa queda representado en la siguiente Figura 4.1.

Es importante destacar que el presente caso, es un caso de estudio real. De hecho, el hotel de estudio ha encargado recientemente a una empresa gallega del sector energético el control del confort hidrotérmico y la eficiencia energética de sus instalaciones. Es por ello que el empleo de todas las técnicas estadísticas utilizadas en este Trabajo Fin de Máster tiene por objetivo resolver un problema real

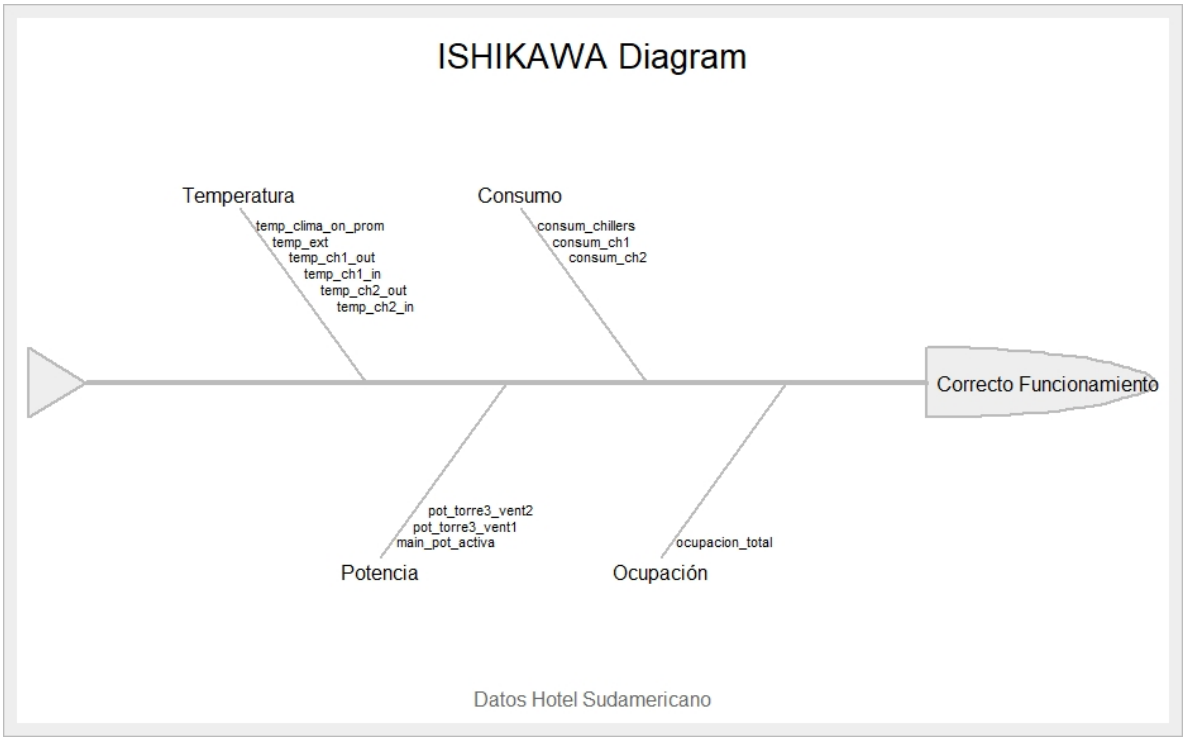


Figura 4.1: Diagrama de Ishikawa de las variables al sistema del hotel.

en el sector servicios, es decir, la mejora, control y mantenimiento de las instalaciones de climatización y aseguramiento de su eficiencia así como el del confort térmico. En consecuencia, este trabajo tiene un marcado carácter aplicado. En el presente estudio, se aporta una solución a la mejora, control, mantenimiento y detección de anomalías en las instalaciones mediante la aplicación de técnicas del control estadístico de la calidad.

A continuación, el hotel sudamericano quiere realizar un análisis de los datos, con el objetivo de mejorar su eficiencia y solucionar posibles problemas empleando técnicas de control de la calidad. A continuación, con el objetivo de conocer primeramente los valores representativos de posición y dispersión de las variables previamente introducidas, se construye la siguiente Tabla 4.1, que muestra un primer resumen de las variables numéricas, calculando sus medias, medianas máximos, mínimos, los cuantiles primeros y terceros y el número de valores perdidos o datos faltantes encontrados.

En la base de datos se puede observar que existen numerosos datos faltantes, lo que implica que existen observaciones multivariantes que no tienen todas sus componentes registradas. Debido a la frecuencia de medida y el gran número de observaciones existentes en la base de datos, se procede a tomar en cuenta solamente aquellas observaciones completas, con registros en todas sus variables. En resumen, la base de datos inicial contaba con un total de 5088 observaciones, que se ven reducidas a 4978 observaciones al tener en cuenta solamente aquellas observaciones completas.

En la Tabla 4.1, se observa en un primer vistazo el rango de valores que adopta cada una de las variables medidas. Resulta llamativo que las medias y las medianas no son muy coincidentes, al igual que los cuantiles no recogen medidas que recuerden a distribuciones normales. Esta diferencia entre la mediana y la media, como ocurre en las variables referentes al consumo de los chiller individual, lleva a pensar que existen valores extremos que hacen que la media se vea mucho más afectada que la mediana.

Tabla 4.1: Tabla resumen de los datos.

Variable	**consum_chillers**	**main_pot_activa**	**ocupacion_total**	**clima_on_total**
Mínimo	0.0	629.7	0.00	3.154
1ºQ	162.3	1015.2	62.63	44.021
Mediana	313.4	1167.8	72.61	51.841
Media	278.8	1150.1	70.62	54.113
3ºQ	351.7	1292.7	83.58	63.671
Máximo	656.3	1578.2	100.00	92.891
Variable	**temp_clima_on_prom**	**temp_ext**	**temp_ch1_out**	**temp_ch1_in**
Mínimo	20.80	18.46	0.00	0.00
1ºQ	22.12	26.20	11.13	14.27
Mediana	22.55	27.75	12.87	15.18
Media	22.56	27.93	13.37	15.34
3ºQ	22.99	29.30	14.81	16.02
Máximo	26.04	38.70	29.63	28.11
Variable	**temp_ch2_out**	**temp_ch2_in**	**consum_ch1**	**consum_ch2**
Mínimo	8.074	10.89	0.0	-2.0335
1ºQ	11.521	14.73	0.0	0.004
Mediana	12.445	15.66	155.9	0.004
Media	12.904	16.15	154.0	124.832
3ºQ	14.037	16.49	284.6	306.047
Máximo	28.905	458.01	388.7	484.663
Variable	**torre3_vent1**	**torre3_vent2**	**pot_torre3_vent1**	**pot_torre3_vent2**
Mínimo	0.000	0.000	0.000	0.000
1ºQ	1.000	1.000	0.900	0.497
Mediana	1.000	1.000	1.410	1.197
Media	0.890	0.797	1.381	1.175
3ºQ	1.000	1.000	1.900	1.717
Máximo	1.000	1.000	3.632	3.704

4.1.1. Funcionamiento del sistema de enfriamiento

A continuación se presenta el funcionamiento del chiller, que es una de las piezas fundamentales del sistema de enfriamiento empleado por el hotel para mantener la temperatura interior dentro del rango de tolerancia especificado por la normativa al respecto, la empresa y/o los clientes.

Los enfriadores o chiller se emplean principalmente para la generación de agua fría y producción de aire acondicionado en edificios. Este aparato opera basándose en el ciclo de Carnot, es decir, parte de un fluido refrigerante en estado líquido, el cual se fuerza a experimentar su evaporación debido a una baja de presión en el evaporador adonde además, toma calor del agua con la que indirectamente se pone en contacto. Es exactamente en ese lugar donde se produce el enfriamiento propiamente dicho del

agua. Posteriormente, el refrigerante en estado de vapor es comprimido por un compresor frigorífico obligándolo a recorrer el circuito de refrigeración. Seguidamente el refrigerante en estado de vapor, llega al condensador donde vuelve al estado líquido, liberando el calor sustraído del evaporador. En estos casos, en los chiller condensados por aire, el calor sale del refrigerante por acción de unos ventiladores.

Como se mencionaba anteriormente, el agua fría se produce en el evaporador y es la bomba centrifuga, bomba 1 en la Figura 4.2, la encargada de impulsar dicha agua fría al resto del edificio que se desea enfriar.

Esta bomba transporta el agua hasta la AHU o unidades manejadoras de aire, que son las encargadas de acumular todo el calor no deseado dentro del edificio. Finalmente, para terminar el recorrido realizado por el sistema 1, el agua en principio fría es devuelta al enfriador, pero en este caso, a una temperatura superior.

Este primer sistema se puede ver esquematizado en la Figura 4.2, donde se observa el recorrido del agua desde que sale del evaporador del chiller hasta el momento en el que regresa al mismo habiendo capturado parte del calor no deseado del edificio. En cuanto a la segunda parte del recorrido, cabe

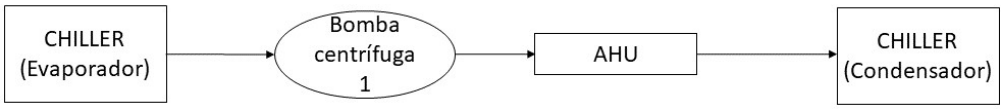


Figura 4.2: Sistema 1 de enfriado.

mencionar que el chiller consta de dos partes diferenciadas y aisladas, el evaporador y el condensador. Estas dos partes del chiller no están en contacto directo, pero existe un refrigerante que se encarga de la transferencia de calor entre ambas partes. El sistema cerrado conformado por las dos partes anteriores se puede observar en la Figura 4.3. Una vez el agua del primer sistema vuelve al chiller, el

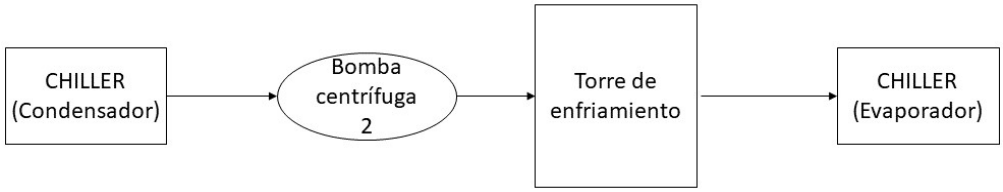


Figura 4.3: Sistema 2 de enfriado.

calor capturado es transferido a la parte del condensador del chiller, para que, acto seguido, mediante la bomba centrifuga o bomba 2, sea transportada a la torre de enfriado. Esta torre suele disponer de varios ventiladores que son los encargados de disipar el calor.

Una vez la torre de enfriado disminuye la temperatura del agua, esta es devuelta al chiller, pero en este caso a la parte del evaporador. De esta manera, el ciclo del agua queda cerrado y se hace posible el objetivo del sistema global, enfriar el edificio, como muestra la Figura 4.4. En resumen, en el evaporador, el agua pierde calor, mientras que el refrigerante absorbe el calor del agua, en la medida exacta. Después, en el condensador, el refrigerante vuelve al estado líquido cediéndole calor al aire (que eleva su temperatura). Esta liberación de calor, al efectuarse en un lugar distinto al original (enfriamiento del agua), consigue un efecto neto de "movimiento de calor"del proceso al ambiente.

4.1.2. Situaciones anómalas registradas

El presente caso real es un estudio controlado, es decir, los encargados del mantenimiento del hotel identificaron previamente una serie de anomalías en el sistema de climatización, consistentes en averías o debidas a encendidos y apagados no programados, entre otras causas. Este trabajo previo permitirá evaluar el desempeño de los gráficos de control aplicados.

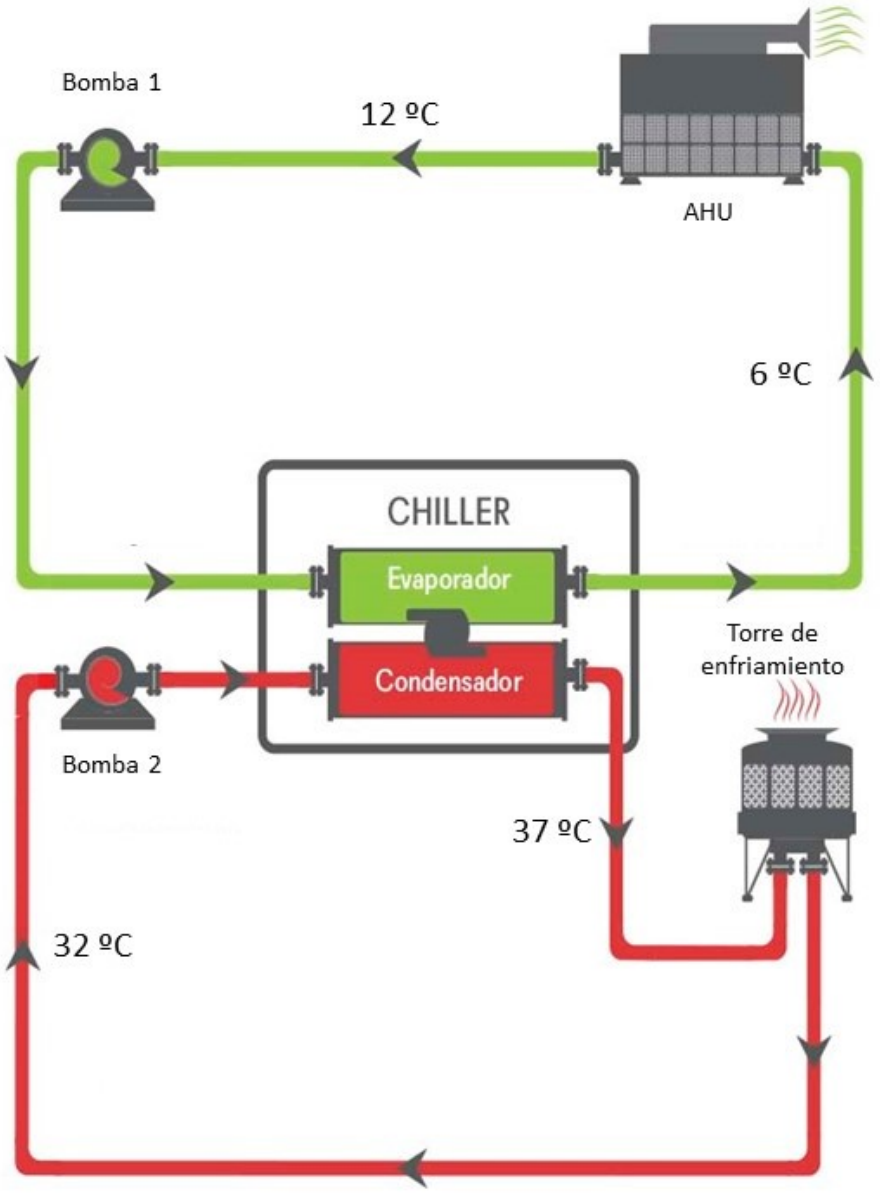


Figura 4.4: Sistema de enfriado completo.

A continuación se proceden a numerar dichas situaciones anómalas, así como las fechas y horas correspondientes.

- **El 07/01/2019 de 3:00 a 4:00 horas**: El día 7 de enero de madrugada ocurre un apagado total del sistema de refrigeración. El apagado total de ambos chillers produce a su vez una disminución obvia de la potencia y un aumento brusco de todas las temperaturas, las del chiller 2 en particular alcanzando el doble de su valor habitual. Aunque el apagado se acontece durante la noche, la temperatura del hotel es casi 1ºC mayor que la media.

- **Desde el 08/01/2019 a las 21:00 hasta el 09/01/2019 a las 7:00 horas**: En este periodo

se produce el cambio de uso del chiller 2 al chiller 1. Es decir, en este intervalo el chiller 2 deja de estar activo y se enciende en su lugar el chiller 1. Debido a este cambio, la temperatura en el interior del hotel sufre un incremento de hasta 1ºC por encima de la media.

- **Desde el 21/01/2019 a las 3:00 hasta el 22/01/2019 a las 9:00 horas**: En este intervalo de tiempo, la temperatura media de los chillers aumenta permanentemente 2ºC, este comportamiento es debido a una asignación de temperatura de consigna nueva.
- **Desde el 15/02/2019 a las 7:40 hasta el 17/02/2019 a las 9:25 horas**: En este caso, se apagó la torre de enfriamiento debido a motivos de mantenimiento del sistema, que lleva a que las medidas de todas las variables de la torre muestran un comportamiento muy irregular.
- **Desde el 18/02/2019 a las 19:00 hasta el 21/02/2019 a las 12:00 horas**: El día 18 de febrero produce el encendido simultáneo de ambos chillers, suceso que ocurre de manera automática debido a los requerimientos del sistema. A partir de este momento, y hasta el 21 de febrero a las 12:00 horas, las temperaturas de ambos chillers varían mucho más deprisa que en ningún otro momento del mes, y con cambios de mayor magnitud, que lleva a la temperatura interior del hotel a sufrir un incremento de casi 1ºC.
- **Desde el 14/04/2019 a las 00:00 hasta el 16/04/2019 a las 10:00 horas**: Hubo un fallo de emisión de datos desde el 14 abril desde las 00:00 hasta el 14 abril hasta las 8:00, pero al iniciarse de nuevo la emisión de los datos del sistema de control, la forma de acumular el dato de la temperatura de entrada no fue correcta por lo que se observa una curva de crecimiento exponencial, hasta la reprogramación del almacenamiento de datos el 16 de abril a las 10:00. A partir de este instante, el sensor, la recogida del datos y su almacenamiento fueron correctos (esto suele suceder cuando las compañías de control no son las mismas que gestionan y almacenan los datos).
- **El 13/05/2019 de 10:00 a 13:00 horas**: El 13 de mayo el sistema de refrigeración sufre un apagado total que produce una disminución esperada de potencia y un aumento brusco de todas las temperaturas, superando en algunos casos el doble de la temperatura habitual registrada en el mes de mayo. Se produce en horario diurno en vez de en mitad de la noche, lo que conlleva a una subida de la temperatura interior del hotel de 3ºC por encima de la media.
- **El 05/06/2019 de 10:00 a 13:00 horas**: El 5 de junio el sistema de refrigeración sufre de nuevo un apagado total. Como ocurre en el apagado general del pasado mes de mayo, el apagado total de ambos chillers produce una disminución obvia de potencia, así como un aumento brusco de todas las temperaturas. Este apagado se produce también en horario diurno en vez de en mitad de la noche, pero en este caso, conlleva a que la temperatura interior del hotel se sitúe hasta 2ºC por encima de la media. El incremento sufrido es algo inferior al sufrido en mayo, posiblemente provocado debido a la diferencia entre ambos meses de la temperatura exterior, que en junio se sitúa 1ºC por debajo de la temperatura exterior media del mes de mayo.

En resumen, si se suman todas las observaciones correspondientes a las anomalías previamente enumeradas, se obtienen un número total de 168.

4.2. Análisis exploratorio

A continuación se muestra un estudio descriptivo de la base de datos que define las instalaciones de climatización del hotel. En el análisis de datos reales, es fundamental realizar un estudio exploratorio antes de aplicar técnicas del control estadístico de la calidad como son los gráficos de control o los índices de capacidad. La identificación de la principales fuentes de variación, posibles grupos, rango normal de funcionamiento, distribución de las variables aleatorias, análisis de tendencias y correlación, entre otras muchas tareas, es esencial para el correcto tratamiento de los datos y para la adecuada aplicación de técnicas estadísticas.

4.2.1. Análisis previo general

En primer lugar, con el objetivo de conocer los valores representativos de las diversas temperaturas que definen el sistema así como su variación mes a mes en el periodo de estudio comprendido entre enero y julio de 2019, se representan las temperaturas medianas de cada mes en la Figura 4.5 (se escoge graficar las medianas para evitar la dispersión debida a los posibles datos atípicos en las medias). La

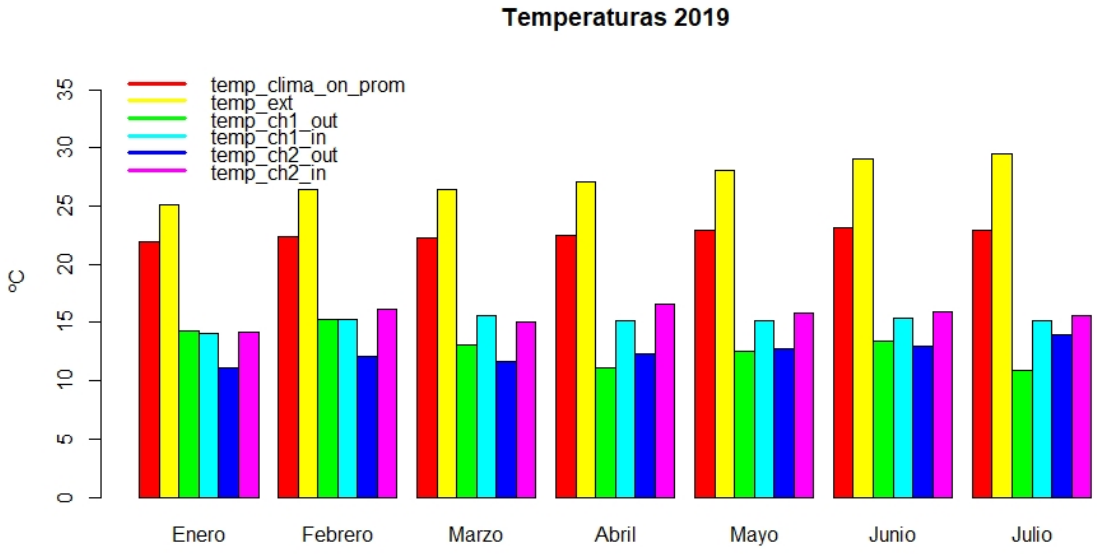


Figura 4.5: Medianas mensuales de las diferentes variables de temperatura que definen el sistema de climatización del hotel.

temperatura exterior es es la que tiene un mayor valor mediano en todos los meses, siento siendo mayor su diferencia con respecto a las demás conforme se acerca el verano. Por otro lado, la temperatura interior del hotel muestra la misma tendencia de incremento, pero de una magnitud menor a la exterior, gracias al sistema de enfriamiento empleado por el hotel. En cuanto a las cuatro temperaturas restantes, son aquellas que corresponden a los procesos de entrada y salida de los chiller. Las temperaturas del proceso de salida, tanto del chiller 1 como del chiller 2, deberían ser superiores a las temperaturas del proceso de entrada, por el propio sistema de funcionamiento de la máquina de refrigeración. Sin embargo, en media parece no existir grandes diferencias entre las temperaturas de entrada y salida en los meses de enero y febrero del chiller 1. No obstante, las temperaturas de entrada y salida asociadas al chiller 2 parecen mostrar mayores diferencias.

En la Figura 4.6, se observa la media mensual de parámetros asociados a los consumos del sistema. La potencia activa es aquella que mayor valores registra en todo el periodo, y sufriendo un ligero incremento en cada uno de los meses. Resulta llamativo el comportamiento del consumo de los chiller, que en los tres primeros meses se observa un incremento en su valor, mientras que de marzo a abril ocurre un descenso de más de un 37 %. En el mes de mayo sigue dándose un incremento de este aunque en menor proporción, sobre un 15.5 %. Finalmente, en los meses de junio y julio vuelve a incrementarse, pero sin llegar a los valores tomados en el mes de marzo, donde se registran su media máxima. Estos comportamientos parecen ajustarse a la ocupación del hotel, siendo mayor cuando la ocupación es mayor, es decir en temporada baja.

A continuación, en la Figura 4.7, se puede ver el comportamiento de las medias generales de los porcentajes de ocupación del hotel y funcionamiento del sistema. La ocupación del hotel alcanza sus

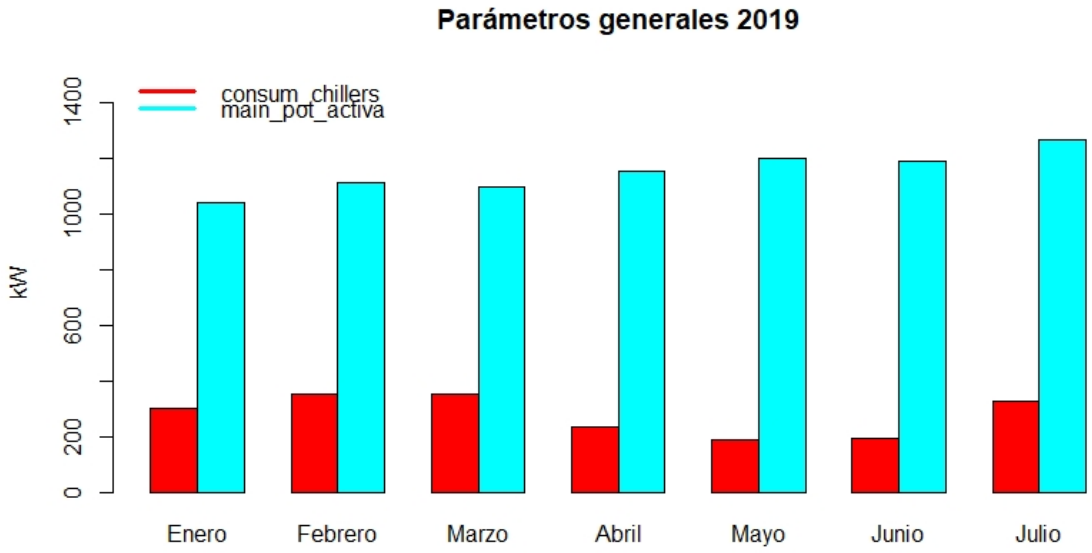


Figura 4.6: Medias generales de los consumos.

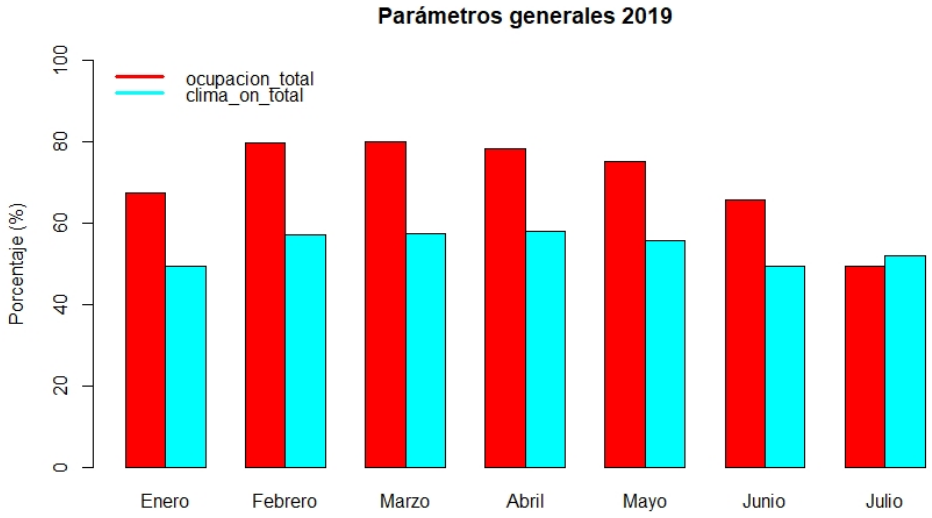


Figura 4.7: Medias generales de los porcentajes de ocupación y funcionamiento del sistema.

máximos en los meses de febrero a abril, disminuyendo considerablemente en los meses de junio y sobre todo julio. El consumo de los chiller, se observa que sigue la tendencia de la variable ocupación, coincidiendo que se observan valores elevados de consumo en temporada alta, de febrero a abril. El aumento del consumo en julio se puede deber a que, aunque haya una menor ocupación, ha aumentado la temperatura media y con ello el consumo de energía en climatización.

En cuanto a la torre de enfriamiento, cuenta con dos ventiladores, los cuales funcionan de manera independiente, como muestra la Figura 4.8, donde se observa que, el consumo debido al funcionamiento de los ventiladores es mayor en aquellos meses en los que el consumo, la ocupación y las temperaturas asociadas a los chiller son elevados. Son resultados esperados, ya que los ventiladores son los encargados

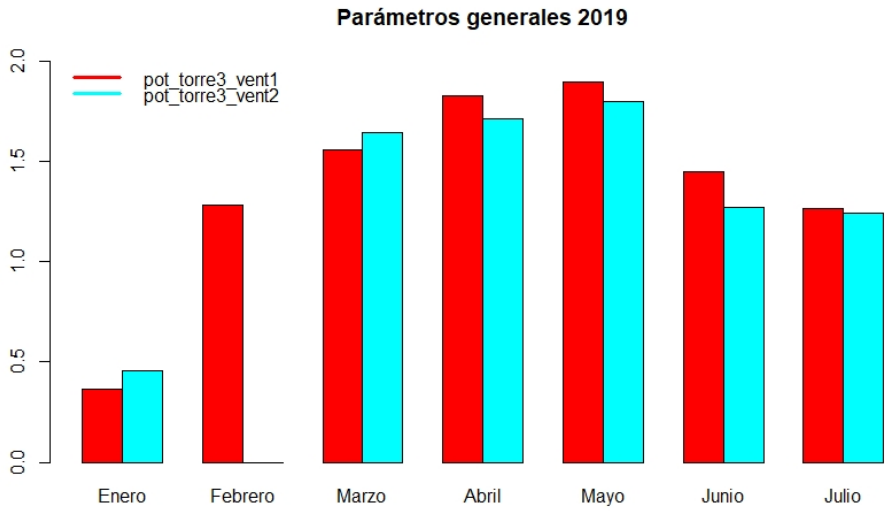


Figura 4.8: Medias generales asociadas al consumo energético de los ventiladores de la torre de enfriamiento.

de disipar en gran parte el calor extraído del interior del hotel. Lo extraño ocurre el mes de febrero, donde el ventilador 2 parece estar apagado, siendo el ventilador 1 el que funciona en este periodo.

En la Figura 4.9 se observa la proporción del consumo energético correspondiente a cada uno de los chiller la que trabajan los chiller para cada uno de los meses. En enero es el chiller 2 el que asume prácticamente todo el consumo. A medida que pasan los meses, esta proporción se va invirtiendo hasta llegar al extremo opuesto en los meses de mayo, junio y julio, donde es el chiller 1 el que supone prácticamente el 100 % del consumo total asociado a los chiller.

A continuación se realiza un estudio exploratorio univariante y multivariante de todas las variables críticas para la calidad de las instalaciones del hotel. El objeto es conocer la distribución de cada variable aleatoria, sus valores característicos, identificar la existencia de varias posibles poblaciones y, por otro lado, detectar y medir la relación de dependencia entre variables. Este análisis exploratorio previo es fundamental para poder aplicar técnicas de control estadístico de procesos univariantes y multivariantes. Nótese que se emplean estimaciones de la densidad no paramétrica tipo núcleo.

En la Figura 4.10 se pueden ver las estimaciones no paramétricas de la función de densidad, así como los histogramas, de las diferentes variables críticas para la calidad de las instalaciones del hotel. Mediante su observación, se puede concluir que las variables no tienden a seguir una distribución normal , pudiendo tener ciertas dudas en el caso de la temperatura asociada al proceso de entrada del chiller 1 y las temperaturas interior y exterior del hotel. En gran parte, muchas de las variables parecen mostrar multimodalidad, que puede provenir del hecho de que el registro de los datos ocurre a lo largo del periodo de enero a julio,periodo en el cual hay diferencias de temperatura exterior de mes en mes, aún estando el hotel en una zona de clima cálido subhúmedo.

Se observa también cómo para muchas de las variables se observan distribuciones multimodales, es decir, que se podría decir que existen dos o más poblaciones. Esto puede deberse a la diferencia ocupacional del hotel en función de los meses según la temporada, alta o baja.

En el caso de la variable referente al consumo de los chiller, existen claramente dos modas, pro-

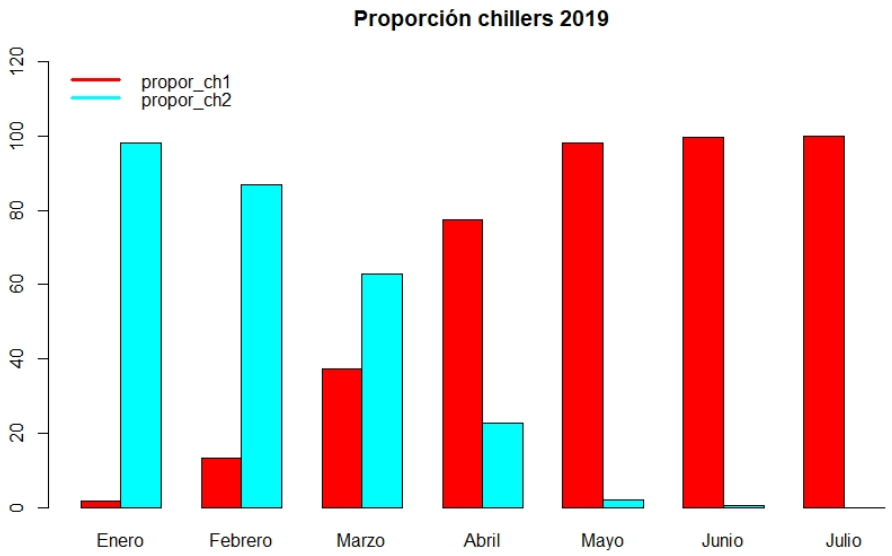


Figura 4.9: Proporciones medias del consumo horario de cada chiller con respecto al consumo total de ambos.

bablemente a la diferenciación entre el consumo del chiller 1 y del chiller 2. En cuanto a las variables del consumo de potencia de los ventiladores de la torre de enfriamiento, observando la multimodalidad que muestran, parece que hay tres modos de funcionamiento: apagado, consumo medio y consumo extremo.

Con idea de conocer el tipo y el grado de relación lineal o correlación entre pares de variables, se ha obtenido la siguiente matriz de diagramas de dispersión, mostrada en la Figura 4.11).

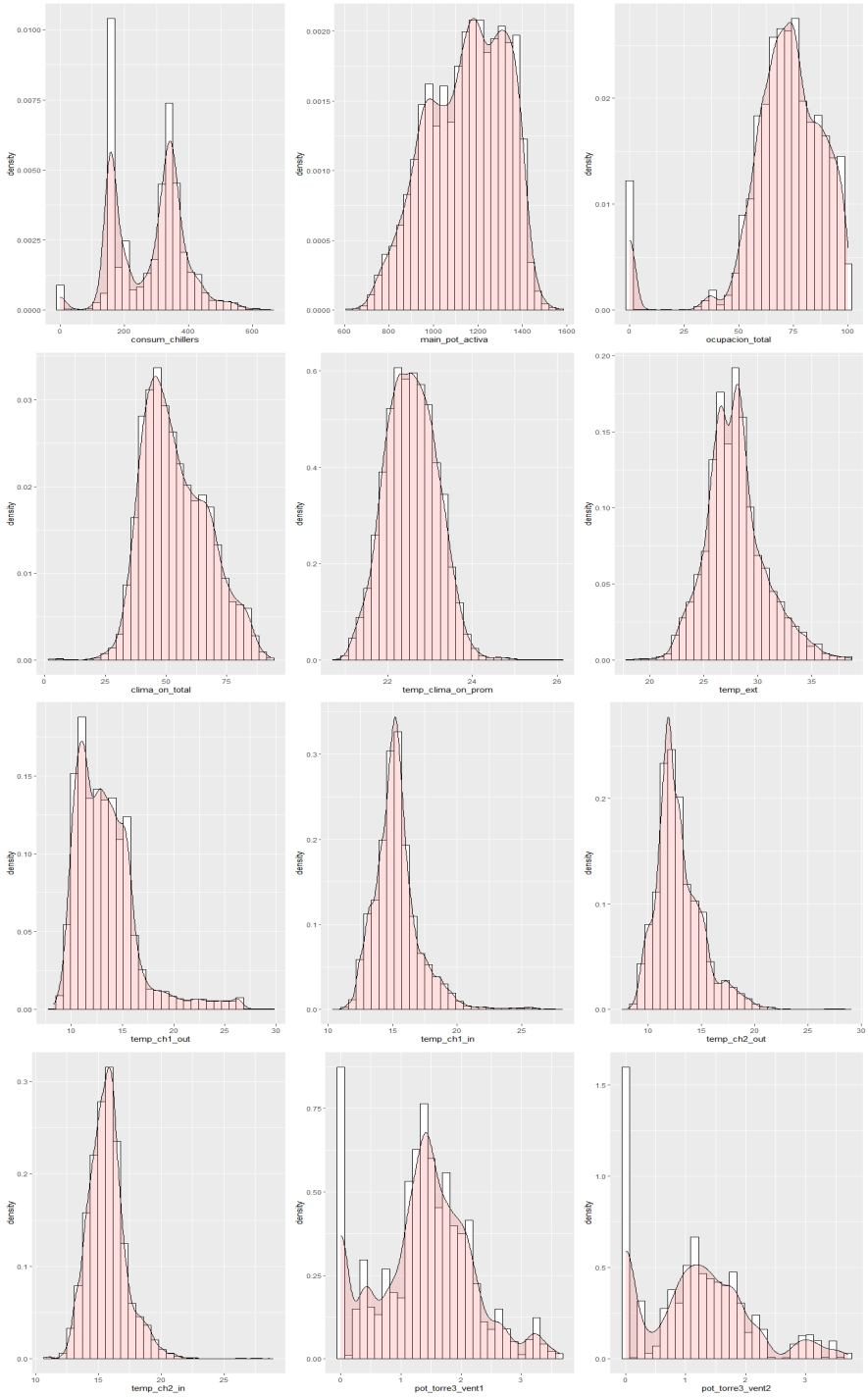


Figura 4.10: Histogramas y estimación de la función de densidad de las distintas variables críticas (mediciones horarias) para la calidad de las instalaciones del hotel.

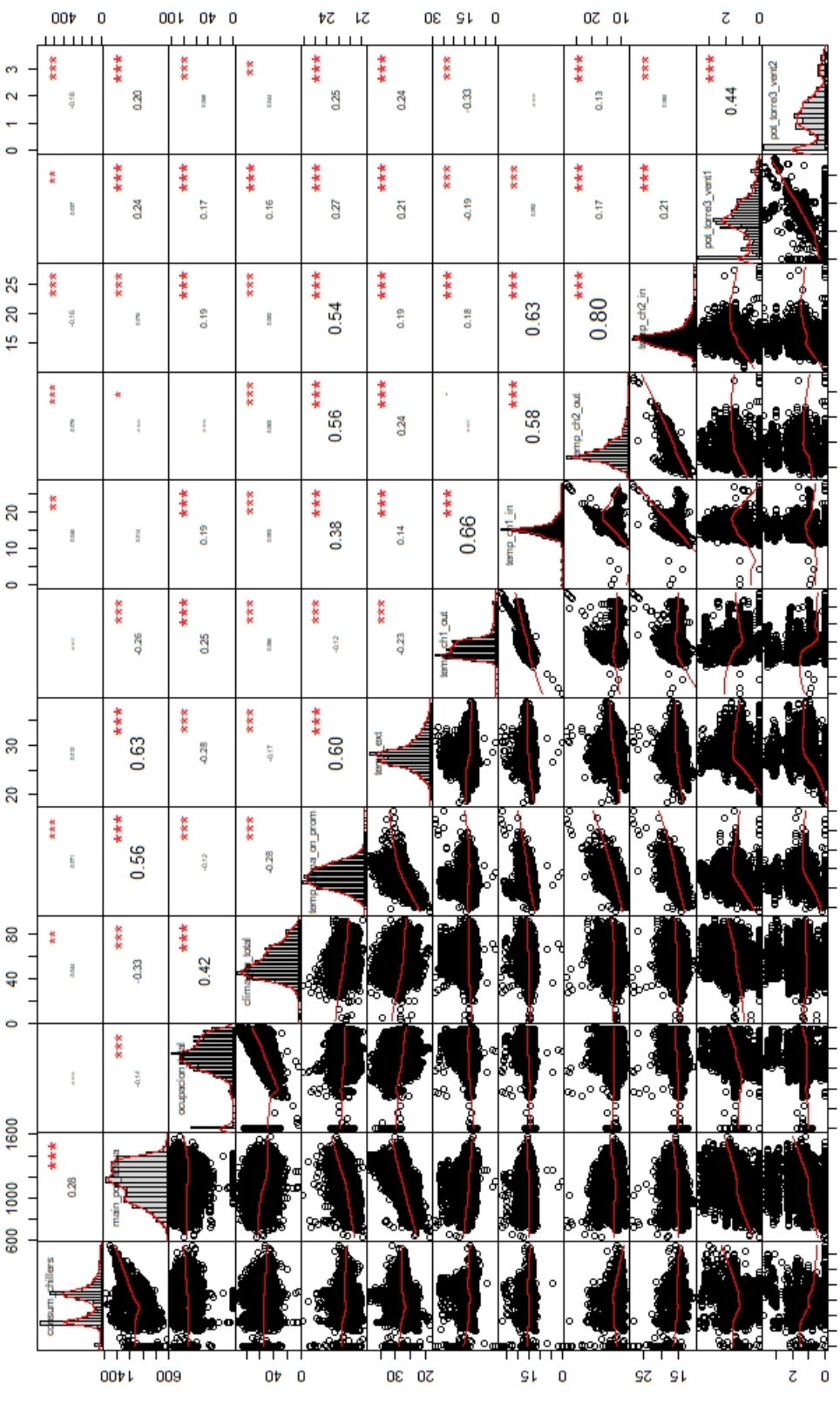


Figura 4.11: Matriz de diagrama de dispersión y correlograma, incluyendo histogramas y estimación no paramétrica de la función de densidad para cada variable.

Como se puede ver en la Figura 4.11, hay diversas variables que están moderadamente correladas. Entre ellas, se puede resaltar la moderadamente alta correlación lineal entre las variables temperatura exterior y la potencia activa del sistema y la temperatura interior del hotel, con índices de correlación en torno a 0.6. Así como las temperaturas del proceso de entrada de los chiller respecto de cada una de las temperaturas de sus procesos de salida. También se muestra una alta correlación lineal entre la potencia de ambos ventiladores de la torre de enfriamiento. Todas estas correlaciones son positivas, es decir, a medida que una de ellas aumenta, la otra también lo hace. También es interesante la relación existente entre la ocupación y el consumo del sistema total, denotando que existen consumos más elevados durante la temporada de alta ocupación en el hotel.

Sin embargo, no se encuentran valores de correlación lineal tan elevados en cuanto a correlaciones negativas se refiere, aunque existe una correlación de entorno a -0.3 entre la potencia activa del sistema y el porcentaje de tiempo que el sistema está activo, la potencia activa también muestra alta correlación negativa con la temperatura de salida del chiller 1. También entorno al mismo valor se obtiene la correlación existente entre las variables ocupación de hotel y temperatura exterior. La temperatura del proceso de salida del chiller 1 muestra una correlación negativa con la potencia del ventilador 2 de la torre de enfriamiento. Finalmente, se resalta también la relación entre temperatura interior del hotel y el porcentaje de tiempo que el sistema de enfriamiento está activo.

Con el fin de poder identificar de una forma rápida y sencilla el grado de dependencia lineal entre variables, en la Tabla 4.2 se muestra la matriz de correlaciones.

Tabla 4.2: Tabla de correlaciones lineales de las variables.

	consum_chillers	main_pot_activa	ocupacion_total	clima_on_total	temp_clima_on_prom	temp_ext	temp_ch1_in	temp_ch2_in	temp_ch1_out	temp_ch2_out	pot_torre3_vent1	pot_torre3_vent2
consum_chillers	1.00	0.28	-0.01	-0.04	-0.07	0.01	-0.02	0.05	-0.08	-0.16	-0.04	-0.18
main_pot_activa	0.28	1.00	-0.14	-0.33	0.56	0.63	-0.26	0.01	-0.03	0.08	0.24	0.20
ocupacion_total	-0.01	-0.14	1.00	0.42	-0.12	-0.28	0.25	0.19	-0.02	0.19	0.17	0.07
clima_on_total	-0.04	-0.33	0.42	1.00	-0.28	-0.17	0.09	0.09	0.08	0.08	0.16	0.04
temp_clima_on_prom	-0.07	0.56	-0.12	-0.28	1.00	0.60	-0.12	0.38	0.56	0.54	0.27	0.25
temp_ext	0.01	0.63	-0.28	-0.17	0.60	1.00	-0.23	0.14	0.24	0.19	0.21	0.24
temp_ch1_in	-0.02	-0.26	0.25	0.09	-0.12	-0.23	1.00	0.66	-0.03	0.18	-0.19	-0.33
temp_ch2_in	0.05	0.01	0.19	0.09	0.38	0.14	0.66	1.00	0.58	0.63	0.09	-0.02
temp_ch1_out	-0.08	-0.03	-0.02	0.08	0.56	0.24	-0.03	0.58	1.00	0.80	0.17	0.13
temp_ch2_out	-0.16	0.08	0.19	0.08	0.54	0.19	0.18	0.63	0.80	1.00	0.21	0.08
pot_torre3_vent1	-0.04	0.24	0.17	0.16	0.27	0.21	-0.19	0.09	0.17	0.21	1.00	0.44
pot_torre3_vent2	-0.18	0.20	0.07	0.04	0.25	0.24	-0.33	-0.02	0.13	0.08	0.44	1.00

4.2.2. Análisis funcional

Debido a la naturaleza del registro de los datos, es posible estudiar las variables críticas para el funcionamiento del sistema HVAC como variables funcionales. Tratando los datos desde este punto de vista, es posible obtener la Figura 4.12, en la que se muestran las curvas de consumo energético reales junto con las curvas suavizadas, empleando el paquete fda.usc de R y aplicando una suavización

b-spline. En estos gráficos se representan los valores del consumo de los chiller en función de la hora del día, es decir, cada curva o dato funcional se refiere a un día del registro.

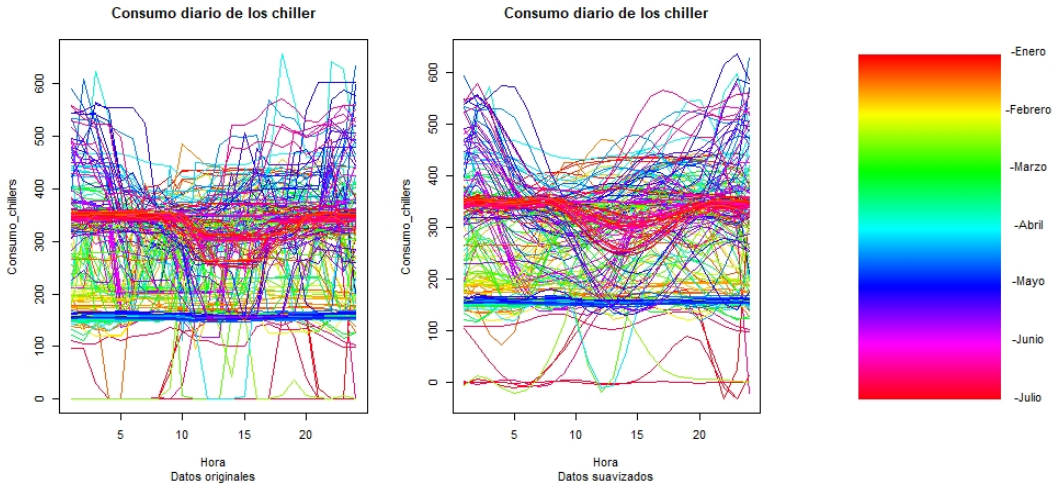


Figura 4.12: Consumo diario de los chiller, incluyendo las curvas originales (panel izquierdo) y las curvas suavizadas (panel derecho).

En los datos suavizados, es posible visualizar de manera más clara la tendencia del consumo en general a lo largo del día, donde se ve que los valores más elevados son aquellos referentes a la noche y la madrugada, obteniendo consumos de los chiller algo inferiores en las horas centrales del día. Esto puede deberse a la ocupación del hotel, que en las horas centrales del día es inferior, incluso que los huéspedes en estas horas no se encuentren en las habitaciones sino en las zonas comunes exteriores.

Asumiendo que la potencia activa consumida es una variable aleatoria funcional, en la Figura 5.10 se representan las curvas diarias de potencia activa consumida, incluyendo las tanto las curvas originales como las curvas suavizadas, empleando nuevamente el paquete fda.usc de R y aplicando una suavización b-spline con parámetro 9.

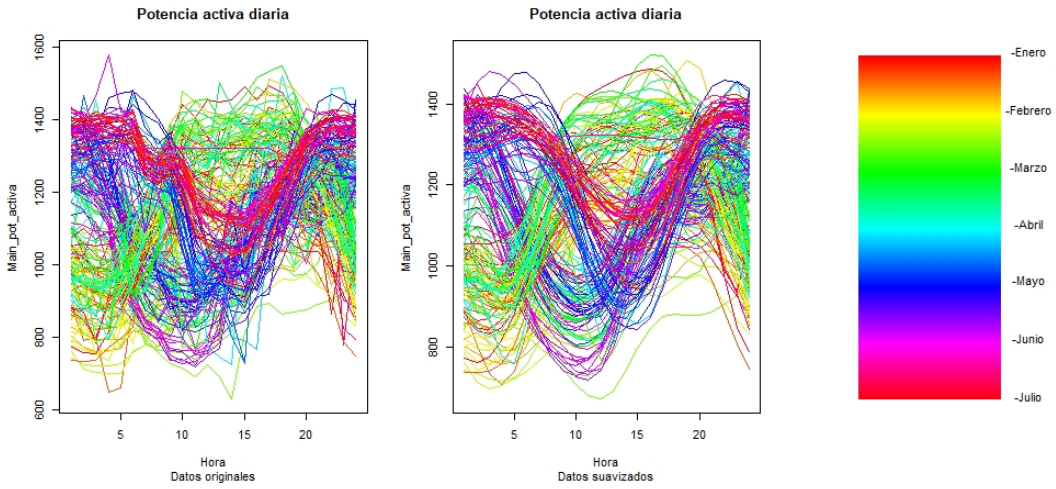
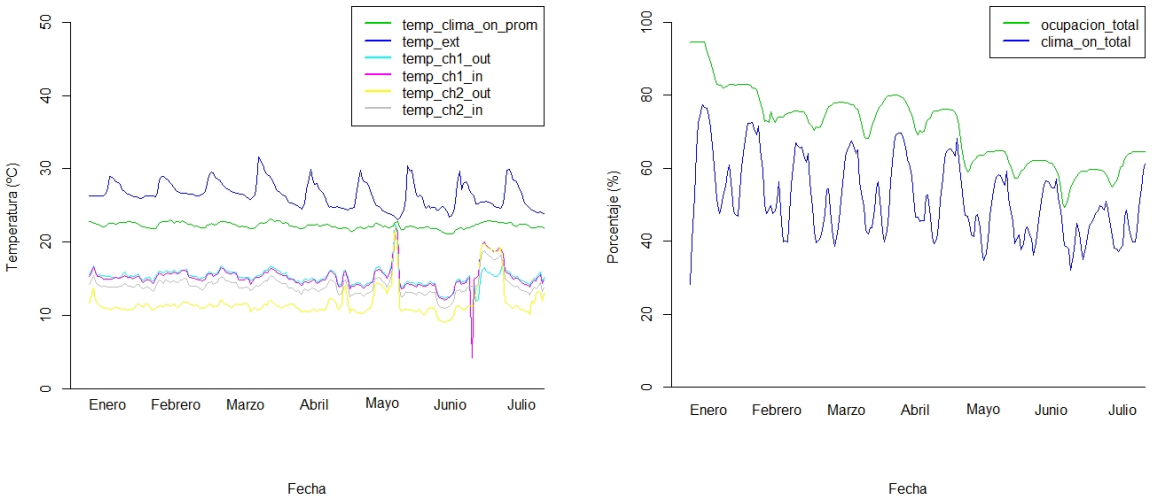


Figura 4.13: Curvas de la potencia activa diaria, incluyendo las curvas originales (panel izquierdo) y las curvas suavizadas (panel derecho).

Observando el valor de dicha potencia en función de la hora del día, donde cada curva o dato

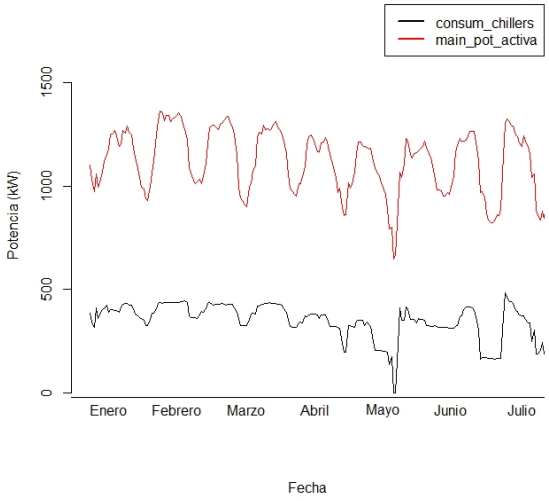
funcional se refiere a un día del registro, se diferencian curvas con forma sinusoidal, desfasadas entre ellas, pero aún así distinguiendo dos grandes grupos, por un lado, aquellas curvas referentes a días en los que en las horas centrales del día de 11:00-16:00 se obtienen valores bajos de potencia y altos en las horas cercanas a la madrugada y, por otro lado, aquel grupo de curvas que se comportan justamente de manera inversa, con valores elevados en las horas centrales del día y mínimos en la madrugada.

Con el objeto de poder estudiar la tendencia de las series correspondientes a cada variable, incluyendo el periodo y magnitud, resulta conveniente observar el comportamiento de las variables a lo largo de todo el intervalo, enero 2019-julio 2019. Estas tendencias se pueden observar en la Figura 4.14, la cual se compone de dos gráficas donde se visualizan las variables de menor magnitud en términos absolutos (panel izquierdo) aparte de las de mayor magnitud (panel derecho). En prácticamente



(a) Temperaturas.

(b) Ocupación y % Funcionamiento.



(c) Potencia.

Figura 4.14: Comportamiento de las variables (enero 2019-julio 2019).

todas las curvas se observa cierta periodicidad, alcanzando de manera simultánea los picos máximos y los mínimos. Existen algunos valores en fechas señaladas que podrían ser anomalías.De hecho, estos valores se corresponden con aquellas fechas en las que el hotel registra anomalías en el sistema por causas asignables apuntadas por el equipo de mantenimiento, como apagados generales o fallos en el sistema. Este tipo de comportamientos se pueden observar en las fechas cercanas al 15 de mayo, donde se produce un apagado general del sistema. En estas fechas se da la situación extraordinaria en la que las temperaturas asociadas a los chiller sufren un gran incremento, mientras que variables como la potencia o el consumo de los chiller disminuyen en magnitud.

Para conocer el comportamiento del consumo de los chiller, se analizan las variables consumo_chillers, consum_ch1 y consum_ch2, que miden el consumo total de ambos chiller, y los consumos marginales del chiller 1 y el chiller 2 respectivamente. En la Figura 4.15 se muestran todas las observaciones

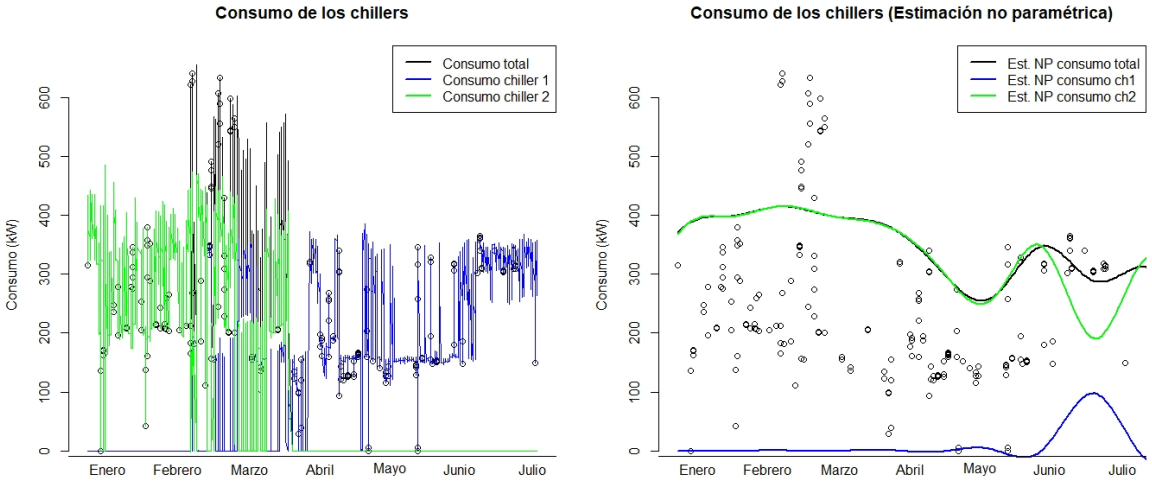


Figura 4.15: Potencia consumida de los chiller (panel izquierdo) y estimación no paramétrica (panel derecho).

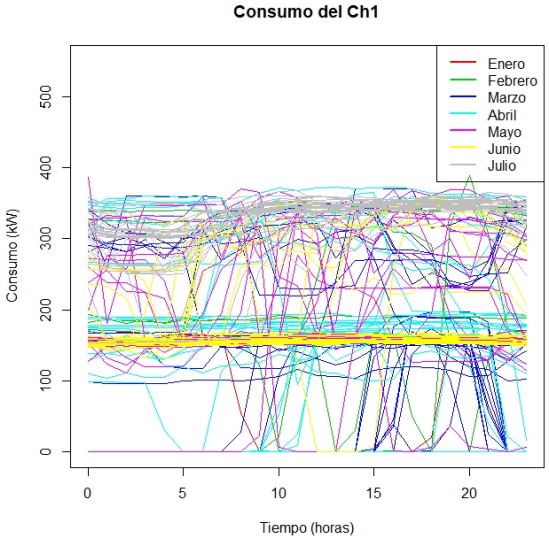
medidas y las estimaciones no paramétricas de cada una de las variables previamente mencionadas, empleando una estimación por splines cúbicos, que nos aporta una representación de la tendencia más clara. Como se observa, en la primera mitad del año, es el chiller 2 el que prácticamente computa el total del consumo de los chiller, mientras que en los meses de mayo, junio y julio, es el chiller 1 el que predomina, siendo en numerosas ocasiones nulo el consumo del chiller 2. No obstante, este último periodo tiene un consumo total inferior a los primeros meses, que vuelve a coincidir con la época de temporada baja del hotel.

Continuando con el análisis del consumo de los chiller, se procede a graficar como dato funcional dicho consumo. En las Figuras 4.16a y 4.16b, se ven las curvas de consumo energético diario y las medias funcionales correspondientes al chiller 1 en función de la hora del día. En las Figuras 4.16c y 4.16d, sin embargo, se grafican aquellos datos referentes al segundo de los chiller.

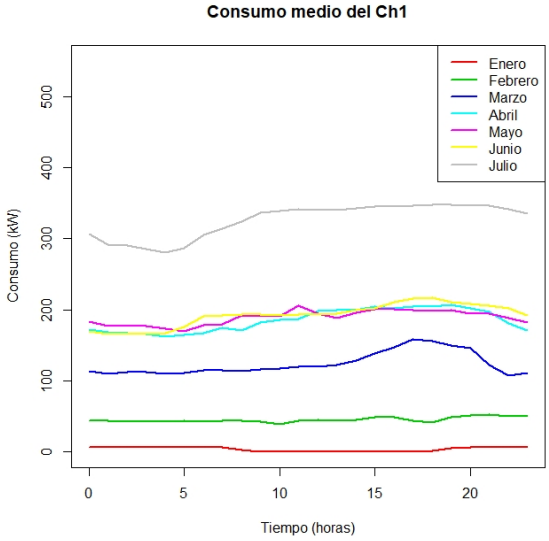
Aquí se ve que los chillers funcionan más o menos dependiendo de la época del año. En abril, mayo, junio y julio es el chiller 1 el que está definido por un mayor consumo,valores muy elevados rondando los 350 kW, mientras que en enero, febrero y marzo consume más el chiller 2, en todo el rango horario.

Además, es particularmente interesante la forma de las curvas medias, pues el chiller que consume más (ya sea el 1 o el 2, dependiendo del mes), tiene un consumo mínimo entre las las 00:00 y las 5:00, comenzando a partir de esta hora un aumento paulatino en el consumo hasta que se alcanza un máximo (alrededor de las 10:00) que más o menos se mantiene hasta las 20:00. El chiller de menor consumo tendrá una curva de potencia más constante, con menos variación de nivel, alcanzando el máximo a

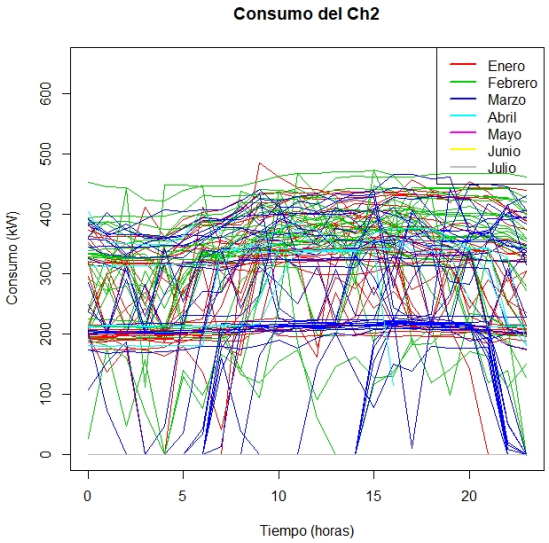
horas intermedias del día, que varían según el mes. Estas franjas horarias donde se incrementan los consumos, ocurren precisamente cuando el hotel realiza las actividades, sobre las 18:00-20:00. En las horas siguientes, hasta la 01:00, se mantiene cierto nivel elevado de consumo, ya que muchos huéspedes suben a las habitaciones, y a medida que pasan las horas de madrugada, al no haber tanta actividad, los consumos vuelven a reducirse y se mantienen más estables en valores bajos.



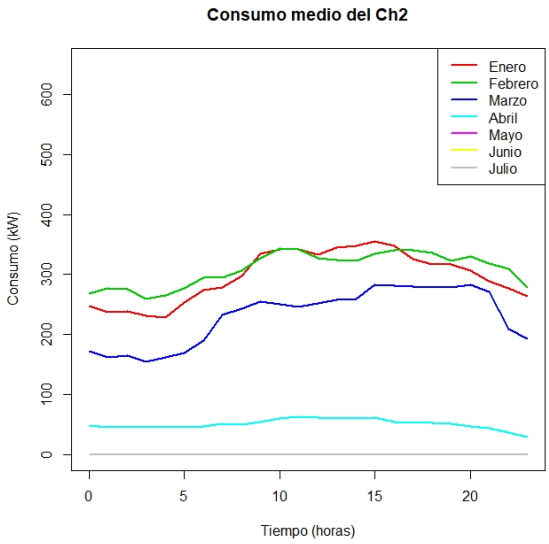
(a) Curvas de consumo energético diario correspondiente al chiller 1.



(b) Medias funcionales del consumo energético diario chiller 1.



(c) Curvas de consumo energético diario correspondiente al chiller 2.

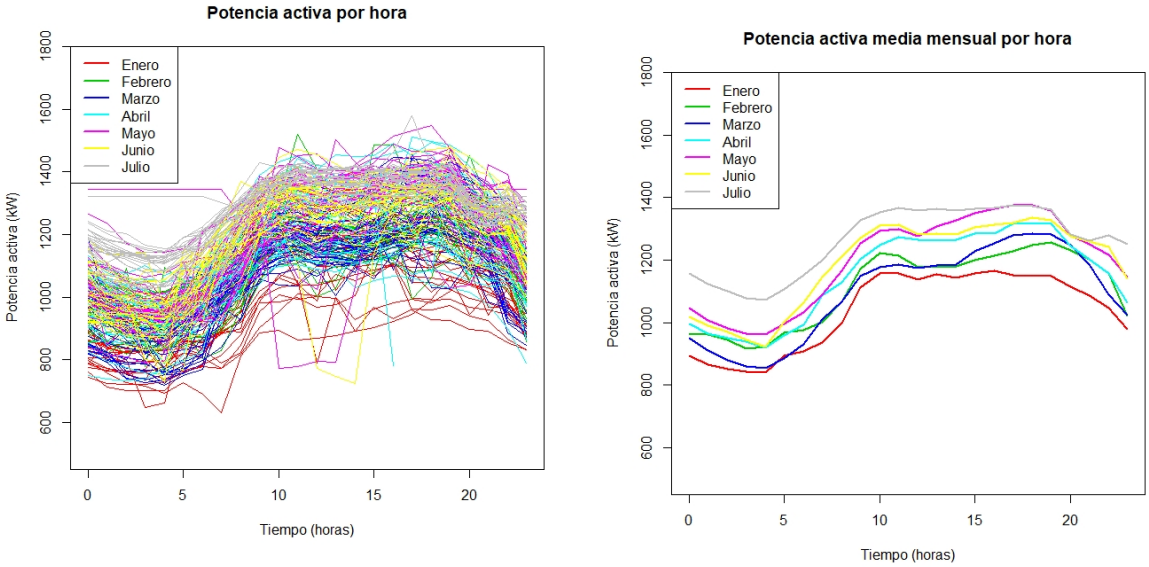


(d) Medias funcionales del consumo energético diario 2.

Figura 4.16: Consumo de los chillers de cada mes por horas.

Otra variable importante para la caracterización del funcionamiento del hotel es la potencia activa total consumida. La potencia activa de cada mes en función de las horas del día, la Figura 4.17 dos paneles, en el izquierdo se representan las curvas diarias de potencia diaria, mientras que en el derecho

se muestran las medias funcionales de la potencia activa para cada mes estudiado. Esta Figura nuestra siete curvas, una para cada una de los meses, y todas ellas con curvatura similar. Sin embargo, se observan diferencias en cuanto al nivel, de hecho, en el panel derecho de la Figura 5.14 se muestra que la potencia activa consumida tiende a aumentar de mes en mes, desde enero hasta julio, que es cuando se alcanza el máximo. Esto puede estar relacionado con la ocupación del hotel y con la necesidad creciente de refrigeración conforme se van acercando los meses de verano en el hemisferio norte. Todos los gráficos anteriores, se refieren a los consumos, no obstante las temperaturas, tanto



(a) Curvas diarias de potencia obtenidas a partir de observaciones horarias.

(b) Medias funcionales para cada mes.

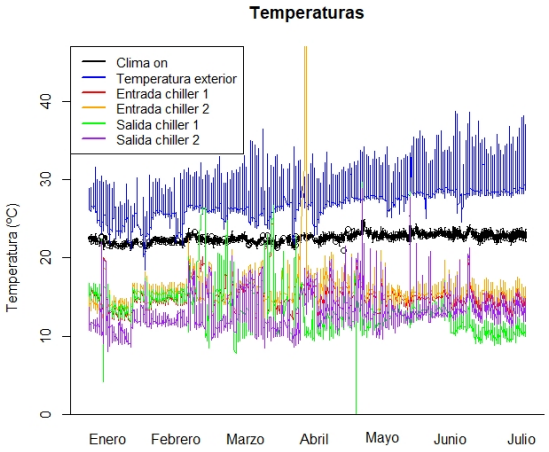
Figura 4.17: Potencia activa de cada mes por horas.

exteriores e interiores como aquellas relacionadas con el funcionamiento interno de los chillers, también tienen una gran importancia, ya sea por ser el principal indicador del comfort térmico de los ocupantes de un edificio ya sea por su relación directa con el consumo energético en las edificaciones. Es decir, es lógico pensar que si en un mes hace más calor, el consumo puede verse incrementado, ya que el sistema debería trabajar en mayor grado para mantener la misma temperatura en el interior del hotel, por ejemplo. En la Figura 4.18 se representan todas las temperaturas de las que se tienen registros: las temperaturas exteriores e interiores y las cuatro temperaturas asociadas a los dos chiller y los dos procesos de entrada y de salida.

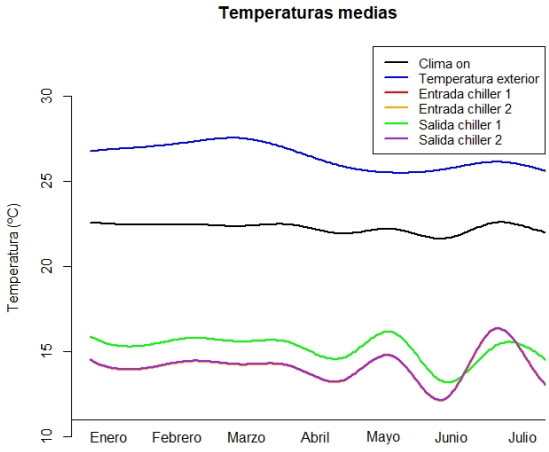
En primer lugar, en la Figura 4.18a se observa un posible dato atípico en la temperatura de entrada al chiller 1, el día 16-04-2019 a las 09:00, con un valor de 458.01ºC. Esta fecha pertenece a una de las anomalías registradas por el hotel.

En cuanto a la Figura 4.18b, se observa que la temperatura dentro del hotel es muy estable en media, a pesar de las fluctuaciones que muestra la temperatura exterior, que tiene un incremento en febrero y marzo, disminuyendo hacia los meses de mayo y junio, comportamiento predecible debido al clima del lugar. Las temperaturas asociadas a los chiller son aquellas que se muestran altamente inestables a partir de mediados de abril, donde sufren intervalos de incremento y decrecimiento bruscos de sus valores en cortos periodos de tiempo.

Por lo que se refiere a las temperaturas exteriores, se observa en la Figura 4.19 una gráfica similar a la obtenida en la Figura 4.17, donde se representaba la potencia activa media diaria funcional de cada mes en función de la hora del día. En esta última gráfica, el periodo en el que se registraban valores altos de dicha variable era 9:00-20:00 horas, mientras que en este caso, la temperatura exterior registra



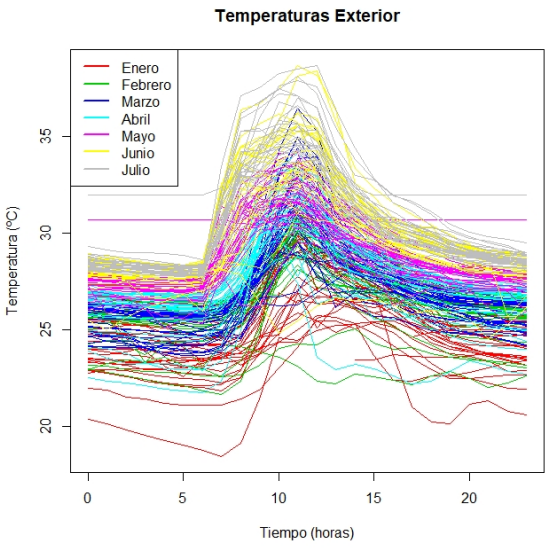
(a) Curvas diarias de temperaturas.



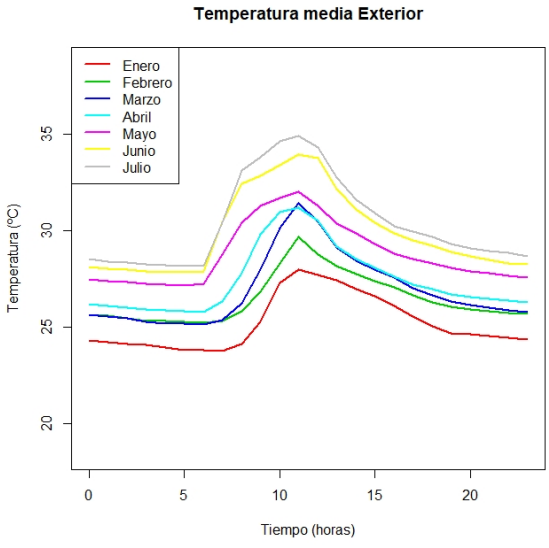
(b) Medias funcionales de las temperaturas.

Figura 4.18: Temperaturas en el periodo (enero 2019-julio 2019).

valores superiores en un periodo más concentrado, entorno a las 9:00-15:00 horas, aunque no vuelve a estabilizarse del todo hasta las 20:00 de la tarde. Esto puede denotar cierta relación lógica entre estas dos variables, implicando que a mayor temperatura exterior, la potencia activa del sistema también se ve incrementada. Se observa un aumento de nivel continuo desde enero a julio. Además, según avanzan



(a) Curvas de temperatura exterior diaria.



(b) Medias funcionales de la temperatura exterior diaria para cada mes.

Figura 4.19: Temperaturas en el periodo (enero 2019-julio 2019).

los meses hacia el verano, el comienzo del aumento de temperatura se produce cada vez a una hora

más temprana.

4.2.3. Análisis previo por meses

En cuanto al mes de enero, se observa como muestra la Figura 4.20, el chiller 1 tiene un consumo nulo en prácticamente todo el periodo, haciendo una excepción en los días 7-10 de enero, donde sí que tiene un pequeño consumo. Esto se puede deber a que el día 7 se dio un apagado total y en los días 8 y 9 se realizó el cambio del chiller 2 al chiller 1. A partir del día 9 de enero, el chiller 2 es el que sigue siendo el chiller que funciona a tiempo completo, por o que la mayoría del tiempo es este el que se encuentra activo, alcanzando el 99 % del funcionamiento total de los chiller en dicho mes.

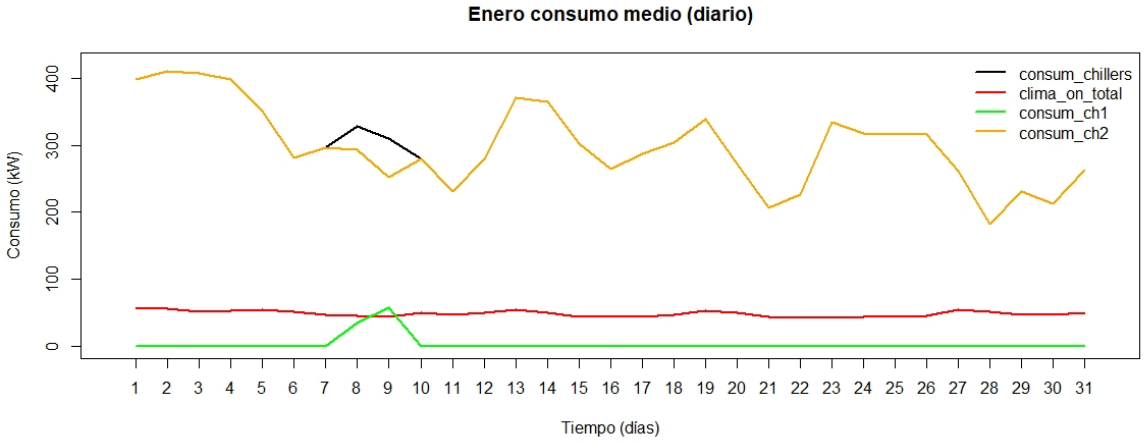


Figura 4.20: Medias de consumo de enero 2019.

En cuanto a las temperaturas en este mes, es obvio que de las temperaturas registradas, la más elevada sea la temperatura exterior, mientras que la temperatura medida en el interior del hotel se muestra muy estable. Como muestra la Figura 4.21, el día 9, que es cuando se produce el cambio del chiller 2 al chiller 1, se da un pequeño incremento de la temperatura en el interior del hotel, pero sin embargo es un incremento que sigue manteniendo una temperatura de confort.

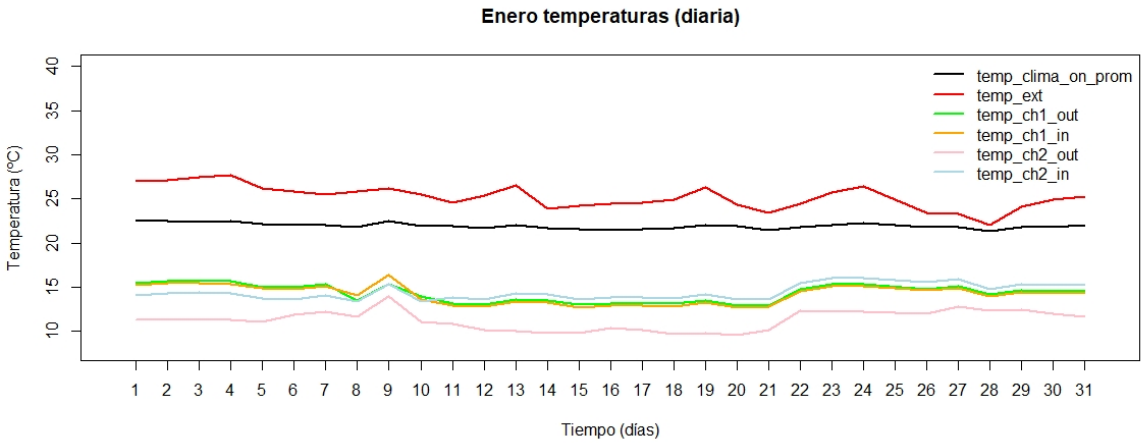


Figura 4.21: Medias de temperaturas de enero 2019.

Las temperaturas de entrada y salida de los chiller, muestran un comportamiento muy diferenciado a lo largo del mes de enero. Por un lado, en los primeros días del mes, las temperaturas asociadas al chiller 1 se muestran muy estables y similares. Sin embargo las temperaturas del chiller 2 se encuentran por debajo de estas, sobre todo la temperatura de salida. Entre los días 8 y 10 de enero, se produce un incremento en todas ellas, y posteriormente, se estabilizan hasta el día 21-22, donde sufren en general un incremento de 2ºC. En este segundo periodo del mes, la temperatura de salida del chiller 2 sigue siendo muy inferior al resto, y es justamente la temperatura de entrada a este la que es superior a las temperaturas asociadas al chiller 1, aunque similares.

Por lo que al funcionamiento de la torre de enfriamiento se refiere, la Figura 4.22 muestra que el ventilador 1 de la torre de enfriamiento se encuentra en funcionamiento la gran parte del mes, sufriendo disminuciones hasta dejar de funcionar en los periodos 5-9 y 28-31 de enero, donde deja de funcionar. En cuanto a los días 10,16 y 20, se reduce su funcionamiento en un 25 % aproximadamente. El ventilador 2 trabaja menos que el ventilador 1, sobre todo a partir del día 24. Cabe resaltar que

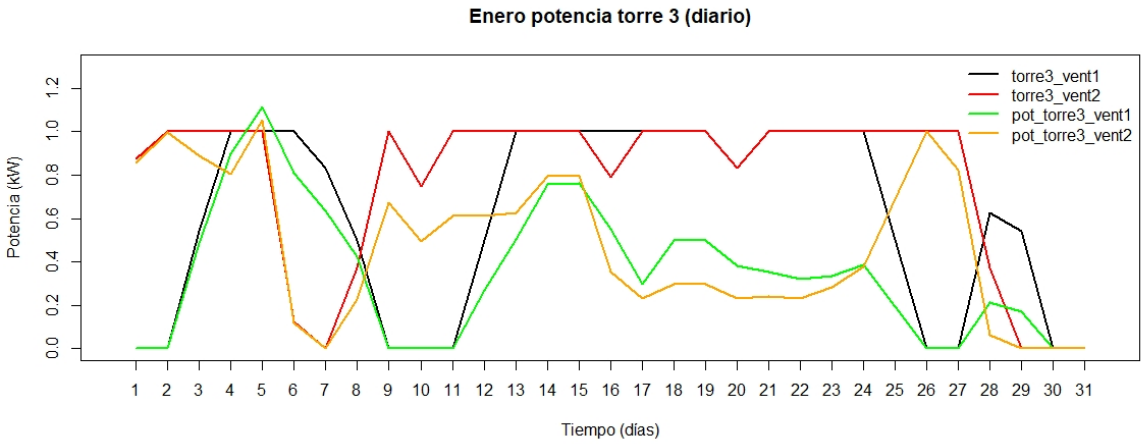


Figura 4.22: Medidas de la torre de enfriamiento de enero 2019.

existen momentos en los que ninguno de los dos ventiladores trabaja, por ejemplo, el día 8 ninguno de los dos ventiladores trabaja de 12:00 a 14:00, y como se puede observar en la Figura 4.23, cuando ambos ventiladores dejan de funcionar, la temperatura interior del hotel sufre un incremento.

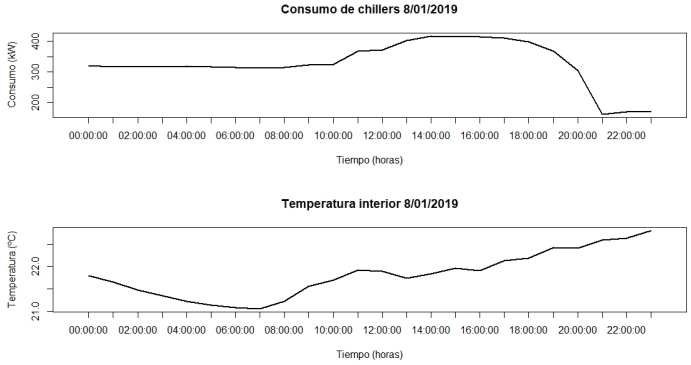


Figura 4.23: Consumo y temperaturas 8 de enero de 2019.

El día 5 la potencia asociada a ambos ventiladores sufre un pico, sobrepasando el valor de 1kW.

Esto puede deberse simplemente a que precisamente el día 04 de enero la temperatura exterior registra una de sus máximos, rondando los 28ºC, lo que puede llevar a que la potencia de los ventiladores se vea incrementada, debido al incremento del esfuerzo a la hora de disipar el calor no deseado del interior del hotel al exterior del mismo.

En cuanto a febrero, en comparación con el mes anterior, el consumo de los chiller se ve incrementado, al igual que ocurre con la ocupación del hotel (del 68 % al 79.77 %) y el porcentaje de tiempo de funcionamiento del sistema de enfriamiento. No obstante, el consumo referente a la potencia activa, es decir, el consumo total se reduce de los 771380.04 kW de enero a los 745307.75 kW consumidos en febrero. Estos datos son referentes a los consumos totales, donde la suma de estos se ve afectado por el número de días del mes, lógicamente. Por ello resulta interesante contemplar los incrementos en porcentajes sobre los valores medios. Esto quiere decir que de enero a febrero ocurre un incremento del 7 % en cuanto al consumo total y un incremento del 17,65 % en cuanto al consumo de los chiller.

En cuanto a los chiller, es el chiller 2 el que presenta un porcentaje de tiempo encendido mayor, entorno al 87.31 %. Esto puede verse reflejado en la 4.24, donde se observa cómo el chiller 2 es el que provoca la gran parte del consumo total, mientras que el chiller 1 solamente tiene un consumo en los días 17-22 y a finales del mes, a partir del día 25.

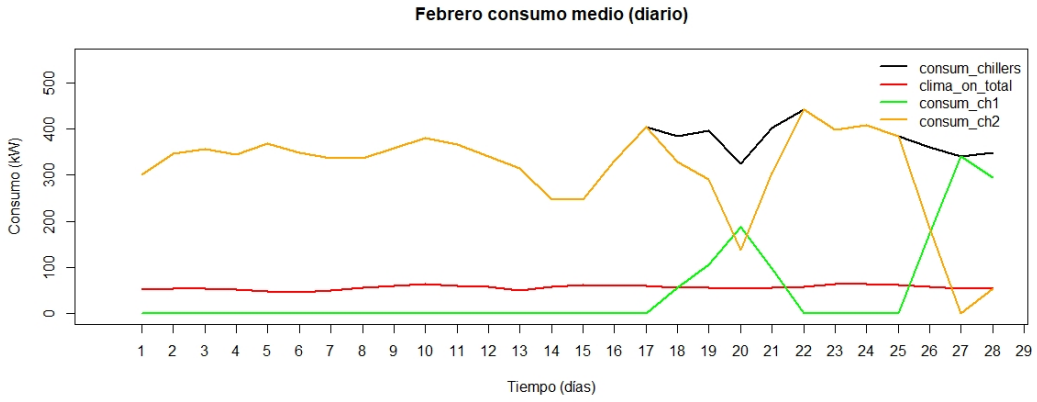


Figura 4.24: Medias de consumo de febrero 2019.

En líneas generales, el consumo total de los chiller es estable, viándose reducido en los días 13-16 de febrero.

En cuanto al comportamiento de las temperaturas, de nuevo ocurre que la temperatura dentro del hotel se mantiene prácticamente constante e entorno a 22.35ºC. Obviamente, la temperatura en el exterior del hotel es superior, con media de 26.48ºC, siendo algo inferior en los días 11-15 de febrero.

En referencia a las temperaturas asociadas a los chiller, son muy estables durante la primera quincena del mes, hasta el día 18 como muestra la 4.25, ya que a partir de dicho día, y sobre todo la temperatura del proceso de salida del chiller, sufre un incremento considerable con un pico máximo que ronda los 25ºC. En todo el periodo, la temperatura de salida del segundo chiller es inferior al resto, y de hecho en los últimos días del mes ésta disminuye, a diferencia del resto de temperaturas.

Por lo que a la torre de enfriamiento se refiere, el ventilador 2 de la torre no funciona, mientras que el ventilador 1 es el que trabaja todos los días, excepto el día 1 y el 16, donde ninguno de los ventiladores se encuentra en funcionamiento. La potencia del ventilador 1, se observa en la 4.26 se observa una inestabilidad considerable a partir del 16 de febrero, donde llega a valores de 3. Precisamente en ese momento es donde se da la desconexión de la torre de enfriamiento, dicha desconexión ocurre desde el 15 a las 7:40 hasta el 17 a las 9:25. Esta torre 3 es la única torre de enfriamiento del sistema, lo que provoca que efectivamente, se den alteraciones de las temperaturas asociadas a los chiller en este periodo.

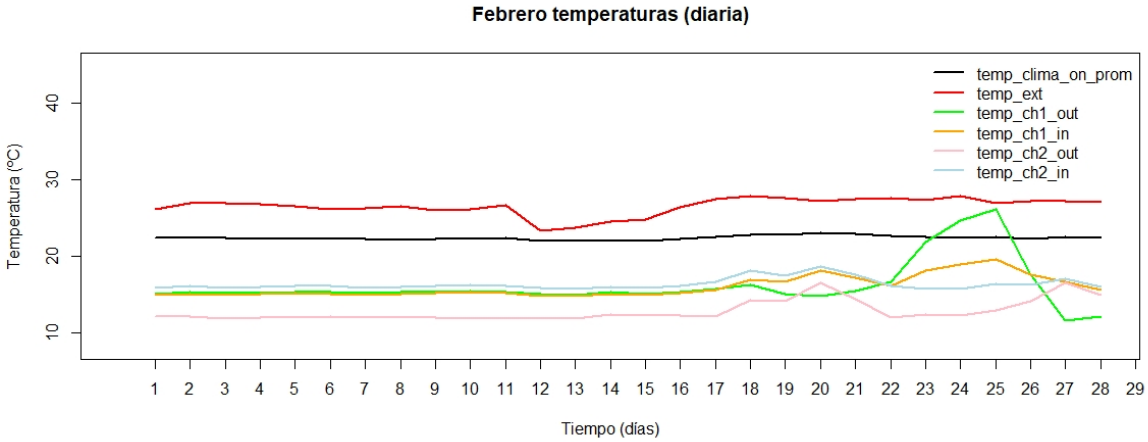


Figura 4.25: Medias de temperaturas de febrero 2019.

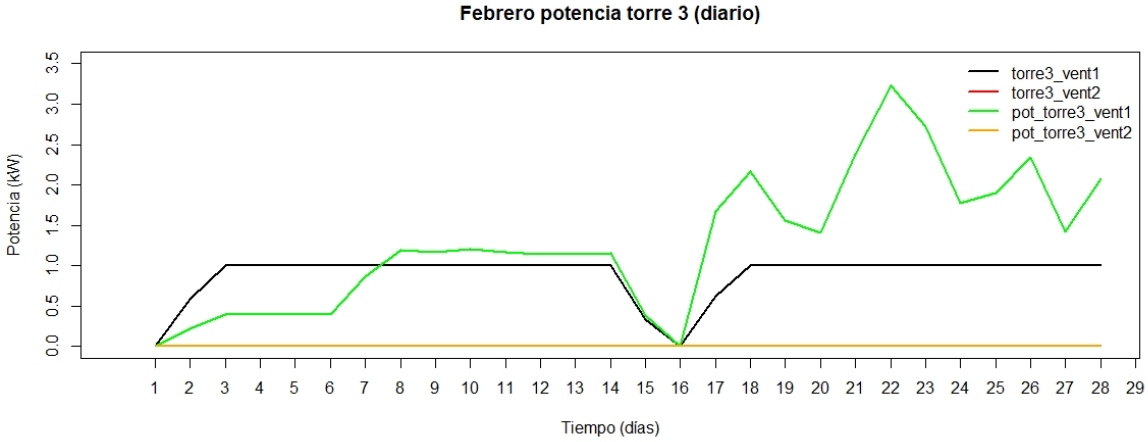


Figura 4.26: Medidas de la torre de enfriamiento de febrero 2019.

Además, desde el 18 a las 19:00 hasta el 21 a las 12:00, se da la puesta en funcionamiento de ambos chiller, lo que puede ser la causa principal de los cambios bruscos en las temperaturas de los chiller, así como la temperatura del hotel. Esta última, la temperatura del interior del hotel es incrementada ligeramente, pero sin dejar de ser una temperatura de confort.

En general, los parámetros relativos a temperaturas y ocupación de marzo son bastante similares a los del mes anterior. Sin embargo, el consumo total aumenta a 814551.97 kW, así como el consumo de los chiller, que pasan de tener una media de 237799.92 kW en febrero a 263554.33 kW en marzo. La temperatura exterior ronda los 27ºC de media, y la temperatura interior del hotel también se mantiene constante respecto a los meses anteriores, situándose de media entorno a 22ºC.

El porcentaje de funcionamiento del chiller 1 a lo largo del mes es del 34.98 %, mientras que el chiller 2 ocupa el 65.02 %. El consumo de estos es algo superior la primera mitad del mes, siendo más constante e inferior en la segunda quincena.

No obstante, la distribución del consumo de los chiller varía considerablemente en este tercer mes del año, como se puede apreciar en la Figura 4.27. Hasta el día 12, es el chiller 2 el que tiene un consumo mayoritario, pero dicho día se produce una inversión, tomando las riendas el chiller 1 hasta

prácticamente el día 25. A finales del mes, de nuevo el chiller 2 es el que provoca en mayor parte el consumo total de los chiller.

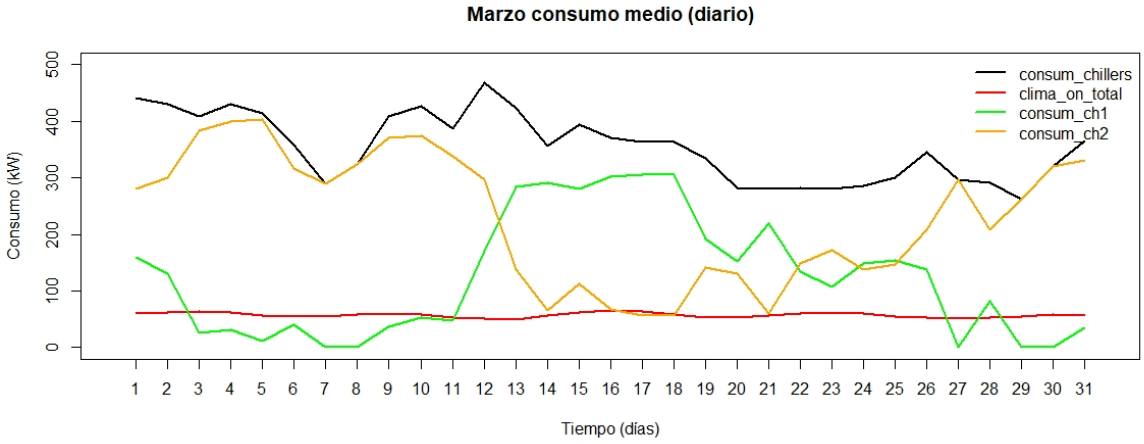


Figura 4.27: Medias de consumo de marzo 2019.

Las temperaturas tanto exterior como interior del hotel son estables y sin grandes cambios. Las temperaturas asociadas a los chiller, en general son estables también, exceptuando los días 26-31 del mes, donde todas ellas aumentan, alcanzando incluso los 23ºC. En particular, la temperatura del proceso de salida del chiller 1 entre los días 2 y 13 de marzo sufre un gran incremento de 10ºC, alcanzando los 20ºC, como muestra la Figura 4.28.

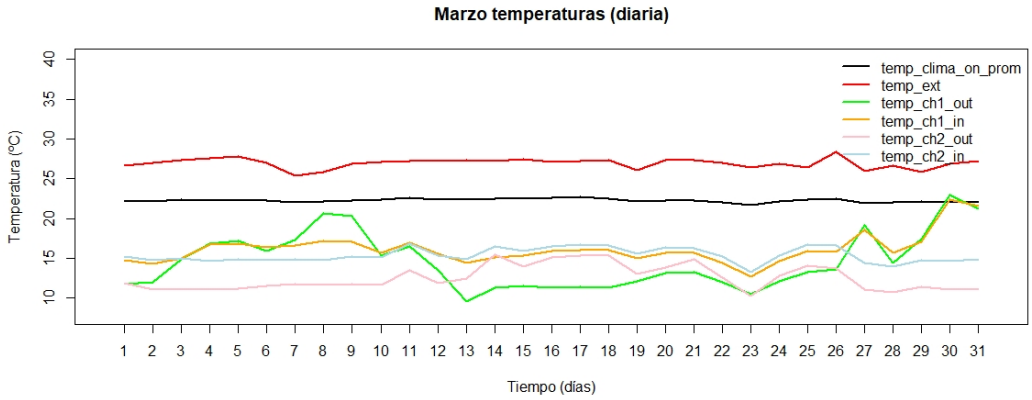


Figura 4.28: Medias de temperaturas de marzo 2019.

Los 6 primeros días sólo funciona el ventilador 1 de la torre de enfriamiento, y por el contrario, a finales es el segundo ventilador el único en funcionamiento. En cuanto a las potencias, los valores máximos son alcanzados por el ventilador 2, llegando a valores de 3.5 en el punto de máxima potencia. Como se ve en la Figura 4.29, el ventilador 1 solo tiene un momento puntual en los que se registran valores que superan el 2, mientras que en el resto del mes muestra valores estables.

Resulta muy interesante observar qué comportamiento tiene tanto el consumo como las temperaturas del hotel cuando solamente funciona uno de los dos chiller. Para ello, se muestra la Figura 4.30. En ella se toman como ejemplo los días 3 de marzo y 8 de marzo, que son días en los que solamente funciona el chiller 1 y el chiller 2 respectivamente. En este mes no existe ningún registro donde ambos chiller

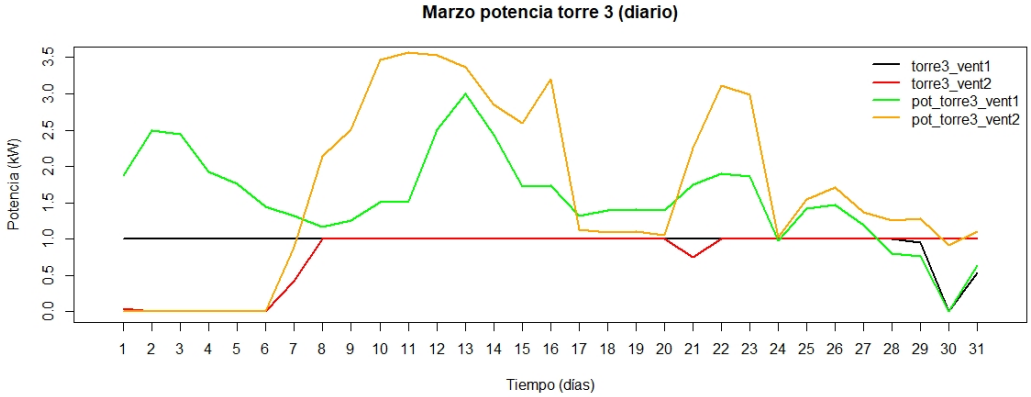


Figura 4.29: Medidas de la torre de enfriamiento de marzo 2019.

se encuentren apagados. Como se puede observar, tanto si trabaja el chiller 1 como si solo trabaja el chiller 2, la temperatura interior del hotel se ve incrementada, así como el consumo. La temperatura dentro del hotel sigue siendo una buena temperatura, pero el consumo se eleva considerablemente respecto de la media, que este mes se encuentra en 1094.828 kW.

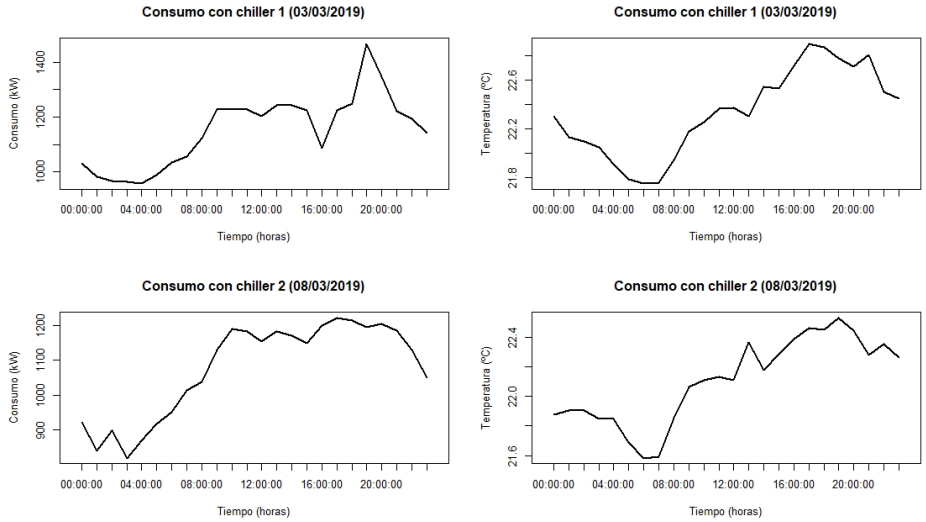


Figura 4.30: Consumo y temperaturas 8 de marzo de 2019.

En abril el consumo total sigue aumentando siendo en este mes, sumando un total de 821562.62 kW, o lo que es lo mismo, de marzo a abril se incrementa en un 4.22 %. Sin embargo, con el consumo de los chiller ocurre algo muy llamativo, ta que del mes de marzo a abril ocurre una disminución del consumo de casi un 35 %. La temperatura exterior sigue incrementándose, superando en media los 27.5ºC. En cuanto a la ocupación en este caso, se reduce en dos puntos, obteniándose una ocupación del 78.4 %.

El consumo del chiller 1 es el 79.68 % del consumo de los chiller, como era de esperar en base a la 4.31. De hecho, se observa cómo el chiller 2 solamente tiene consumo los primeros 7 días de abril. En estos 7 primeros días el consumo total es superior al resto del mes. Queda reflejado, por tanto, cómo el tener trabajando los dos chiller de manera simultánea provoca consumos más elevados que cuando

trabaja uno solo. Sin embargo, el chiller 1 tiene un consumo un tanto inestable el resto del periodo con valores bajos en los días 7-14, y con valores altos en los días 15-21 y 27-28.

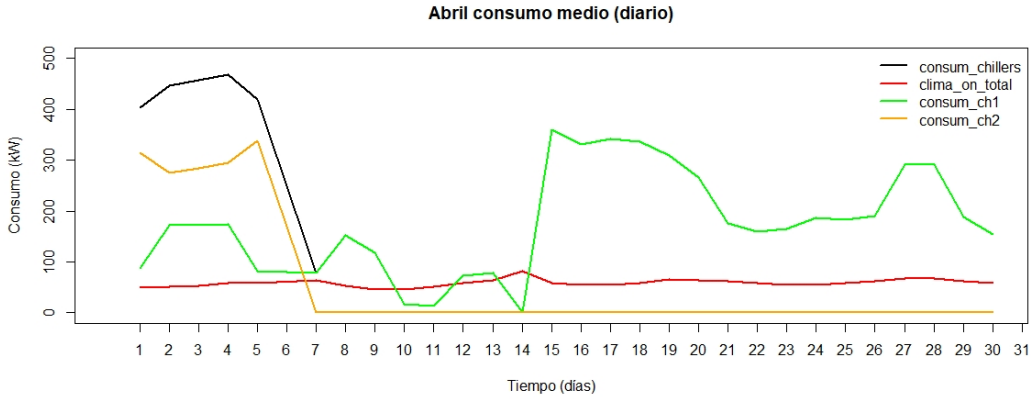


Figura 4.31: Medias de consumo de abril 2019.

Por lo que a las temperaturas respecta, la temperatura exterior ronda los 27.5ºC, pero en el hotel se mantiene una temperatura media confortable, entorno a los 22.5ºC de media. Las temperaturas en los chiller se muestran estables, excepto la asociada al proceso de entrada del chiller 2, que sufre un pico en los días 13-17, como muestra la Figura 4.32 alcanzando valores extremos en los valores superiores a los 60ºC. Este suceso tiene que ver con la existencia de un fallo en la emisión de los datos desde el 14 de abril a las 00:00 hasta las 08:00 horas. Además, al iniciarse de nuevo la emisión de los datos del sistema de control, la forma de acumular el dato de la temperatura del proceso de entrada del chiller 2 no es correcta, por lo que se ve ese incremento exponencial hasta que ocurre la reprogramación de almacenamiento de datos, hecho que ocurre el 16 de abril a las 10:00 horas. A partir de esta hora, la recogida del valor de la temperatura d entrada al chiller 2 es correcta. Estos sucesos suelen darse con cierta facilidad cuando las compañías de control no son las mismas que gestionan y almacenan datos.

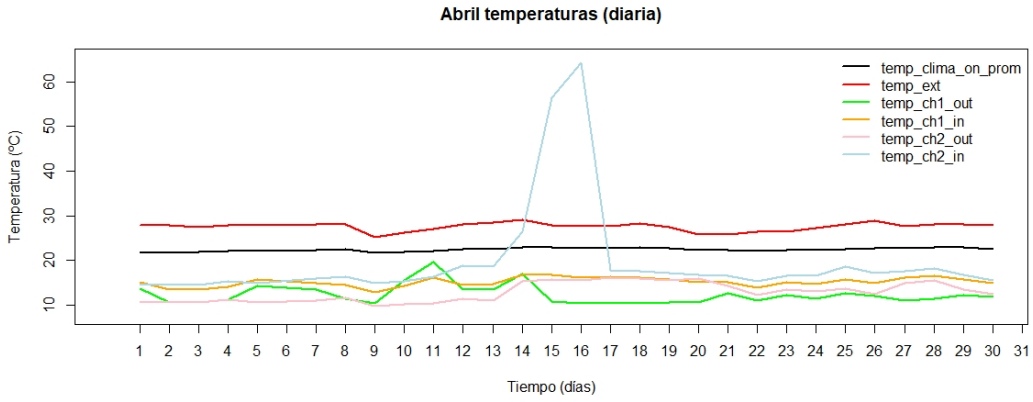


Figura 4.32: Medias de temperaturas de abril 2019.

El ventilador 1 y el ventilador 2 de la torre de enfriamiento se comportan de manera muy similar en todo el mes, y además, se puede observar en la Figura 4.33 que no hay ningún momento en el que los ventiladores se encuentren apagados. En los días 9-11, los ventiladores trabajan a potencias bajas entorno a 0.75 kW. Por el contrario, en los días 18-20 y 27-30 los ventiladores trabajan a potencias muy elevadas, llegando a cuadruplicar el valor anterior, es decir, superando los 3 kW.

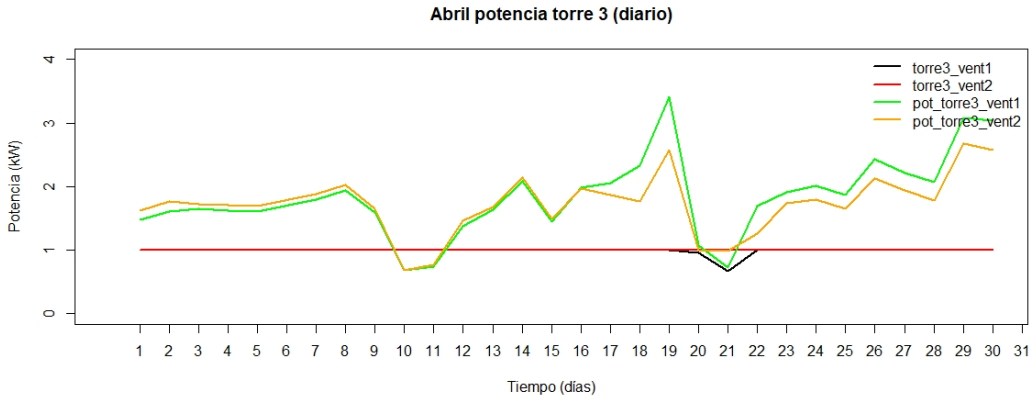


Figura 4.33: Medidas de la torre de enfriamiento de abril 2019.

Nótese que el gran incremento de la potencia de los ventiladores coincide con el incremento en la temperatura del proceso de entrada del chiller 2.

El consumo medio de mayo se incrementa en un 4.87 % respecto del mes anterior, pero el consumo de los chiller se reduce en un 18.31 %. El consumo de los chiller proviene íntegramente del consumo realizado por el chiller 1 que es el único que se encuentra en funcionamiento. Por ello, en la Figura 4.34 se muestra que el consumo total de los chiller es igual al consumo del chiller 1. Estos valores son bastante estables hasta llegado el dia 18 de mayo, donde comienza a incrementarse el valor hasta prácticamente el 18 de mayo. Respecto a la ocupación, esta sigue descendiendo, siendo del 75.1 % en este mes.

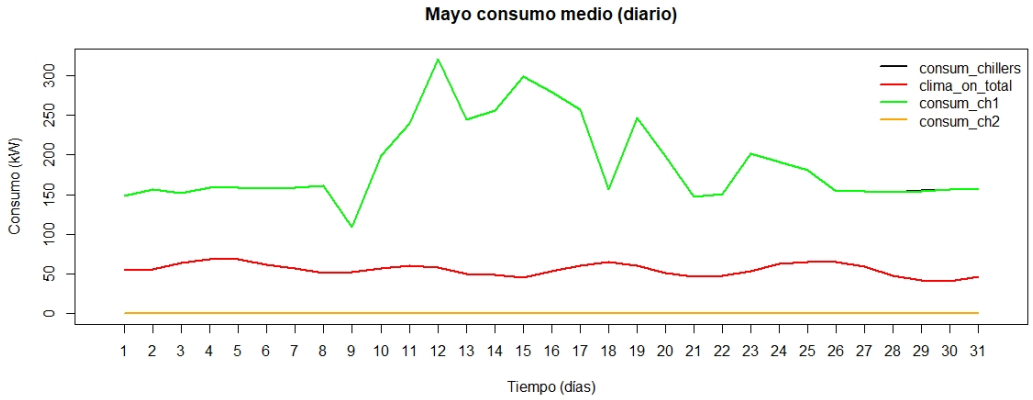


Figura 4.34: Medias de consumo de mayo 2019.

Por lo que a las temperaturas de mayo respecta, tanto la temperatura exterior como la del interior del hotel se elevan en este mes, alcanzando unas temperaturas exteriores medias de 28.8ºC y la temperatura interior media es de 23.25ºC. Las temperaturas asociadas a los chiller, muestran grandes variaciones en los días 11-15 de mayo, y en líneas generales las temperaturas asociadas al proceso de entrada de ambos chiller son superiores a las del proceso de salida. En base al gráfico anterior, el chiller 2 no se encuentra en funcionamiento en mayo, pero sin embargo, las temperaturas asociadas al chiller 2 mostradas en la Figura 4.35, no muestran los mismo.

Cono ocurría ya en abril, ambos ventiladores de la torre de refrigeración se encuentran encendidos en todo momento. Mayoritariamente, la potencia asociada al ventilador 1 es superior a la potencia del

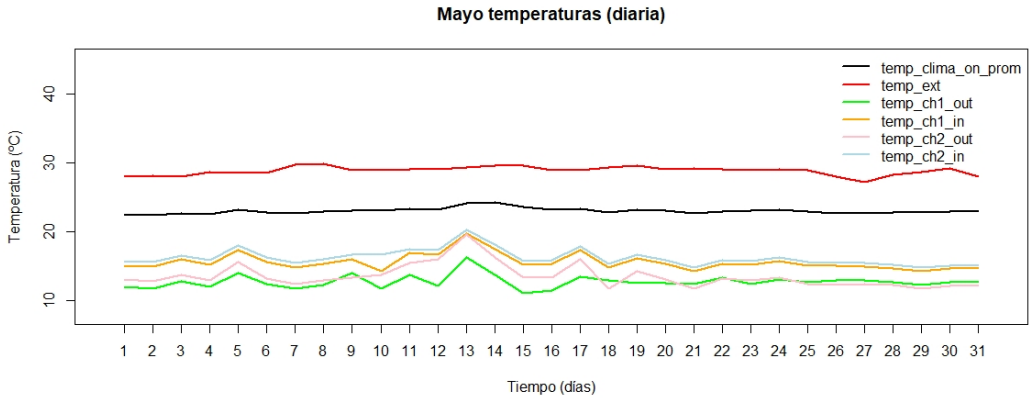


Figura 4.35: Medias de temperaturas de mayo 2019.

segundo ventilador. En los días 4-8 de mayo se registran los valores de potencia máxima, ver Figura 4.36, alcanzando los 3 kW para el ventilador 2 y os 2.5 kW para el ventilador 1.

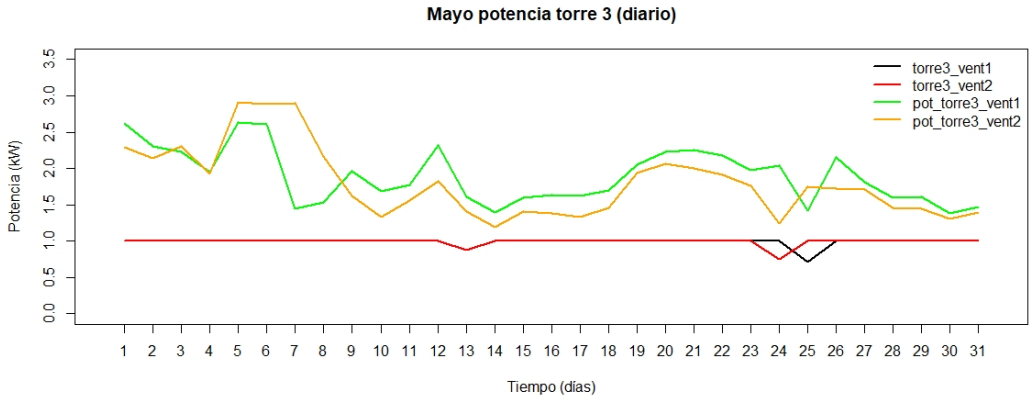


Figura 4.36: Medidas de la torre de enfriamiento de mayo 2019.

El día 13 de mayo ocurre un apagado desde las 10:00 hasta las 13:00, por lo que resulta intrigante observar qué ocurre con el consumo y las temperaturas tras dicho apagado. Como muestra la Figura 4.37, el apagado provoca grandes alteraciones en las temperaturas, y de hecho, la temperatura en el interior del hotel se incrementa hasta 3ºC pasando de los 23ºC hasta los 26ºC.

Este mes, a diferencia de lo que venía ocurriendo, el consumo de los chiller aumenta en un 1.54 % respecto el mes anterior. Como era de esperar, la temperatura exterior sigue en aumento, cerrando el mes con una media de casi 30ºC. La ocupación, por su lado, se encuentra en disminución siendo en junio del 64.39 %.

Como se observa en la Figura 4.38, el consumo total viene dado por el consumo del chiller 1, que es el que trabaja el 100 % de días del mes. El gráfico muestra tres puntos de incremento, dos muy puntuales en los días 6-9 y 14-16, y uno más prolongado, que se da desde el día 23 hasta final de mes.

En la temperatura interior del hotel no se dan grandes variaciones como muestra la Figura 4.39, y la media es de algo más de 23ºC. Las temperaturas asociadas a los chiller son estables en líneas generales, la temperatura del proceso de salida del chiller 1 es algo menos estable, y a partir del día 23 sufre un descenso. Este descenso de la temperatura del proceso de salida del chiller 1 puede tener relación con el aumento del consumo de potencia de este chiller en dicho periodo.

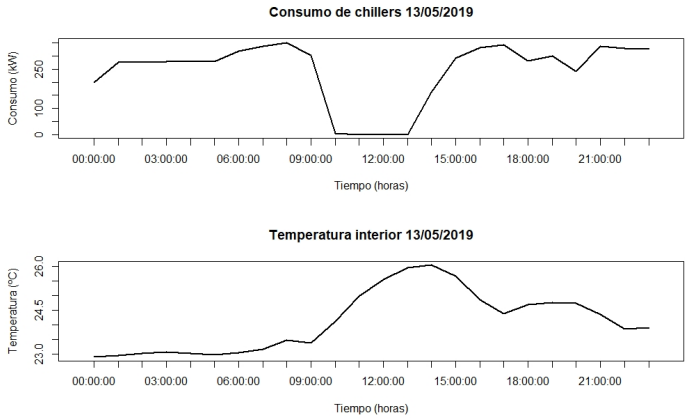


Figura 4.37: Consumo y temperaturas 8 de mayo de 2019.

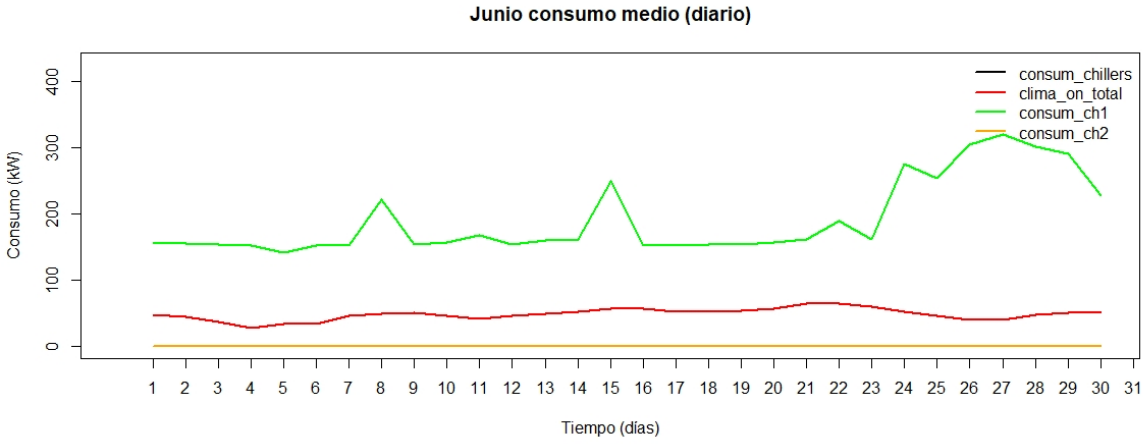


Figura 4.38: Medias de consumo de junio 2019.

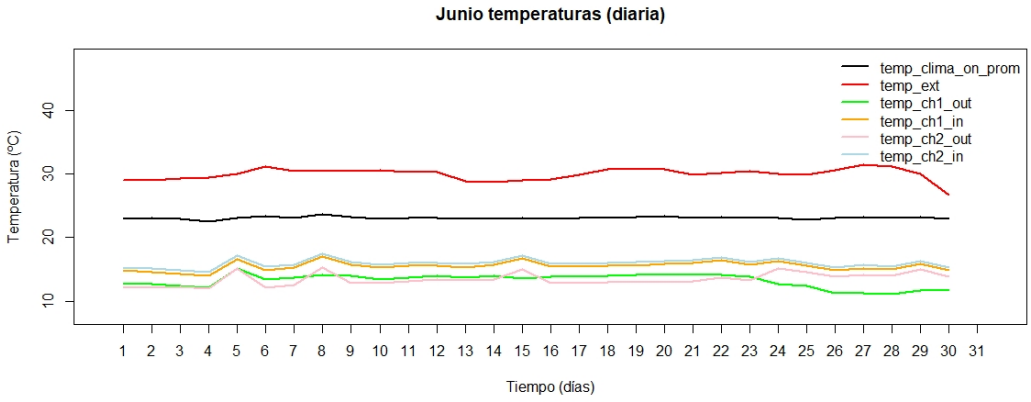


Figura 4.39: Medias de temperaturas de junio 2019.

Sin embargo, todas las temperaturas sufren un ligero incremento en los días 5-6 de junio,que es precisamente cuando se da el apagado del sistema, que ocurre exactamente el día 5 de junio de 10:00 a 13:00. Antes de pasar al análisis de las potencias relativas a la torre de enfriamiento, vuelve a llamar la atención lo que ya ocurría el mes de mayo, y es que el chiller 2 se supone que no está en funcionamiento, pero sin embargo, sus temperaturas de entrada y salida parecen corresponder a un chiller en funcionamiento.

En cuanto a la torre de enfriamiento, como se ve en la Figura 4.40, mantiene sus dos ventiladores activos en todo el mes. Ambas potencias, las asociadas al ventilador 1 y el ventilador 2, sufren una caída en los días 13-14. En la primera parte del mes, ambos ventiladores muestran potencias similares pero tras el pico sufrido en los días 13-14, es el ventilador 1 el qe toma valores superiores. El día 5-6 la potencia parece ser afectada también por el apagado general, reduciéndose en dichas fechas notablemente.

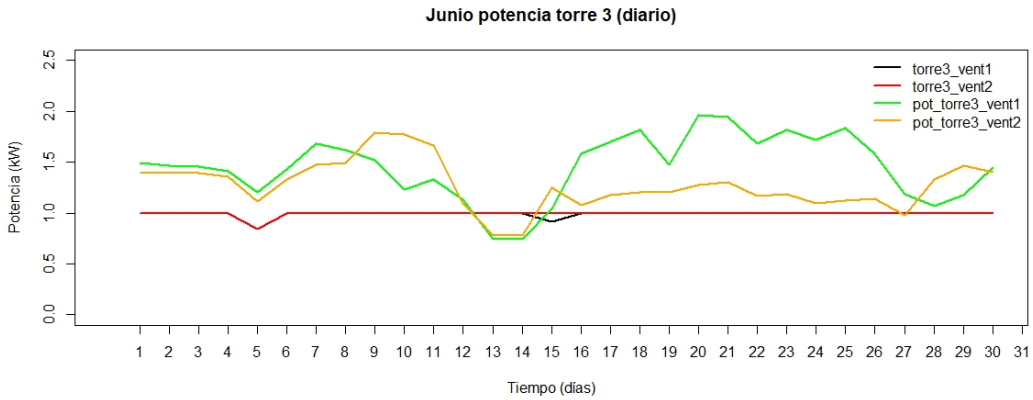


Figura 4.40: Medidas de la torre de enfriamiento de junio 2019.

Para observar qué es lo que ocurre exactamente el día 5 cuando ocurre el apagado general desde las 10:00 hasta las 13:00, se grafica el consumo y la temperatura interior del hotel por hora en ese día. Como muestra la Figura 4.41, cuando se da el apagado, la temperatura aumenta considerablemente, y a su vez, el consumo se reduce inicialmente para posteriormente sufrir un incremento.

Nótese que el gráfico referente a la temperatura interior del hotel a partir de las 18:00 horas se encuentra incompleto, ya que en este periodo de tiempo existen valores faltantes, es decir, no se procedió a la lectura de la temperatura interior, no obstante puede apreciarse la tendencia que ocurre en el periodo de 10:00 hasta las 13:00 horas, que es donde ocurre el apagado general, a partir de las 19:00 todo apunta a que es posible considerar el restablecimiento del sistema.

El mes de julio deja un dato bajo de ocupación, que no alcanza el 50 %. El consumo total sufre un incremento considerable del 6.10 %, pero lo más llamativo es el incremento del consumo de los chiller. En este caso también es el chiller quien más tiempo está en funcionamiento siendo casi del 100 %.

Como muestra la 4.42, los valores de consumo son muy estables en todo el periodo de julio, pero son valores elevados. Estos valores elevados se alcanzan desde el día 2 del mes, ya que los últimos días de junio no mostraban valores de esa magnitud.

La temperatura exterior en este mes es de 30.4ºC, y el hotel mantiene una media de 22.94ºC. A rasgos generales, las temperaturas son más elevadas, tanto la del exterior como la obtenida en el interior del hotel. Sin embargo, llama la atención que la temperatura del proceso de salida asociado al chiller 1 es inferior al resto, lo que tendría sentido, ya que es el chiller que trabaja prácticamente el 100 % de las veces.

En cuanto a la torre de enfriamiento, en prácticamente todo julio mantiene sus dos ventiladores en funcionamiento, como la Figura 4.44 muestra, aunque el 27 de julio el ventilador deja de funcionar hasta finales de mes. La mayoría del tiempo, es el ventilador 1 el que mayor potencia mantiene. Se observa

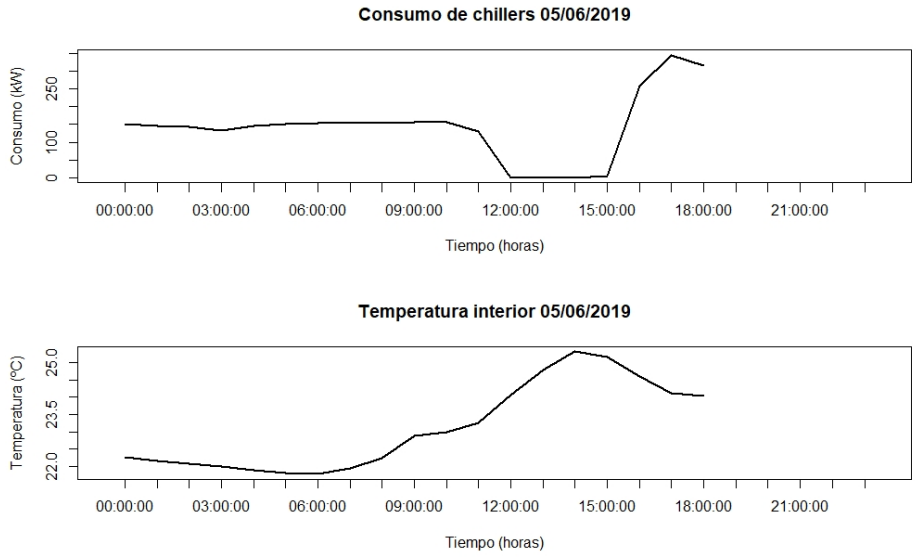


Figura 4.41: Consumo y temperaturas 5 de junio de 2019.

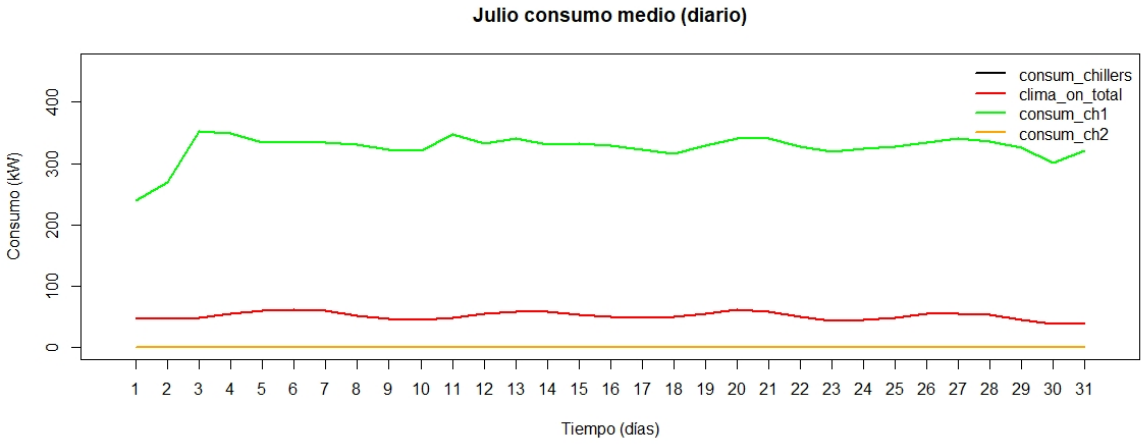


Figura 4.42: Medias de consumo de julio 2019.

que cuando el ventilador 1 deja de funcionar, el ventilador 2 sufre un incremento de su potencia. El hecho de que el ventilador 1 deje de funcionar, puede ser un factor a tener en cuenta en el incremento del consumo del chiller 1 en esos últimos días de julio.

En líneas generales, se detecta un suceso extraño respecto al chiller 2. Y es que a principios del mes de abril, el chiller 2 deja de funcionar si se observa el consumo de éste. A partir de abril, no vuelve a tener consumo, pero sin embargo, las temperaturas del proceso de entrada y salida asociadas a este chiller se muestran de manera que parecen referirse a un chiller en funcionamiento. Esto indica que en abril el chiller 2 deja de emitir correctamente su consumo, posteriormente se detecta que es la placa de control interior la que no indica correctamente el estado ni el consumo. A partir del día 07 de abril a las 01:00 horas el consumo del chiller 2 se registra como 0.003464 kW, por lo que es lógico que el consumo de marzo a abril se vea reducido en un porcentaje de casi un 35 %, por que el chiller 2 no está registrando el consumo de potencia que realmente está teniendo.

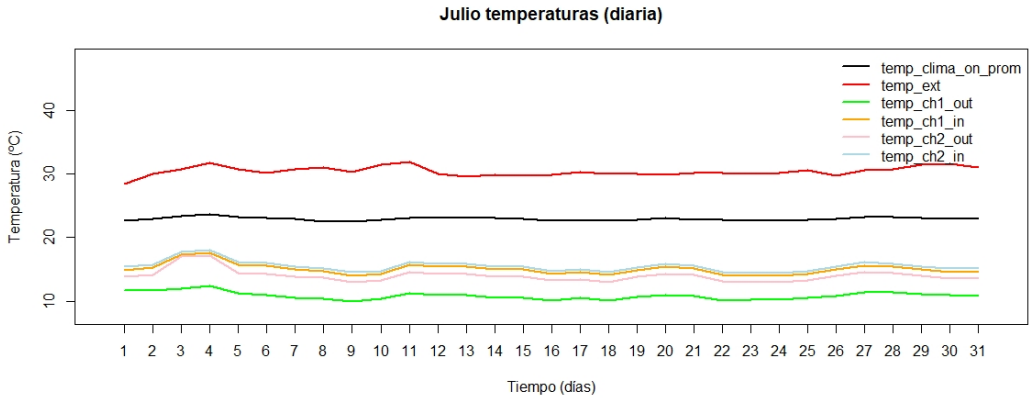


Figura 4.43: Medias de temperaturas de julio 2019.

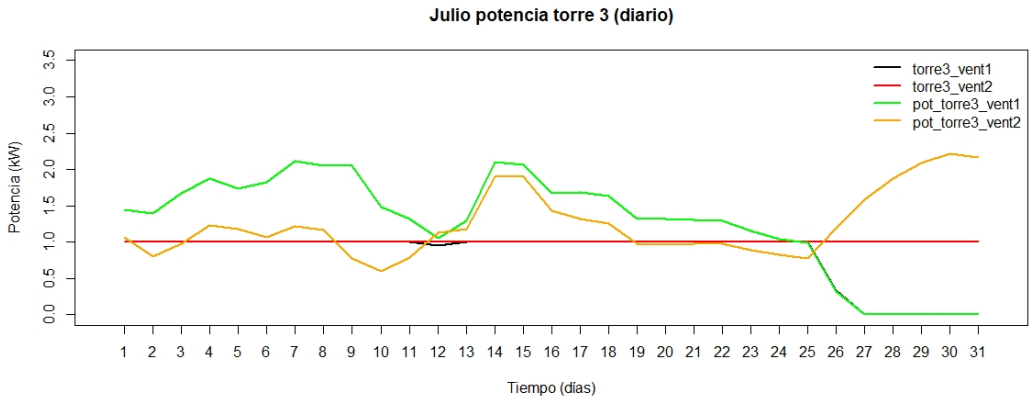


Figura 4.44: Medidas de la torre de enfriamiento de julio 2019.

Estudio de datos atípicos

Cabe pensar en la posibilidad de datos atípicos o también conocidos como outliers. Estos outliers son datos que simplemente son muy distintos del resto de los valores, lo que implica directamente que son valores que de manera excepcional se encuentran alejados del centro. En la mayoría de los casos estos valores extremos tienen influencia en la media, pero no en la mediana, o la moda. Por lo tanto, los outliers son importantes en su efecto en la media.

Realizando un pequeño estudio de detección de outliers, empleando como medida la profundidad de los datos. La profundidad es una herramienta estadística que puede emplearse, entre otras cosas, para obtener medidas de localización. Existen multitud de medidas de profundidad, como la profundidad de **Fraiman y Muniz (2001)**, la profundidad modal **Cuevas et al. (2007)** o la medida de profundidad de proyecciones aleatorias **Cuesta et al. (2007)**. Para este caso, se emplea la profundidad modal de **Cuesta et al. (2007)**:

$$MD(x_i) = \sum_{j=1}^{N} K\left(\frac{d(x_i, x_j)}{h}\right) \tag{4.1}$$

Esta medida de profundidad, trata de medir la profundidad empleando una función tipo kernel asimétrico, donde h es el parámetro de la ventana.

A la hora de la detección de outliers, se emplea un recorte del 1 %, donde se quiere obtener dicho

porcentaje de outliers. Aplicando esta técnica ea los datos en cuestión, el procedimiento detecta 49 datos atípicos. Estos datos se distribuyen según los meses como muestra la Figura que se muestra a continuación, en la Figura 4.45. Como se observa, de los 49 outliers detectados, 19 pertenecen a

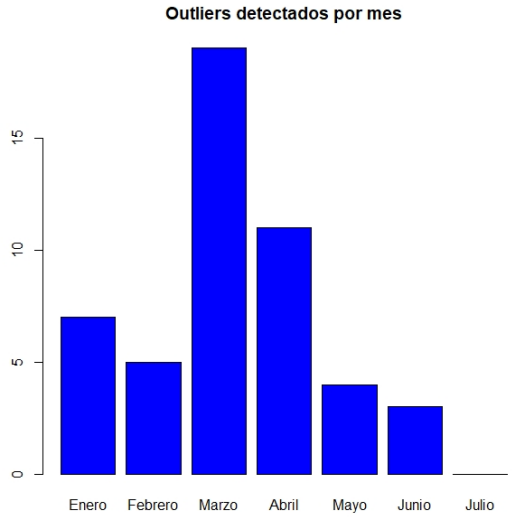


Figura 4.45: Distribución según los meses de os outliers detectados (trim=1 %).

mediciones realizadas en marzo y 11 a abril, mientras que julio no parece contener ningún outlier. Puede resultar interesante observar la distribución de estos outliers en las variables de estudio, como muestra la siguiente Figura 4.46. Como se observa, los outliers detectados pertenecen principalmente a valores extremos (excesivamente grandes o excesivamente pequeños) de las variables main_pot_activa y consum_chillers, que son precisamente dos de las variables más importantes en el control de calidad de este sistema. Sin embargo, al mismo tiempo son precisamente estos valores atípicos los que pueden tener relevancia en el control multivariante que se quiere llevar a cabo. El análisis de outliers debe ser complementado por los gráficos de control, ya que los gráficos de control, aparte de detectar anomalías, permiten estimar posición y variabilidad de las variables que definen al proceso, además de permitir detectar patrones cuando las observaciones son obtenidas secuencialmente en el tiempo. Los outliers detectados en el análisis no son valores extremadamente alejados de las nubes de puntos formada por las observaciones, por lo que no se procede a la eliminación de los mismos, con el simple objetivo de ver más adelante en los gráficos de control multivariante si realmente son observaciones fuera de control o no.

Análisis preliminar de agrupación por meses

Se han aplicado técnicas de clasificación no supervisada o cluster para observar si existen similitudes o diferencias en en comportamiento de las instalaciones del hotel dependiendo del intervalo temporal, en este caso medido en meses.

Para ello, se realiza un análisis a grandes rasgos de la formación de grupos empleando un método clúster jerárquico aglomerativo (Anexo B) basado en distancias euclídeas a las media y a las medianas mensuales para observar cómo se comportan las variables la gran parte de cada uno de los meses. Pudiendo verificar de esta manera si realmente se observan meses con comportamientos similares y si existen meses con comportamientos distintos del resto. A continuación, en la Figura 4.47 se muestra el dendograma derivado de dicha formación de grupos.

Si se observan las alturas donde el dendograma aplica la primera ramificación, se observa que el salto

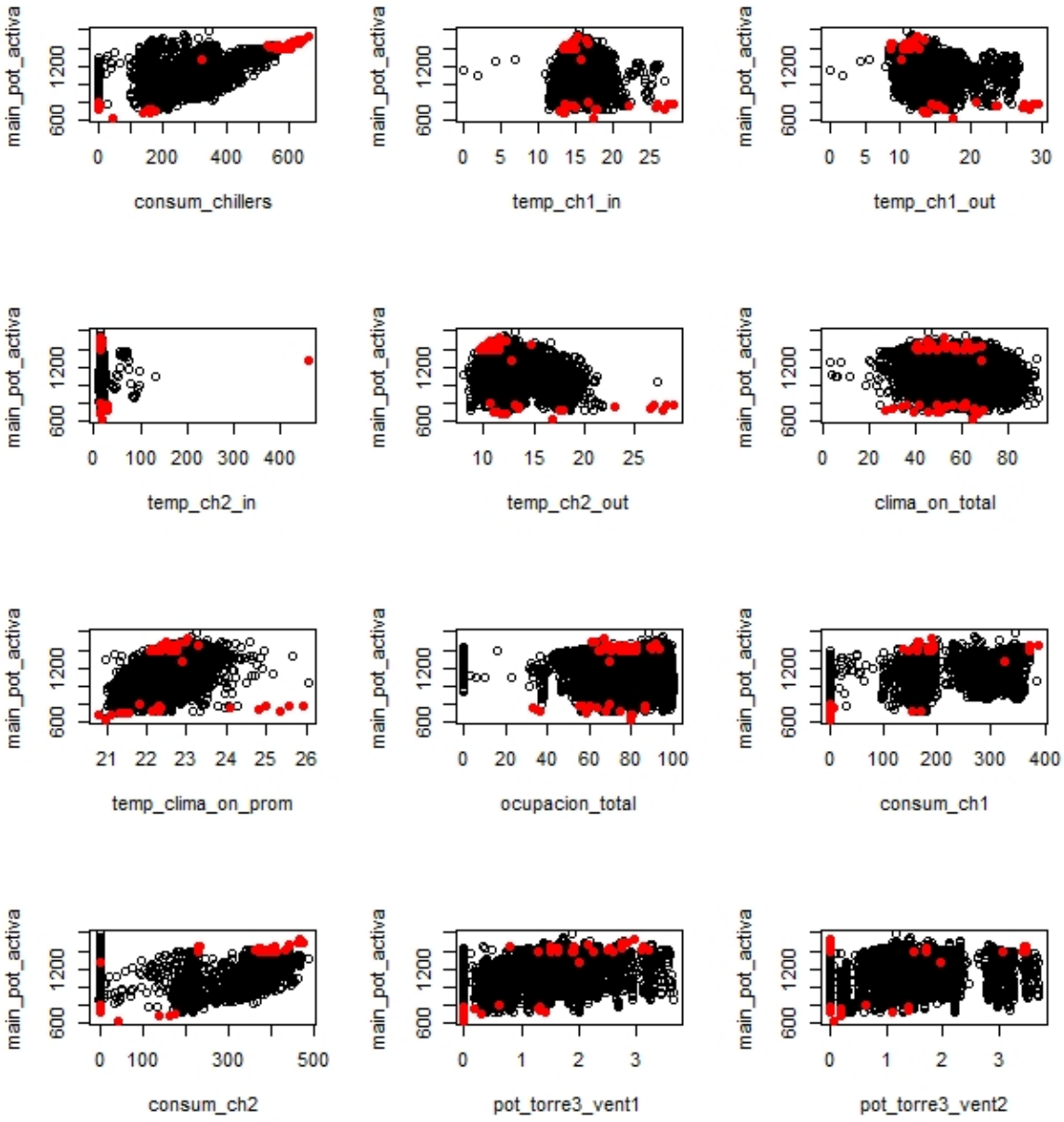


Figura 4.46: Valores atípicos detectados mediante el método de la profundidad moda (trim 1 %), mostrados sobre los diagramas de dispersión de la potencia activa principal en función de las demás variables.

desde la primera ramificación hasta la siguiente es grande, lo que denota que las distancias que pueden traducirse a su vez en similitud entre grupos (meses) también es elevada por lo que se diferencian 2 grandes grupos. Por un lado, hay un grupo compuesto por los meses enero, febrero y marzo, y por otro, los meses restantes, abril, mayo, junio y julio.

A su vez, marzo se diferencia de enero y febrero dejando una gran altura en la bifurcación, lo que implica la detección de diferencias notables con los otros dos meses. Lo mismo ocurre con julio, que se separa de los meses abril, mayo y junio a una altura bastante elevada.

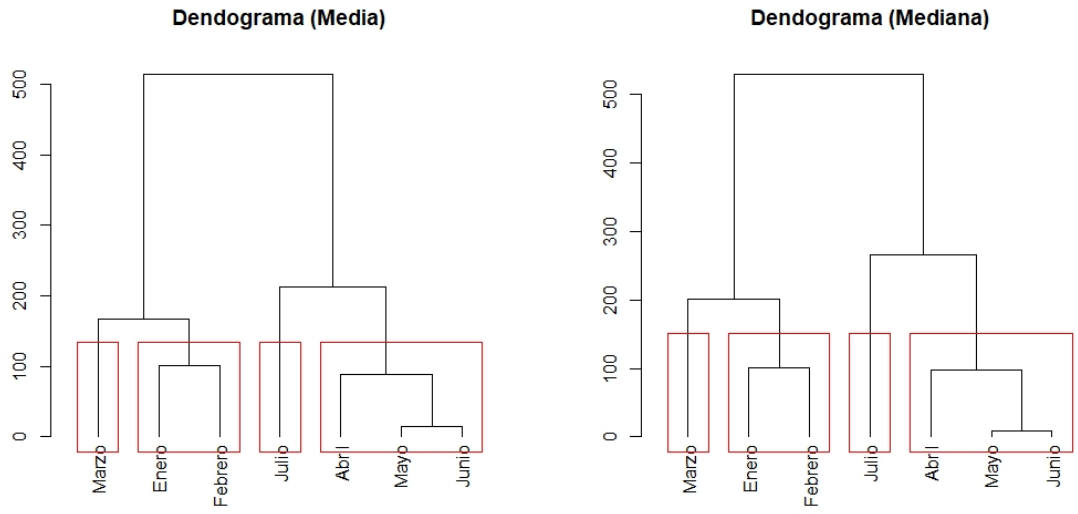


Figura 4.47: Dendograma por meses (Método jerárquico aglomerativo con distancias euclídeas).

Esta lógica puede seguir siendo aplicada con las siguientes ramificaciones de los niveles inferiores, pero sin embargo las diferencias de alturas entre el resto de meses ya no es tan elevada, por lo que puede resultar prudente e considerar estos 4 grupos en cuanto a comportamientos generales por meses se refiere.

Por lo que se concluye que, tanto empleando las medias mensuales de las variables como sus medianas, se diferencian cuatro comportamientos diferenciables, cuatro grupos. Un grupo formado por enero y febrero, otro por los meses abril, mayo y junio y finalmente otros dos grupos individuales formados por marzo y julio respectivamente.

Con idea de comprender las similitudes y/o diferencias se procede a comparar en las Tablas 4.3 y 4.4 las medias y las medianas de las variables por mes respectivamente:

Tabla 4.3: Tabla resumen medias por mes.

Variable	consum_chillers	main_pot_activa	ocupacion_total	clima_on_total	temp_clima_on_prom	temp_ext	temp_ch1_out	temp_ch1_in
Enero	300.351	1038.112	67.508	49.423	21.924	25.231	14.270	14.112
Febrero	353.869	1109.089	79.772	57.101	22.355	26.478	16.079	15.932
Marzo	354.240	1094.828	80.129	57.508	22.257	26.937	14.632	16.161
Abril	235.618	1150.991	78.184	57.921	22.376	27.461	12.091	15.072
Mayo	188.783	1196.652	75.098	55.755	22.960	28.802	12.669	15.533
Junio	192.427	1189.995	65.577	49.303	23.082	29.939	13.217	15.491
Julio	326.198	1262.622	49.341	51.971	22.936	30.408	10.832	15.041
Variable	**temp_ch2_out**	**temp_ch2_in**	**consum_ch1**	**consum_ch2**	**torre3_vent1**	**torre3_vent2**	**pot_torre3_vent1**	**pot_torre3_vent2**
Enero	11.223	14.378	3.178	297.173	0.623	0.765	0.363	0.456
Febrero	12.772	16.296	44.905	308.964	0.876	0.000	1.281	0.000
Marzo	12.512	15.376	123.921	230.318	0.952	0.781	1.561	1.647
Abril	12.841	19.683	184.773	50.845	0.986	1.000	1.829	1.711
Mayo	13.476	16.157	188.779	0.003	0.987	0.988	1.896	1.799
Junio	13.429	15.969	192.424	0.003	0.993	0.996	1.453	1.275
Julio	13.990	15.493	326.194	0.003	0.816	1.000	1.265	1.243

Tabla 4.4: Tabla resumen medianas por mes.

Variable	**consum_chillers**	**main_pot_activa**	**ocupacion_total**	**clima_on_total**	**temp_clima_on_prom**	**temp_ext**	**temp_ch1_out**	**temp_ch1_in**
Enero	322.175	1044.988	66.227	48.226	21.902	25.140	14.268	14.064
Febrero	356.497	1134.479	79.954	54.371	22.340	26.388	15.303	15.300
Marzo	358.648	1122.625	80.754	55.322	22.261	26.422	13.067	15.569
Abril	189.890	1144.250	76.330	56.155	22.430	27.071	11.100	15.100
Mayo	158.372	1234.162	73.461	52.926	22.953	28.098	12.498	15.200
Junio	157.274	1233.992	66.543	47.353	23.086	29.032	13.418	15.418
Julio	340.323	1290.496	61.321	49.050	22.959	29.425	10.900	15.121
Variable	**temp_ch2_out**	**temp_ch2_in**	**consum_ch1**	**consum_ch2**	**torre3_vent1**	**torre3_vent2**	**pot_torre3_vent1**	**pot_torre3_vent2**
Enero	11.126	14.183	0.000	321.905	1.000	1.000	0.300	0.300
Febrero	12.044	16.141	0.000	348.679	1.000	0.000	1.189	0.000
Marzo	11.690	15.098	149.084	219.610	1.000	1.000	1.494	1.420
Abril	12.296	16.600	176.261	0.003	1.000	1.000	1.802	1.708
Mayo	12.720	15.763	158.368	0.003	1.000	1.000	1.903	1.740
Junio	12.969	15.893	157.270	0.003	1.000	1.000	1.472	1.300
Julio	14.003	15.579	340.320	0.003	1.000	1.000	1.324	1.104

A priori, nótese que la información que se obtiene al calcular las medias y las medianas de cada una de las variables en referencia a cada uno de los meses de estudio es muy semejante. No obstante, en las medianas se observan valores en general más igualados entre meses, ya que esta medida no se ve tan afectada por los valores extremos, como sí que ocurre con las medias. Es por ello que anteriormente también se presentaban ambos dendograma, uno para las medias y otro para las medianas, para verificar que ambas medidas conllevaban a la misma agrupación de meses.

Observando con detenimiento las tablas resumen anteriores, es posible ver cómo ya en la primera variable, consum_chillers, se observa cierta diferenciación por meses. Esta diferenciación coincide con parte de la ramificación obtenida en el dendograma, ya uqe los meses mayo, junio y abril muestran consumos bajos en comparación con el resto de meses. De hecho, en la Figura 4.16, ya se podía intuir que existen diferencias notables en cuanto a los meses en referencia a dicha variable.

Para la variable main_pot_activa, julio se muestra muy diferenciado, registrando la potencia más elevada entre todos los meses, seguido de los meses abril, mayo y junio, que registran valores entre 1150 y 1192 kW, finalizando con enero febrero y marzo donde se tienen potencias inferiores, entorno a 1038 y 1109 kW. Julio es un mes que prácticamente en todas las variables despunta del resto, con la mínima ocupación, la temperatura externa más elevada, la temperatura del proceso de salida del chiller 1 mínima y la temperatura de salida del segundo chiller máxima. Esto deja claro que julio tiene un comportamiento distinto del resto de meses, lo que claramente es coincidente con los grupos obtenidos por el método de partición jerárquico aglomerativo.

Sin embargo, la diferenciación de marzo es algo más costosa de apreciar a simple vista, pero en líneas generales marzo registra valores similares a los de enero y febrero, a excepción de la proporción en la que trabajan los chiller en dicho mes, que mientras que en enero y febrero casi todo el consumo de los chiller viene dado por el chiller 2, en marzo esta proporción es mucho más equilibrada. Esta situación se repite en lo que a los ventiladores de la torre de enfriamiento se refiere, ya que enero y febrero registran potencias casi nulas para el ventilador 2, mientras que marzo lo hace de manera equilibrada, empleando prácticamente el ventilador 1 y el ventilador 2 por igual.

Teniendo en cuenta la formación de grupos por el método jerárquico empleado, y las medidas resumen calculadas, se observa que efectivamente es posible detectar que julio y marzo son meses que se diferencian de manera individual, mientras que los dos primeros meses, enero y febrero, al igual que los meses abril, mayo y junio, registran condiciones diferentes en el funcionamiento del sistema.

4.2.4. Análisis de componentes principales

Como se se ha mencionado anteriormente, el análisis de componentes principales (PCA) es una técnica multivariante basada en la transformación ortogonal de bases de datos compuestas por variables correladas para la reducción de la dimensionalidad obteniendo como resultado unas nuevas variables, las componentes principales, combinación lineal de las variables originales.

En el control estadístico de la calidad, el análisis de componentes principales resulta muy útil como paso previo a la aplicación de gráficos de control para datos multivariantes. De hecho, los gráficos de control para datos multivariantes son un método efectivo siempre y cuando el número de variables, p, no sea muy grande. Cuando p aumenta, el desplazamiento a detectar en el proceso tiende a diluirse el en espacio p-dimensional, aumentando el ARL (número de observaciones hasta detectar que el proceso está realmente fuera de control) de forma significativa **Hubele et al., (1991)**. Debido a esto, y teniendo en cuenta que pasar de $p = 2$ a $p = 12$ variables podría doblar el valor del ARL, podría ser útil e interesante reducir la dimensión del presente caso de estudio, definido por 12 variables, más si cabe teniendo en cuenta que este tipo de procesos son también recomendables cuando se sospecha que la variabilidad del procesos no está equitativamente distribuida entre las p variables.

Otra utilidad es el análisis descriptivo, es decir, ver qué variables son las que definen mejor el sistema en cuanto a variabilidad, pero para ello es importante asumir variables escaladas previamente. Como ocurre en el presente estudio, se tienen variables en diferentes escalas, por lo que se recomienda escalarlas previamente o utilizar la matriz de correlaciones en lugar de la matriz de varianzas-covarianzas.

Antes de comenzar con dicho análisis cabe determinar las variables que se tendrán en cuenta en

este apartado. Las variables que se van a emplear para el PCA son todas las variables numéricas, exceptuando las variables binarias correspondientes a los ventiladores de la torre de enfriamiento, que son las encargadas de indicar 1 si el ventilador está en funcionamiento o 0 si no lo está. Esta información viene dada indirectamente por las variables *torre*3_*vent*1 y *torre*3_*vent*2, las cuales sí que se contemplan en el análisis que se realiza a continuación. Además, se obvian las variables *consum_ch*1 y *consum_ch*2, ya que la suma de estas resultan en la variable *consumo_chillers*, que sí que se tendrá en cuenta. En resumen, el análisis de componentes principales tiene como variables originales de partida las 12 variables resultantes de la eliminación de aquellas variables no numéricas (*date* y *time*), las variables binarias (*torre*3_*vent*1 y *torre*3_*vent*2) y las variables de los consumos marginales de los chiller (*consum_ch*1 y *consum_ch*2).

Primeramente, para aplicar correctamente el análisis de componentes principales a los datos en cuestión, conviene aplicar un contraste de hipótesis de homogeneidad, donde se debe plantear la siguiente hipótesis:

$$\begin{aligned} H_0: &\quad \textit{Datos homogéneos} \\ H_0: &\quad \textit{Datos no homogéneos} \end{aligned} \tag{4.2}$$

Aplicando el test de Levene, efectivamente se obtiene un p-valor cercano al valor unitario, lo que implica que no existen evidencias suficientes para cualquier nivel de significación usual ($\alpha = 1\,\%, 5\,\%, 10\,\%$) para rechazar la hipótesis nula de homogeneidad en los datos. Como el test realizado resulta estadísticamente no significativo, se procede a la aplicación del análisis de componentes principales basado en la matriz de varianzas-covarianzas.

Al aplicar el PCA, el primer resultado interesante que se obtiene es la desviación estándar y la proporción de varianza explicada dada por cada una de las componentes principales. A continuación se presenta la Tabla 4.5, donde se muestran las desviaciones estándar, la varianza explicada por cada una de las componentes y la varianza explicada acumulada de ellas. Como es posible observar, se

Tabla 4.5: Tabla resumen de las componentes principales.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12
Desv. estándar	1.71	1.48	1.29	1.09	0.99	0.89	0.81	0.76	0.72	0.50	0.40	0.26
Var. explicada	0.24	0.18	0.14	0.10	0.08	0.07	0.05	0.05	0.04	0.02	0.01	0.01
Var. acumulada	0.24	0.43	0.56	0.66	0.75	0.81	0.87	0.92	0.96	0.98	0.99	1.00

obtiene que no se obtiene una variabilidad explicada acumulada del 90 % hasta la octava componente principal, y con solamente las dos primeras componentes se obtiene una variabilidad explicada de poco más del 40 %, lo que es un porcentaje bajo. De manera gráfica, se muestra la Figura 4.48, donde se muestra la varianza explicada por cada una de las componentes principales. Se observa fácilmente cómo a partir de la séptima componente principal la varianza explicada obtenida por añadir una variable adicional, se sitúa prácticamente en 0. Para conocer cómo se obtienen dichas componentes principales, se calculan los pesos de las variables originales en cada una de las componentes principales. Esta información viene resumida en la siguiente Tabla 4.6, acompañada del siguiente Figura 4.49, que muestra de manera gráfica los pesos de las variables originales sobre las siete primeras componentes principales.

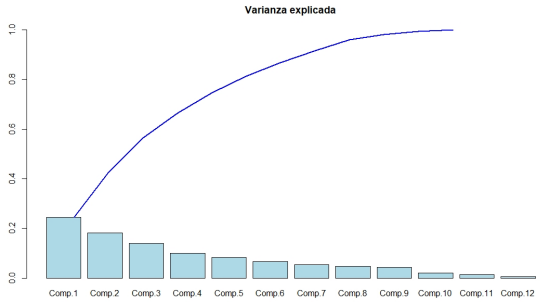


Figura 4.48: Gráfico de los pesos asociados a cada una de las variables originales en las componentes principales.

Tabla 4.6: Tabla resumen de los pesos asociados a cada una de las componentes principales.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12
consum_chillers			0.221	0.698	0.355	0.350	0.320		0.184	0.112	0.242	
main_pot_activa	0.423	0.208	0.111	0.408		-0.211	-0.193			-0.433	-0.523	0.208
ocupacion_total	-0.144	-0.330	-0.335	0.380	-0.114	-0.246	-0.134	-0.647		0.319		
clima_on_total	-0.173	-0.284	-0.426	0.160	0.107	0.389	-0.541	0.246	0.116	-0.380		
temp_clima_on_prom	0.506	-0.136	0.121					-0.308	-0.147	-0.356	0.667	
temp_ext	0.465				-0.107		-0.521	0.241	0.153	0.627		
temp_ch1_out	-0.192	-0.446	0.332	0.133	-0.289	-0.324		0.311			0.190	0.554
temp_ch2_in	0.157	-0.581	0.242		-0.109		0.135	0.154	0.148		-0.282	-0.646
temp_ch2_out	0.290	-0.398		-0.292	0.187	0.485	0.156	-0.236		0.126	-0.298	0.462
temp_ch2_in	0.102	-0.180		-0.105	0.818	-0.503		0.105				
pot_torre3_vent1	0.262		-0.463	0.207	-0.101		0.313	0.409	-0.612	0.108		
pot_torre3_vent2	0.271		-0.486		-0.131	-0.148	0.354		0.706			

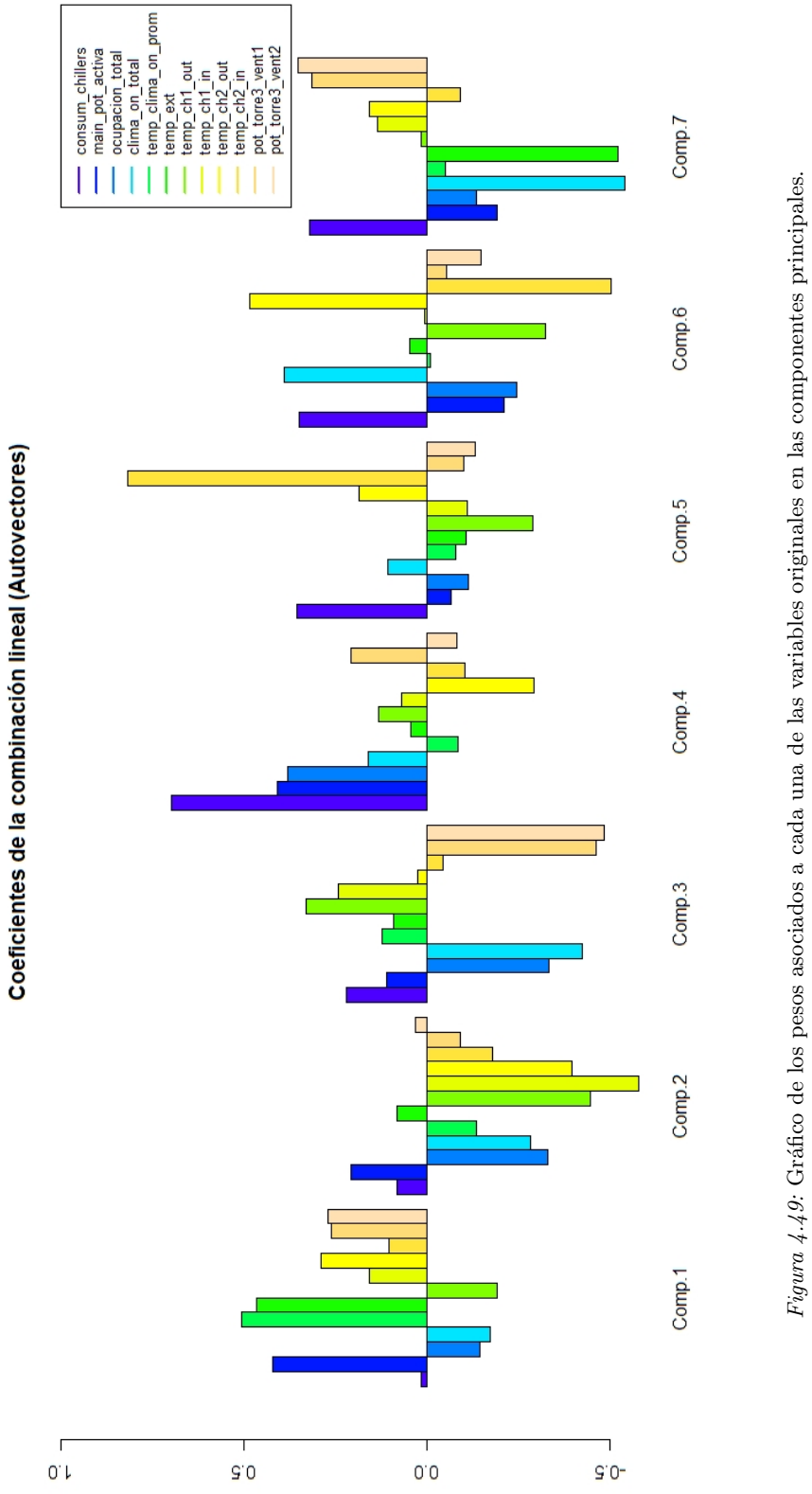


Figura 4.49: Gráfico de los pesos asociados a cada una de las variables originales en las componentes principales.

Centrando la atención en la primera componente, se tiene que esta es la nueva variable resultante principalmente por la combinación lineal de las variables *main_pot_activa*, *temp_clima_on_prom* y *temp_ext*, es decir, el consumo de potencia del sistema, la temperatura interior del hotel y finalmente la exterior. Respecto de la segunda componente principal, esta viene dada por combinación lineal de las mismas variables, solo que en este caso el mayor peso recae sobre las variables asociadas a los procesos de entrada y salida de los chiller, *temp_ch*1*_in*, *temp_ch*1*_out*, *temp_ch*2*_in* y *temp_ch*2*_out*. A diferencia de la primera componente principal, en esta segunda las variables predominantes computan negativamente. Cabe resaltar que la tercera componente principal es producto de la combinación lineal de las variables *ocupacion_total*, *clima_on_total*, *pot_torre*3*_vent*1 y *pot_torre*3*_vent* mayoritariamente, y que la variable *consum_chillers* no predomina hasta la cuarta componente principal. Esto se refleja de manera ilustrativa en la siguiente Figura 4.50, que muestra las observaciones de la muestra situada en el nuevo sistema de coordenadas formado por la primera componente principal en el eje X y la segunda en el eje de ordenadas. Además, muestra los vectores de las variables originales en este nuevo sistema de coordenadas.

La longitud de los segmentos de los vectores de las variables originales muestran la cantidad de información aportada a cada una de estas dos primeras componentes principales, y como se puede observar, prácticamente todas variables vienen representadas por segmentos de longitud considerable, excepto la variable *consum_chillers*, *pot_torre*3*_vent*1 y *pot_torre*3*_vent*. En cuanto a la dirección y el sentido de estos vectores ocurre lo mismo, que en este caso, por ejemplo, la variable *main_pot_activa* se encuentra en el primer cuadrante, debido a que los scores tanto para la componente principal como para la segunda son positivos. Empleando como ejemplo la variable $temp_ch2_out$, por el contrario, se encuentra en el cuarto cuadrante, lo que indicaban los scores asociados, que eran positivos para la primera componente y negativos para la segunda. Las observaciones originales se encuentran distribuidas según este nuevo sistema de coordenadas, lo que implica que por ejemplo, la observación número 3183, referente a la medición realizada el 13 de mayo a las 14:00 horas, está situada en la esquina inferior derecha, indicando que tendrá valores elevados de temperaturas de entrada y salida de los chiller, y sin embargo la ocupación del hotel será bajo, como muestra la Figura 4.51.

En dichos gráficos se muestran tres paneles, el primero hace referencia al consumo de los chiller, el segundo a la potencia consumida por el sistema y finalmente la temperatura interior del hotel en dicho día de mayo. Como es posible apreciar en todos ellos, a las 14:00 horas se observan valores elevados de consumo, potencia y también de temperatura interior, que se corresponde con la posición tomada en la Figura 4.50 por la observación número 3183.

4.2.5. Normalidad e independencia

La mayoría de los métodos del control estadístico de la calidad parte de las suposiciones de normalidad e independencia de los datos. Para este caso particular, tras realizar el análisis exploratorio, es totalmente lógico que las variables no resulten normales, partiendo de que con el análisis preliminar ya se intuye la existencia de varias poblaciones, cada una con su distribución. Por otro lado, la suposición de independencia significa que el valor de una observación no influye ni afecta el valor de otras observaciones, lo que rara vez se cumplirá en este tipo de datos, donde se recogen muestras en intervalos equiespaciados del tiempo.

- Normalidad:

Que un conjunto de variables de un vector X siga una distribución normal multivariante, $N_d(\mu, \Sigma)$,

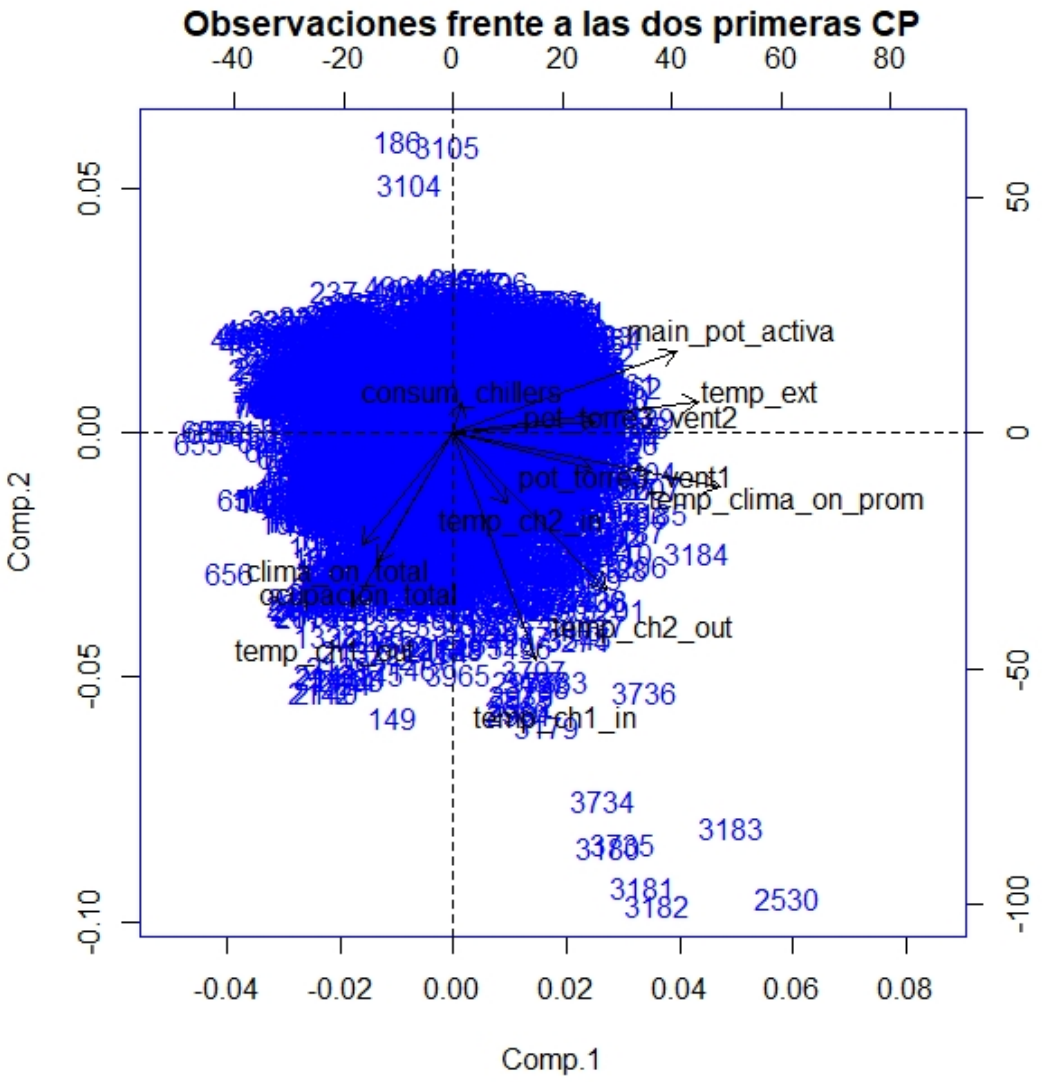


Figura 4.50: Gráfico de los pesos asociados a cada una de las variables originales en las componentes principales.

tendrá un vector de medias y una matriz de varianzas-covarianzas tal que:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1p} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{p1} & \Sigma_{p2} & \cdots & \Sigma_{pp} \end{pmatrix} \tag{4.3}$$

En este caso particular, partiendo de las 12 variables contempladas en el análisis, el vector de medias y la matriz de varianzas covarianzas, representada en la Tabla 4.7, toman los siguientes

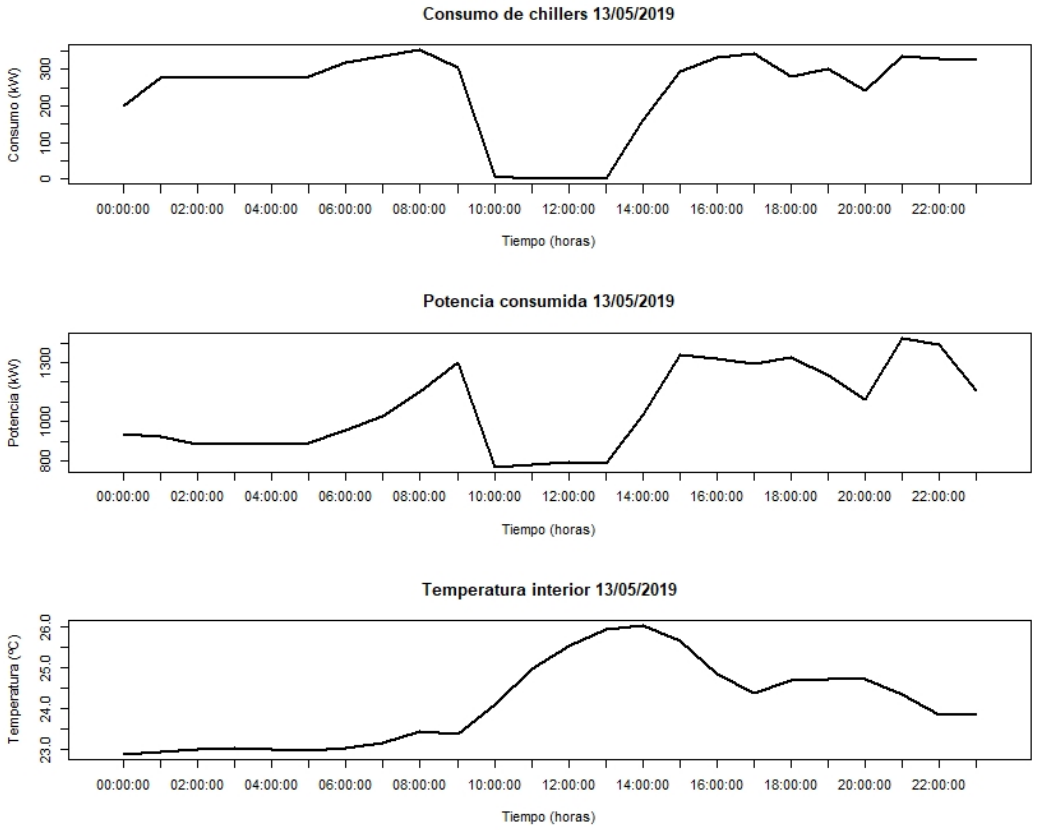


Figura 4.51: Gráfico de los pesos asociados a cada una de las variables originales en las componentes principales.

valores:

$$\mu' = \Big(278{,}84,\ 1150{,}07,\ 70{,}62,\ 54{,}11,\ 22{,}56,\ 27{,}93,\ 13{,}37,\ 15{,}34,\ 12{,}90,\ 16{,}15,\ 1{,}38,\ 1{,}17\Big) \tag{4.4}$$

Comenzando por la suposición de normalidad, se representan gráficos QQ-Plot para cada una de las variables, tratando de comparar los cuantiles de las muestras con los cuantiles teóricos de una distribución normal, ver Figura 4.52.

Como es posible observar, la mayoría de las variables se distancian de la recta representativa de los cuantiles teóricos normales, sobre todo en los extremos. No obstante, para conocer si estos datos son normales de una manera más analítica, se requiere de herramientas estadísticas como lo son los contrastes de normalidad, cuya hipótesis nula es sencillamente la siguiente:

$$\begin{aligned} H_0: &\quad \textit{Datos normales},\ X \sim N_d(\mu, \Sigma) \\ H_1: &\quad \textit{Datos no normales} \end{aligned} \tag{4.5}$$

Sin embargo, antes de aplicar este contraste cabe recordar el lema que se menciona a continuación. Y es que si se tiene un vector X normal multivariante tal que $X = (x_1, x_2, \cdots, x_p$, cada componente $x_i, \quad i = 1, 2, \cdots, p$ tiene distribución normal $N(\mu_i, \sigma_i^2)$. No obstante, es importante conocer que no necesariamente la condición recíproca sea cierta, es decir, el hecho de que la

Tabla 4.7: Tabla representativa de la matriz de varianzas-covarianzas estimada.

	consum_chillers	main_pot_activa	ocupacion_total	clima_on_total	temp_clima_on_prom	temp_ext	temp_ch1_in	temp_ch2_in	temp_ch1_out	temp_ch2_out	pot_torre3_vent1	pot_torre3_vent2
consum_chillers	11997.04	5266.94	-29.55	-58.80	-4.78	3.17	-7.92	9.23	-17.53	0.58	-3.08	-16.84
main_pot_activa	5266.94	29717.49	-470.23	-749.87	58.87	298.99	-140.05	3.08	-15.21	25.07	33.55	30.56
ocupacion_total	-29.55	-470.23	397.43	111.37	-1.44	-15.49	15.37	6.76	-0.64	6.33	2.67	1.20
clima_on_total	-58.80	-749.87	111.37	176.15	-2.26	-6.13	3.30	2.32	2.53	5.14	1.75	0.54
temp_clima_on_prom	-4.78	58.87	-1.44	-2.26	0.38	1.02	-0.22	0.44	0.76	0.63	0.13	0.13
temp_ext	3.17	298.99	-15.49	-6.13	1.02	7.63	-1.89	0.71	1.44	0.84	0.46	0.59
temp_ch1_out	-7.92	-140.05	15.37	3.30	-0.22	-1.89	9.46	3.71	-0.17	-0.25	-0.47	-0.91
temp_ch1_in	9.24	3.08	6.76	2.32	0.44	0.71	3.71	3.42	2.38	2.27	0.13	-0.03
temp_ch2_out	-17.53	-15.21	-0.64	2.53	0.76	1.44	-0.17	2.38	4.86	3.65	0.29	0.26
temp_ch2_in	0.58	25.07	6.33	5.14	0.63	0.84	-0.24	2.27	3.65	61.87	0.42	0.38
pot_torre3_vent1	-3.08	33.55	2.67	1.75	0.13	0.46	-0.47	0.13	0.29	0.42	0.64	0.31
pot_torre3_vent2	-16.84	30.56	1.21	0.54	0.13	0.59	-0.91	-0.03	0.26	0.38	0.31	0.78

densidad de cada uno de esos x_i sea normal univariante no implica necesariamente que el vector multivariante X sea normal.

A continuación se calculan algunos de los contrastes de normalidad más empleados, para conocer si la distribución de los datos es normal. Para todos ellos, la hipótesis nula a plantear es exactamente la misma:

$$\begin{aligned} H_0: &\quad \textit{Los datos son normales.} \\ H_1: &\quad \textit{Los datos no son normales.} \end{aligned} \tag{4.6}$$

- Test de Shapiro-Wilk (1965) En primer lugar se calculan los estadísticos de contraste univariantes para cada una de las 12 variables, donde se obtienen los resultados que se muestran a continuación en la siguiente tabla 4.8: Como se observa, ninguna de las variables muestra normalidad univariante.

  Sin embargo, este tipo de base de datos requiere de un contraste de normalidad multivariante, y aplicando el test de Shapiro-Wilk multivariante se obtiene el siguiente valor del estadístico y su p-valor asociado:

$$WVM = 0{,}85336 \qquad p - valor = 2{,}2 - 10^{-16} \tag{4.7}$$

  Se concluye por lo tanto que como el p-valor es inferior a cualquier valor usual del nivel de significación (1 %, 5 %, 10 %), el test resulta estadísticamente significativo, indicando que existen evidencias suficientes para rechazar la hipótesis nula de normalidad multivariante.

- Test de Mardia (1970)
  Aplicando el test de Mardia para contrastar la normalidad multivariante, se obtiene en la Tabla 4.9 el siguiente valor de curtosis y asimetría, junto con su p-valor asociado al realizar la comparación con la curtosis y asimetría de una normal: Se observa que en este caso, al

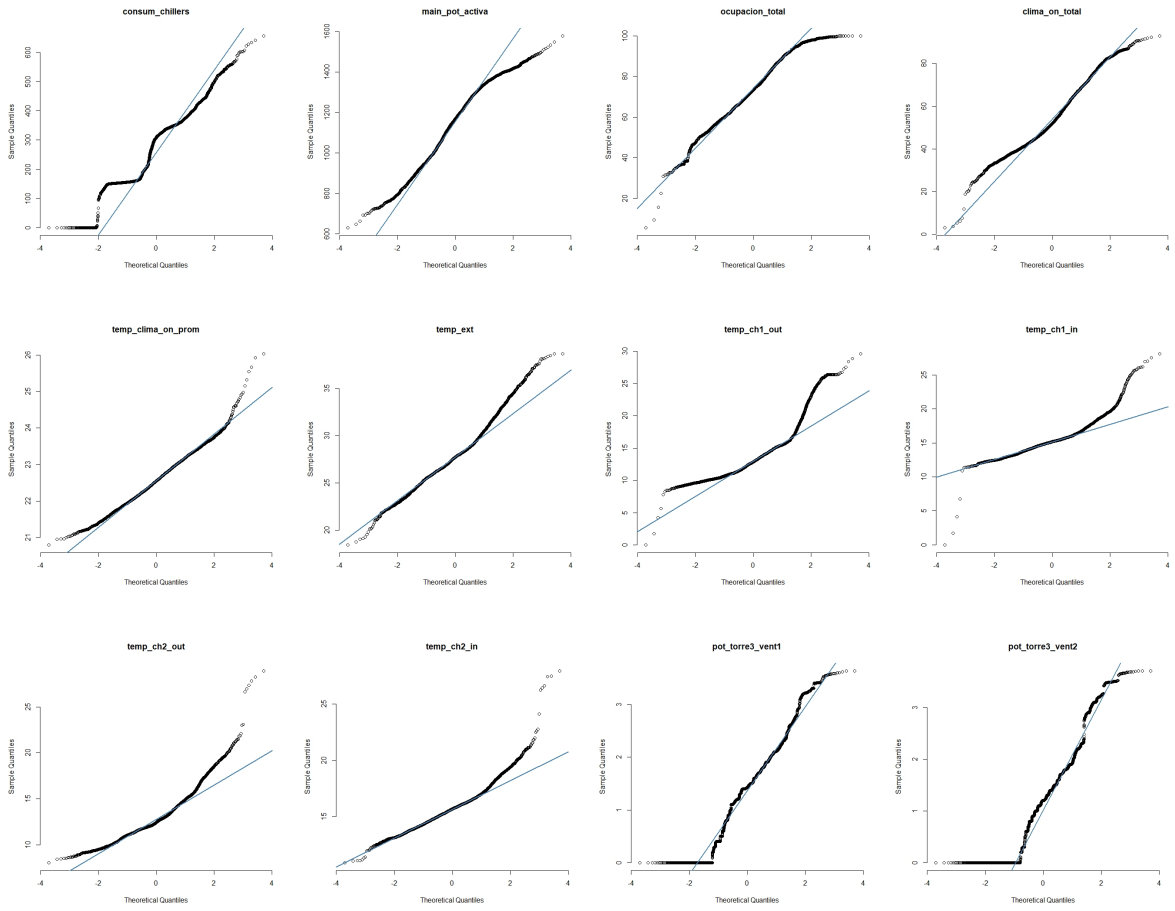


Figura 4.52: QQ-Plots de las variables (comparativa frente a la distribución normal).

igual que ocurría con el test de Shapiro-Wilk se rechaza la hipótesis nula para cualquier nivel de significación usual.

- Test de Henze y Zirkler (1990)
  Cuando se emplea el test de Henze y Zirkler, los resultados del estadístico y el p-valor asociado son los siguientes:

$$HZ = 9{,}8011 \qquad p - valor = 0 \tag{4.8}$$

  Nuevamente, con un p-valor nulo, existen evidencias suficientes para rechazar la hipótesis nula, concluyendo que los datos no se distribuyen según una normal multivariante.

- Test de Royston (1992)
  Finalmente, se aplica el test de Royston, cuyos resultados son los siguientes:

$$H = 1193{,}217 \qquad p - valor = 3{,}4282 - 10^{-248} \tag{4.9}$$

  El test de Royston devuelve un p-valor prácticamente nulo, apoyando la decisión tomada hasta el momento de rechazar la hipótesis nula planteada.

En resumen, de acuerdo a todos los contrastes realizados, todos ellos mostraban un rechazo de la hipótesis nula lo que lleva a la conclusión final de que los datos no se distribuyen según una normal

Tabla 4.8: Tabla de resultados del contraste de normalidad univariante.

Variable	Estadístico	p-valor	H_0 cierta
consum_chillers	0.9430	0.001	NO
main_pot_activa	0.9754	0.001	NO
ocupacion_total	0.8442	0.001	NO
clima_on_total	0.9759	0.001	NO
temp_clima_on_prom	0.9921	0.001	NO
temp_ext	0.9783	0.001	NO
temp_ch1_in	0.9025	0.001	NO
temp_ch2_in	0.1236	0.001	NO
temp_ch1_out	0.8680	0.001	NO
temp_ch2_out	0.9283	0.001	NO
pot_torre3_vent1	0.9695	0.001	NO
pot_torre3_vent2	0.9388	0.001	NO

Tabla 4.9: Tabla de resultados del contraste de normalidad de Mardia.

	Estadístico	p-valor	H_0 cierta
Asimetría	1441182.9917	0	NO
Curtosis	4403.1249	0	NO

multivariante. De manera que esto hace pensar que deberán realizarse ciertas agrupaciones en las muestras para poder realizar un estudio desde el punto de vista paramétrico, basándose en el Teorema Central del Limite (TCL), que dice que en condiciones muy generales, si S_n es la suma de n variables aleatorias independientes y de varianza no nula pero finita, entonces la función de distribución de S_n se aproxima bien a una distribución normal. Normalmente, es la práctica se considera una suma suficientemente grande de variables aleatorias a partir de 30. En todo caso, el no cumplimiento de la normalidad, siempre y cuando la distribución no sea muy asimétrica, no invalidaría totalmente la aplicación de técnicas como los gráficos de Shewhart. En todo caso, teniendo en cuenta este resultado, en la siguiente sección también se aplicarán alternativas no paramétricas de gráficos de control, como son los gráficos r, Q y S.

- Independencia:
  Una vez chequeado el supuesto de normalidad multivariante o multinormalidad, se procede a contrastar la independencia de las observaciones. Para ello se emplean gráficas de autocorrelación, también conocidas como correlogramas.

  Estas gráficas son una buena manera gráfica de chequear la independencia de los datos. Los correlogramas se pueden comprender de manera semejante a un contraste de hipótesis con hipótesis nula tal que:

$$\begin{aligned} H_0: &\quad \textit{Las observaiones son independientes.} \\ H_1: &\quad \textit{Las observaciones no son independientes.} \end{aligned} \tag{4.10}$$

En primer lugar, la Figura 4.53 muestra el correlograma de la muestra al completo de las 12

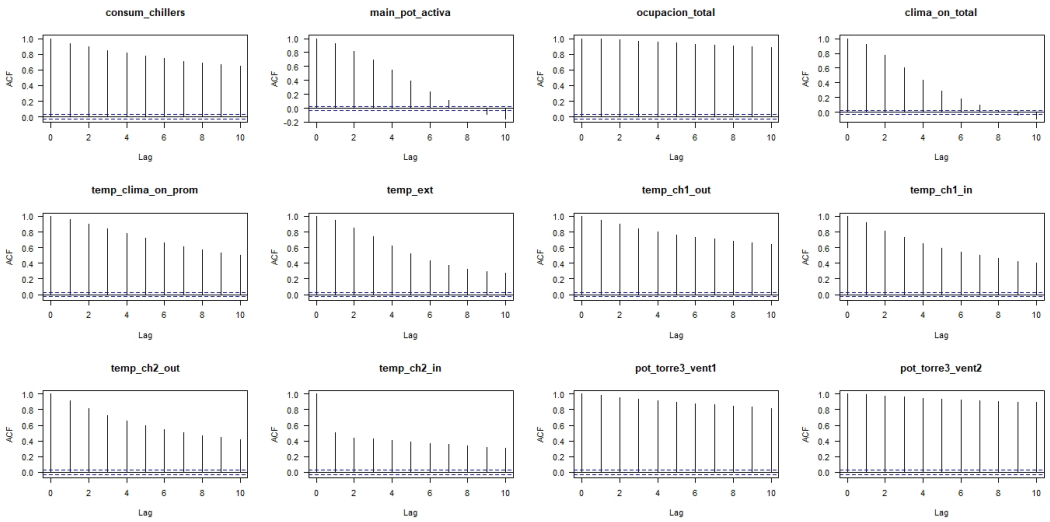


Figura 4.53: Gráfico de autocorrelaciones de los datos.

variables y todas aquellas observaciones que son completas, donde se ve que existe una fuerte dependencia temporal en todas las variables.

De igual manera, otro procedimiento que puede ayudar es tomar sólo ciertas horas del día, en ese caso, estas horas han de ser aquellas que sea más representativas. Teniendo en cuenta la ocupación, actividades del hotel y unidades activas en el sistema de refrigeración, se consideran las franjas horarias más representativas los rangos horarios que van de 2:00-4:00 horas de la madrugada y también el rango de 18:00-20:00 horas, siendo las de madrugada aquellas en las que mayor unidades activas (habitaciones con el aire acondicionado encendido), y las de la tarde correspondientes al rango de horas donde menos, ya que a esas horas de la tarde los clientes no suelen permanecer en las habitaciones sino que se encuentran en las salas comunes o asistiendo a las actividades ofrecidas por el hotel, llevando a un menor número de unidades activas.

Dentro de este contexto, puede ser de gran utilidad verificar la independencia para cada uno de los meses por separado, aplicando un sencillo test de Ljung-Box multivariante.

En este caso, como muestra la Tabla 4.10, existen numerosos p-valores resultantes del test de Ljung-Box de independencia cercanos al cero, lo que indica que para cualquier nivel de significación usual como el 1 %, 5 % o 10 % existen evidencias suficientes para rechazar la hipótesis nula de independencia. Variables como las asociadas a las potencias de los ventiladores de la torre de enfriamiento, la potencia consumida o la ocupación son aquellas variables que para todos los meses presentan dependencia.

De manera análoga, empleando el test de Ljung-Box multivariante, es posible agrupar los datos de dichos rangos horarios a su vez por otros grupos que no sean sencillamente los diferentes meses, por ejemplo por grupos de acuerdo al análisis clúster, véase B. De acuerdo al análisis clúster, y con idea de mantener la cronología de las observaciones, se pueden considerar cuatro grupos. El primer grupo queda compuesto por el mes de enero, el segundo grupo por el conjunto de datos provenientes de febrero y marzo, el tercero compuesto por el mes de abril y finalmente el cuarto grupo, compuesto por los tres meses restantes de mayo, junio y julio que presentaban comportamientos similares.

Empleando esta agrupación, se obtienen los p-valores representados en la Tabla 4.10, donde se observan numerosos p-valores resultantes del test de Ljung-Box de independencia cercanos o

Tabla 4.10: Tabla de correlaciones lineales de las variables.

Mes	consum_chillers	main_pot_activa	ocupacion_total	clima_on_total	temp_clima_on_prom	temp_ext	temp_ch1_in	temp_ch2_in	temp_ch1_out	temp_ch2_out	pot_torre3_vent1	pot_torre3_vent2
Enero	0.053	0.000	0.000	0.527	0.052	0.000	0.000	0.002	0.030	0.132	0.000	0.000
Febrero	0.294	0.000	0.000	0.333	0.501	0.000	0.000	0.000	0.003	0.822	0.000	0.000
Marzo	0.128	0.000	0.000	0.485	0.158	0.827	0.005	0.013	0.136	0.758	0.000	0.000
Abril	0.000	0.000	0.000	0.763	0.040	0.000	0.002	0.032	0.000	0.000	0.000	0.000
Mayo	0.000	0.000	0.000	0.411	0.169	0.409	0.267	0.123	0.366	0.126	0.000	0.000
Junio	0.064	0.000	0.000	0.063	0.003	0.014	0.000	0.727	0.260	0.721	0.000	0.000
Julio	0.204	0.000	0.000	0.199	0.029	0.562	0.278	0.304	0.559	0.305	0.000	0.000

Tabla 4.11: Tabla de correlaciones lineales de las variables.

Grupo	consum_chillers	main_pot_activa	ocupacion_total	clima_on_total	temp_clima_on_prom	temp_ext	temp_ch1_in	temp_ch2_in	temp_ch1_out	temp_ch2_out	pot_torre3_vent1	pot_torre3_vent2
Grupo 1	0.053	0.000	0.000	0.527	0.052	0.000	0.000	0.002	0.030	0.132	0.000	0.000
Grupo 2	0.186	0.000	0.000	0.207	0.273	0.000	0.000	0.000	0.917	0.464	0.000	0.000
Grupo 3	0.000	0.000	0.000	0.763	0.040	0.000	0.002	0.032	0.000	0.000	0.000	0.000
Grupo 4	0.000	0.000	0.000	0.064	0.138	0.011	0.000	0.434	0.080	0.427	0.000	0.000

iguales a cero, lo que indica que para cualquier nivel de significación usual existen evidencias suficientes para rechazar la hipótesis nula de independencia. Variables como las asociadas a las potencias de los ventiladores de la torre de enfriamiento, la potencia consumida o la ocupación son aquellas variables que para todos los mese presentan dependencia, mientras que variables como las temperaturas asociadas a los procesos de entrada y salida de los chiller muestran p-valores superiores a los niveles de significación usuales 1 % o 5 %.

4.3. Control estadístico de las instalaciones

Una vez realizado el análisis anterior, queda evidenciado que los datos a tratar son datos no normales. Por lo tanto, es recomendable la aplicación de alternativas no paramétricas de gráficos de control como son los gráficos r, S y Q.

En todo caso, también se evaluarán los resultados proporcionados por la aplicación de gráficos de control paramétricos multivariantes, para observar hasta que punto son robustos frente al incumpli-

miento de las hipótesis de partida, además de comparar su desempeño con el correspondiente a los gráficos no paramétricos.

4.3.1. Gráficos de control multivariantes paramétricos

En este momento conviene primeramente diferenciar los datos empleados para la Fase I y la Fase II del control estadístico. El primer conjunto de datos, empleado para la fase de calibrado Fase I y denominado datos1, viene dado por un conjunto de datos que se encuentran bajo control, compuesto por exactamente 130 observaciones horarias compuesta por las 12 variables. Por otro lado, el grupo de datos empleado para la segunda de las fases, está compuesto por el resto de observaciones. En este segundo conjunto de datos se incluyen también todas aquellas observaciones correspondientes a los cambios de chiller y apagados generales realizados por la organización del hotel, es decir, que la base de datos datos2 contiene de manera intencionada ciertos valores que se encuentran fuera de control por estas situaciones anómalas mencionadas anteriormente.

En otras palabras, en la Fase I se estiman los límites de control naturales de las variables y en la Fase II se monitorizan las nuevas observaciones, contrastándose si es verosímil o no que pertenezcan a la misma población que la muestra de calibrado (muestra de la Fase I).

Primeramente, se procede a obtener el gráfico de control multivariante paramétrico Chi-Cuadrado, empleando el paquete de R MSQC, cuyos resultados tanto para la Fase I como para la Fase II se encuentran representados a continuación en la Figura 4.54. En primer lugar, se observa cómo el gráfico

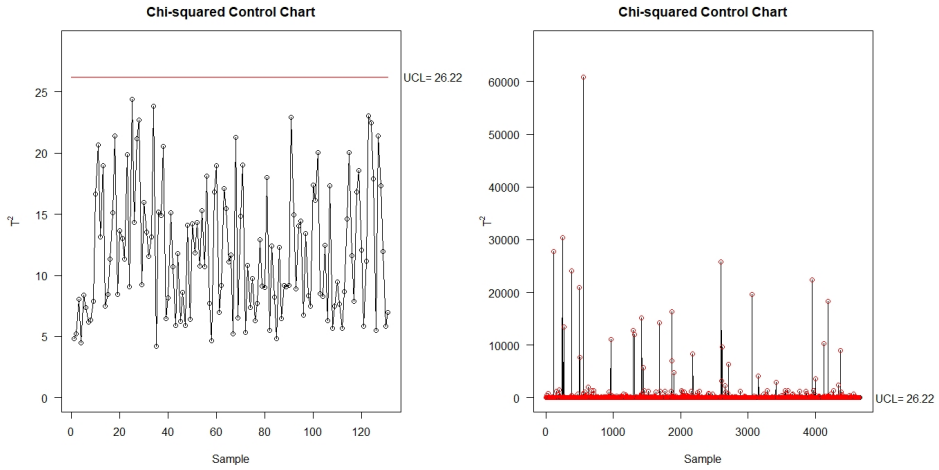


Figura 4.54: Gráfico de control Chi-Cuadrado para la Fase I (panel izquierdo) y para la Fase II (panel derecho).

de Chi-Cuadrado muestra para la Fase I que todas las observaciones se encuentran bajo control. Por otro lado, para la Fase II se muestran varias observaciones que el método gráfico clasifica como fuera de control.

El límite UCL de control alcanza un valor de 26.22, siendo observaciones fuera de control todas aquellas observaciones que obtienen un valor del estadístico superior a este valor. Se obtiene un porcentaje de fuera de control del 20.84 %, de los cuales solamente coinciden con las anomalías detectadas por el hotel en un 7.9 %, lo que implica que del total de observaciones un 19.11 % son detectadas como falsas alarmas.

El gráfico de control multivariante paramétrico T^2 de Hotelling se representa en la Figura 4.55, con la misma diferenciación de datos y empleando el mismo paquete estadístico de R, empleando datos1 para la Fase I y el conjunto datos2 ara la Fase II. En este caso, para el gráfico de control T^2 de Hotelling el límite UCL de control alcanza un valor de 30.88, clasificando como observaciones fuera de control todas aquellas observaciones que obtienen un valor del estadístico superior a dicho valor.

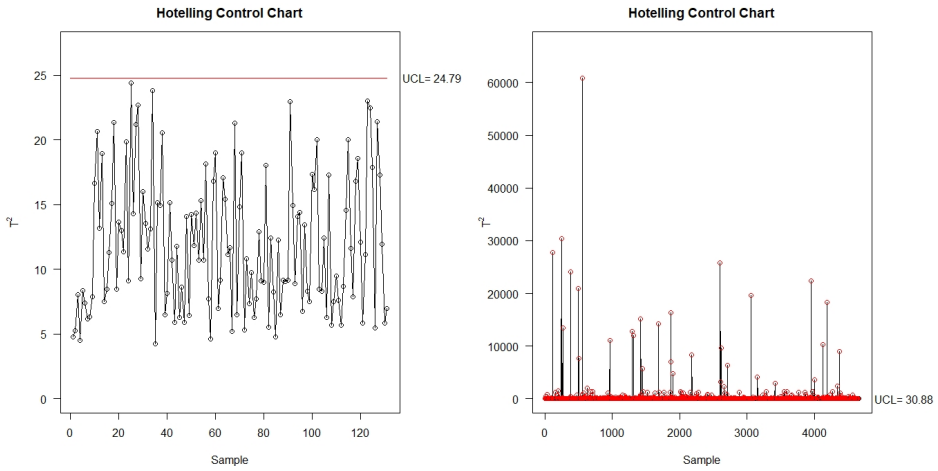


Figura 4.55: Gráfico de control T^2 Hotelling para la Fase I (panel izquierdo) y para la Fase II (panel derecho).

Se obtiene un porcentaje de fuera de control del 17.25 % sobre el total de observaciones, de los cuales solamente coinciden con las anomalías detectadas por el hotel en un 9.7 %, lo que implica que del total de observaciones un 15.52 % sobre el total de observaciones son detectadas como falsas alarmas.

A continuación, se emplea el análisis de componentes principales para posteriormente aplicar los gráficos Chi-Cuadrado y T^2 de Hotelling sobre las nuevas componentes. El uso de las 3 primeras componentes principales viene motivado por el elevado número de variables a tener en cuenta. Se procede primero a realizar el gráfico de control Chi-Cuadrado basado en el análisis de componentes principales, de donde se obtiene el siguiente resultado de la Figura 4.56. En este caso, se observa

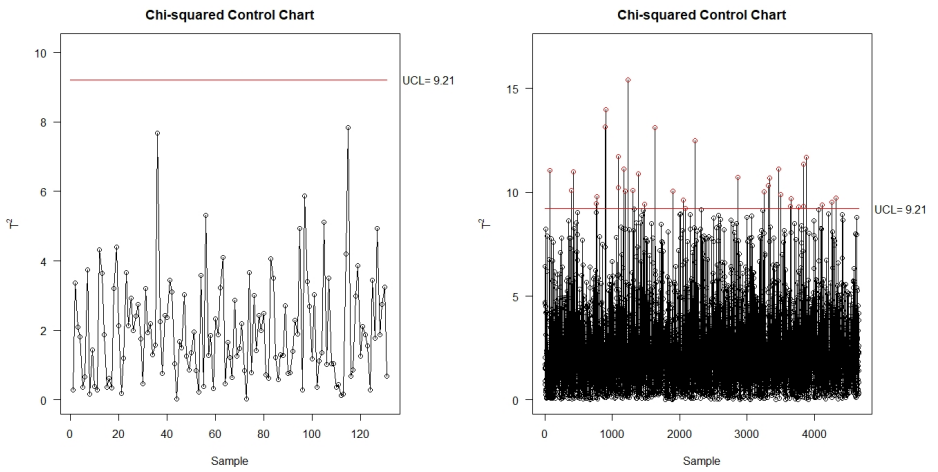


Figura 4.56: Gráfico de control basado en PCA Chi-Cuadrado para la Fase I (panel izquierdo) y para la Fase II (panel derecho).

nuevamente que el análisis sobre la primera fase, con datos1, muestra todas sus observaciones bajo control, y en el caso de la segunda fase, existe un total de 0.75 % de observaciones clasificadas como datos fuera de control, siendo el límite de control UCL=9.21, de las cuales solamente coinciden con las anomalías detectadas por el hotel en un 11.43 %, lo que implica que del total de observaciones un 0.72 % son detectadas como falsas alarmas.

Análogamente, empleando el mismo análisis de componentes principales, se obtiene en la Figura

4.57 el gráfico de control T^2 Hotelling basado en PCA. En este caso, se obtiene un resultado bastante

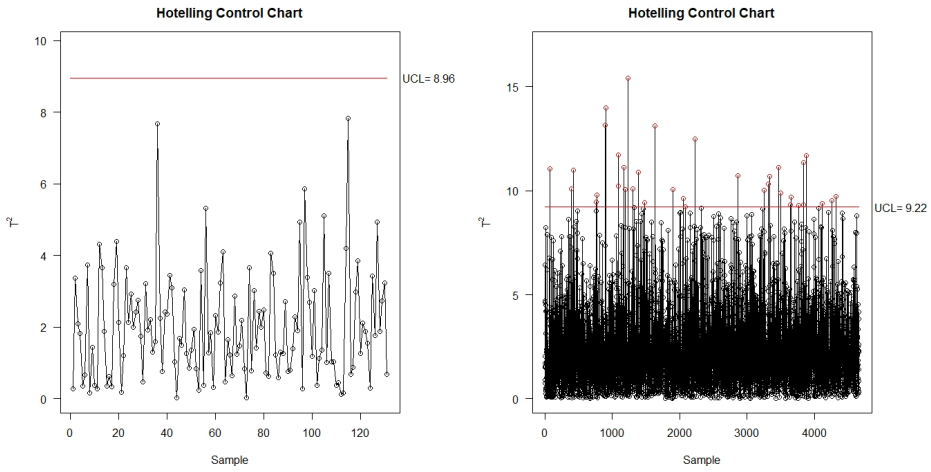


Figura 4.57: Gráfico de control basado en PCA T^2 Hotelling para la Fase I (panel izquierdo) y para la Fase II (panel derecho).

similar al obtenido en el análisis previo del gráfico de control Chi-Cuadrado basado en el análisis de componentes principales, con un porcentaje de observaciones de fuera de control del 0.75 %, con un límite de control UCL=9.22, de los cuales solamente coinciden con las anomalías detectadas por el hotel en 14.28 %, lo que implica que del total de observaciones un 0.69 % son detectadas como falsas alarmas.

A continuación, se muestran los gráficos MEWMA, empleando en este caso la librería mvdalab de R, primeramente para los datos empleados para el calibrado del proceso, representado en la Figura 4.58. Tanto los gráficos MEWMA como los gráficos MCUSUM, pertenecen a la tipología de gráficos con memoria, es decir, que detectan mejor los pequeños cambios ocurridos en el proceso en comparación por ejemplo con los gráficos de control T^2 de Hotelling. En la Figura 4.58, se observa la disposición del

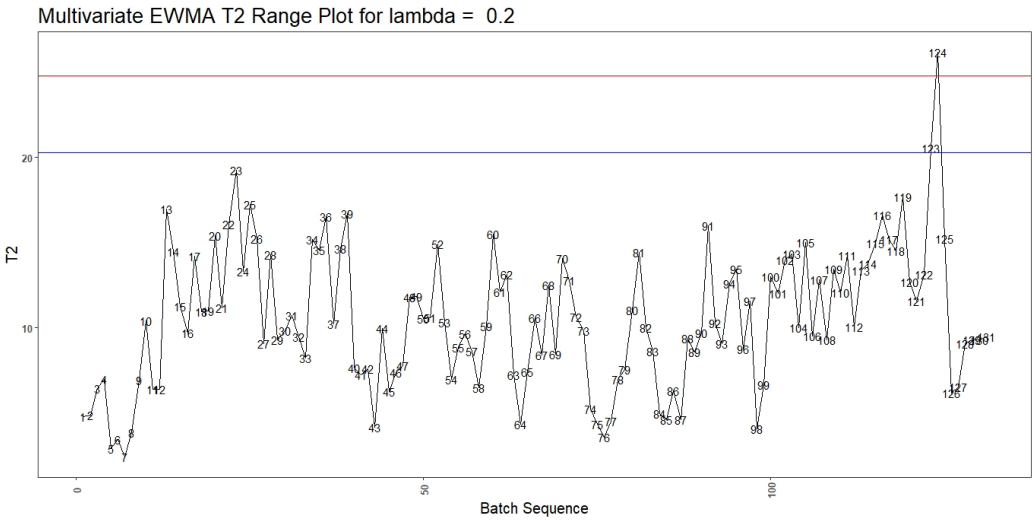


Figura 4.58: Gráfico de control MEWMA para Fase I.

estadístico junto con dos líneas horizontales, correspondientes a los límites de control asociados al nivel de confianza del 95 % y al 99 %. Para el caso del 95 % de confianza, existen dos datos que superan dicho

límite, lo que lleva a un 1.52 % de observaciones que se clasifican como observaciones fuera de control. De manera análoga, para el caso del 99 % de confianza, existen en este caso solamente un dato que lo supera, implicando que el 0.76 % de observaciones clasificadas como observaciones fuera de control.

Para la segunda fase, se obtiene en la Figura 4.59, donde como se observa, se obtienen muchos puntos fuera de control para ambos niveles de confianza, como también venía ocurriendo en los gráficos paramétricos anteriores. En este caso, para obtener una idea de la cantidad de observaciones fuera de

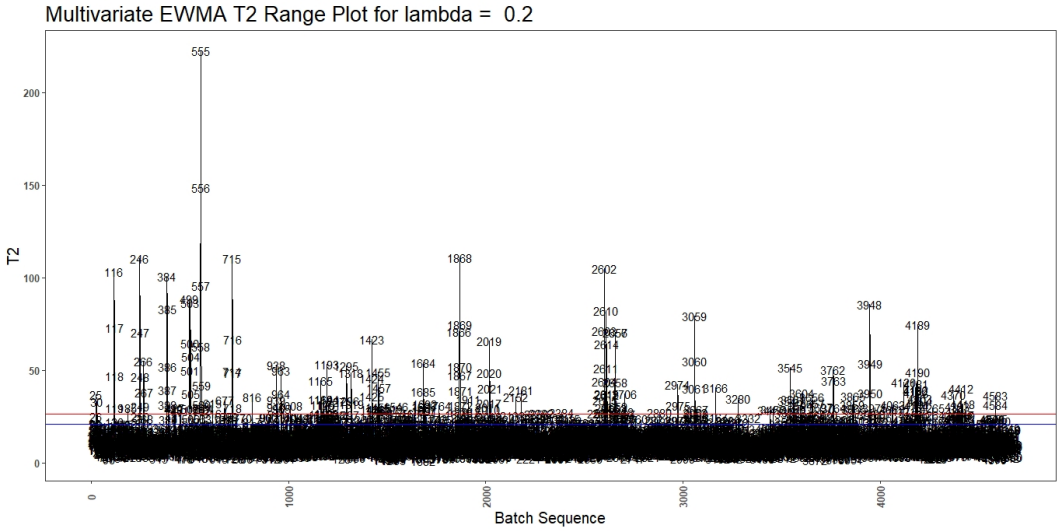


Figura 4.59: Gráfico de control MEWMA para Fase II.

control se obtiene un porcentaje de observaciones fuera de control de 7.62 % y 4.03 % para los niveles de confianza del 95 % y 99 % respectivamente. De dichas observaciones fuera de control, solamente coinciden con las anomalías detectadas por el hotel el 10.56 % de las anomalías detectadas por el gráfico MEWMA, lo que implica que del total de observaciones un 2.95 % son detectadas como falsas alarmas.

A modo de resumen, a continuación se presenta la Tabla 4.12, donde se muestran los porcentajes de fuera de control y de falsas alarmas sobre el número total de observaciones tras aplicar los gráficos de control. paramétricos.

Tabla 4.12: Tabla resumen de los gráficos de control multivariantes paramétricos

Método	% Fuera de control	% Falsas alarmas
Chi-Cuadrado	20.84	19.11
T^2 Hotelling	17.25	15.52
PCA-Chi-Cuadrado	0.75	0.72
PCA-T^2 Hotelling	0.75	0.69
MEWMA	7.62	2.95

Como se puede ver, tanto Chi-Cuadrado como T^2 Hotelling obtienen un número muy elevado tanto de observaciones fuera de control como de falsas alarmas. Empleando el análisis de componentes

principales, por otro lado, a penas se obtrienen detecciones de observaciones fuera de control, además de que las pocas detectadas son falsas alarmas. El método MEWMA, sin embargo muestra mejores resultados, con un porcentaje de observaciones fuera de control más moderado y un porcentaje de falsas alarmas entorno al 3 %.

4.3.2. Gráficos de control multivariantes no paramétricos

En este apartado, se van a aplicar los gráficos de control multivariantes no paramétricos,dado que presentan la ventaja de no asumir una distribución paramétrica a priori. Los gráficos que se van a emplear a continuación son aquellos propuestos por **Regina y Liu (1995)**. Se representarán los gráficos r, Q y S^* para la base de datos compuesta por las 12 variables no normales tratadas hasta el momento, véase el Anexo C.

En primer lugar se emplea el gráfico de rangos r, que se muestra a continuación en la Figura 4.60.

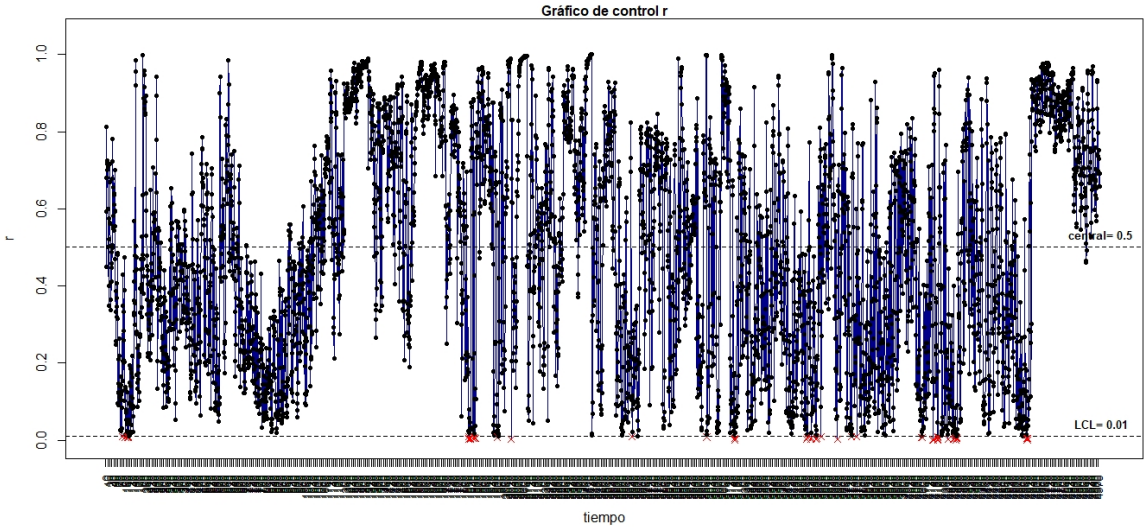


Figura 4.60: Gráfico de control multivariante r.

En este caso, considera nuevamente un total del 0.98 % de las observaciones como observaciones fuera de control, con un límite de control LCL=0.01.

En la Tabla 4.13 se muestran las fechas asociadas a las observaciones detectadas como observaciones fuera de control por este procedimiento. Es posible realizar una comparación de estas fechas con las fechas referentes a las anomalías registradas por el hotel, donde se observa que ninguna fecha es coincidente, esto lleva a concluir que mediante este método todas las alarmas detectadas son falsas alarmas.

A continuación se procede a aplicar a los datos el método del gráfico S^* propuesto por **Regina y Liu (1995)**. Este gráfico es el equivalente al gráfico paramétrico de sumas acumuladas, lo que hace lógico que el número de observaciones clasificadas como fuera de control sea superior a los dos métodos no paramétricos anteriores, ya que estos son altamente sensibles a las pequeñas modificaciones del proceso. En este caso, el método considera un total del 38.73 % de las observaciones como observaciones fuera de control, con un límite de control LCL=-2.33.

Las observaciones consideradas como fuera de control en la Figura 4.61 son aquellas que se muestran en la Tabla 4.14, donde se observan dos periodos de rechazo de la hipótesis nula de proceso bajo control. Esto indica que a partir del 15 de enero comienza ha haber modificaciones en el proceso grandes, y que el proceso no se termina de estabilizar hasta pasada la primera semana de marzo. Ocurriendo lo

Tabla 4.13: Tabla resumen con las fechas de las observaciones fuera de control detectados por Gráfico r no paramétrico.

Fecha y hora
2019-01-04 a las 08:00
2019-01-05 a las 01:00, 13:00, 15:00 y 16:00
2019-03-19 de 9:00 a 12:00 y de 18:00 a 19:00
2019-03-20 a las 9:00 y de 18:00 a 19:00
2019-03-25 a las 11:00
2019-03-28 a las 8:00
2019-04-24 a las 20:00
2019-05-10 a las 10:00
2019-05-16 de 6:00 a 8:00
2019-05-31 a las 7:00 y a las 9:00
2019-06-01 de 5:00 a 6:00
2019-06-02 a las 5:00 y a las 8:00
2019-06-03 a las 7:00
2019-06-07 de 5:00 a 6:00
2019-06-10 a las 4:00
2019-06-11 a las 6:00
2019-06-24 a las 20:00 y a las 23:00
2019-06-27 de 5:00 a 6:00
2019-06-28 de 3:00 a 6:00
2019-06-30 a las 7:00
2019-07-01 a las 10:00, 21:00 y 23:00
2019-07-02 de 00:00 a 2:00
2019-07-16 de 19:00 a 21:00

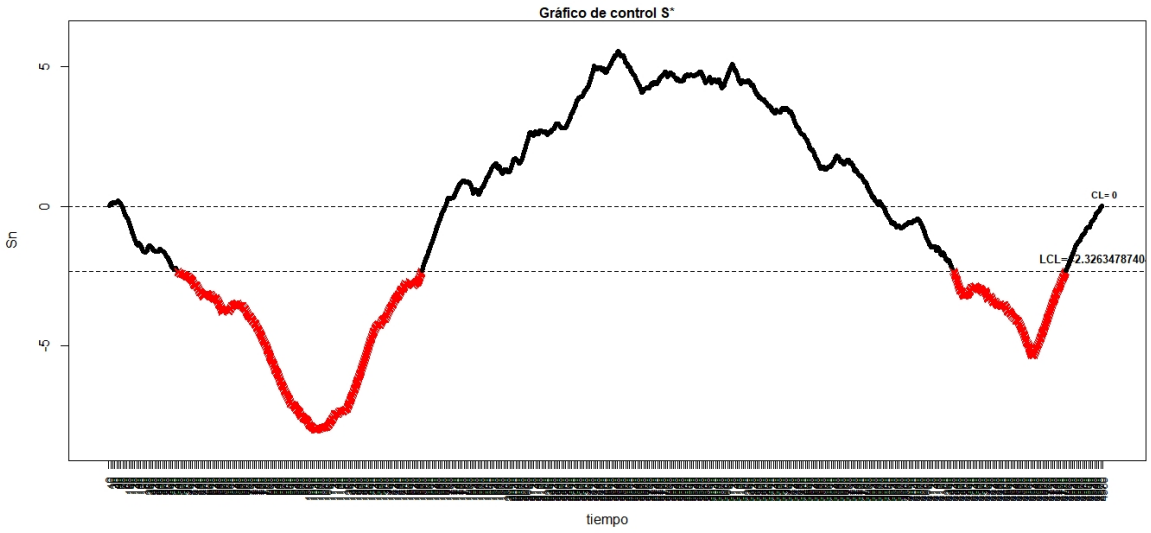


Figura 4.61: Gráfico de control multivariante S^*.

Tabla 4.14: Tabla resumen con las fechas de las observaciones fuera de control detectados por Gráfico S^* no paramétrico.

Fecha y hora
Desde el 2019-01-15 a las 10:00 hasta el 2019-03-09 a las 01:00
Desde el 2019-06-30 a las 21:00 hasta el 2019-07-24 a las 08:00

mismo a finales de junio, y estabilizándose a finales de julio, entorno al día 24. Se observa que este procedimiento obtiene un porcentaje de observaciones fuera de control elevado, del 38.73 %, lo cual es lógico ya que este gráfico es más sensible a pequeños cambios en comparación con el anterior o el T^2 de Hotelling, sin embargo, dentro de esas observaciones fuera de control el 11.92 % coinciden con fechas registradas como anomalías por el hotel. En definitiva, esto lleva directamente a un porcentaje del 34.17 % de falsas alarmas sobre el total de observaciones, lo cual es elevado. Estos resultados se pueden observar en la Tabla 4.15 que se muestra a continuación.

Tabla 4.15: Tabla resumen de los gráficos de control multivariantes no paramétricos

Método	% Fuera de control	% Falsas alarmas
Gráfico r	0.98	0.98
Gráfico S	38.73	34.17

4.3.3. Gráficos de control multivariantes paramétricos con agrupación de datos

El número de datos, teniendo en cuenta la cantidad de observaciones y variables de la base de datos en cuestión, puede resultar muy práctico emplear cierta agrupación. En este caso, el hotel cree conveniente estudiar las variables asociadas a las temperaturas de entrada y salida de los chiller, centrando el estudio en aquellos rangos de horas con mayor y menor actividad. Este método puede favorecer el cumplimiento de la suposición de independencia requerida en la aplicación de los gráficos de control. Por ello, el hotel determina como horas de actividad mínima del sistema de enfriamiento el rango horario de 2:00-4:00 de la madrugada, y el rango de 18:00 a 20:00 como el rango de máxima actividad y por lo tanto mayor unidades de funcionamiento en el hotel, siendo estas unidades habitaciones, restaurantes y salones.

De esta manera, se establecen como observaciones la media de las variables de temperaturas de entrada y salida de los chiller calculadas a partir del rango de actividad mínima del sistema, de 2:00 a 4:00, y por otro lado, la media resultante del rango de actividad máxima de 18:00 a 20:00. De esta manera, se tendrán solamente dos observaciones diarias, en lugar de las 24 correspondientes a las observaciones originales horarias.

Al mismo tiempo, de acuerdo al análisis exploratorio realizado, queda evidenciado que existen diferencias en los comportamientos a lo largo del periodo de estudio. De acuerdo al análisis clúster, véase el Apéndice B, y con idea de mantener la cronología de las observaciones, se pueden considerar cuatro grupos. El primer grupo queda compuesto por el mes de enero, el segundo grupo por el conjunto de datos provenientes de febrero y marzo, el tercero compuesto por el mes de abril y finalmente el cuarto grupo, compuesto por los tres meses restantes de mayo, junio y julio que presentaban comportamientos similares.

Con esta nueva agrupación, si se plantea un análisis de independencia, como puede ser un test de

Ljung-Box, se obtiene que las observaciones de las variables son independientes exceptuando alguna del mes de abril o grupo tres. Por otro lado, por lo que a normalidad se refiere, es posible asumir que por el Teorema Central del Límite, al estar agrupando un número elevado de observaciones ($n \geq 30$), esta converge a una normal. Es decir, al trabajar con medias de las observaciones horarias, las variables tenderán a acercarse más a la normalidad.

En la siguiente Figura 4.62 se puede observar el gráfico paramétrico de T^2 de Hotelling tanto para la fase I (panel izquierdo) como para la fase II (panel derecho) aplicado al primer grupo, o lo que es lo mismo a las observaciones del mes de enero. Para el grupo primero se obtiene un 13.56 %

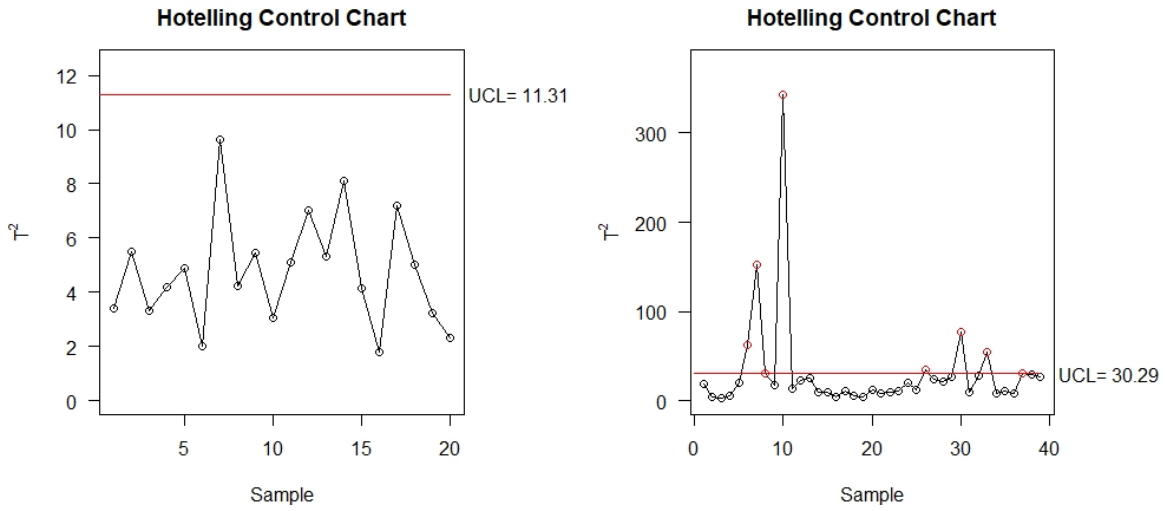


Figura 4.62: Gráfico de control multivariante T^2 de Hotelling del grupo 1.

de observaciones fuera de control, asociadas a los días 6,7,8,9,22,24,27 y 29 de enero. Es decir el procedimiento es capaz de detectar los días 7,8,9 y 22 como observaciones fuera de control, pero sin embargo no detecta el día 21, que sí que es considerado como dato anómalo por el hotel. No obstante la anomalía es la asociada al tiempo contenido entre el día 21 de enero a las 3:00 hasta las 9:00 horas del día 22, por lo que puede decirse que detectando el día 22, sería suficiente para detectar la causa asignable a dicha anomalía, que corresponde con un reajuste manual de la temperatura requerida para el interior del hotel.

Siguiendo el mismo procedimiento para el segundo de los grupos, compuesto por las observaciones de los meses de febrero y marzo, se obtienen los gráficos de control representados en la Figura 4.63. En este caso, se obtiene que del total de observaciones pertenecientes al grupo dos, un 29.66 % se detectan como observaciones fuera de control. En este caso, en el mes de febrero los días con observaciones fuera de control son los días 16,18,19,20,21,26,27 y 28, mientras que en el mes de marzo son los días 1,2,3,4,11,13,15,16,17,18,19,20,21,22,23,24,25,26,30 y 31. Este gran número de detectados en marzo puede deberse a que el hotel se da cuenta de que el sensor de temperatura de uno de los chiller está fallando el día 29 de marzo, pero no se conoce exactamente desde cuando estaba registrando datos erróneos. No obstante, se detectan los días 16,18,19,20 y 21 de febrero que son aquellos días en los que se producen anomalías por encendidos forzados del sistema de enfriamiento. De nuevo, el procedimiento no es capaz de detectar los días 15 y 17 de febrero, pero al detectar el resto de días colindantes, la causa asignable quedaría descubierta.

Por lo que al tercer grupo respecta, se detectan como observaciones fuera de control los días 1,5,11,13,15,16,17,18,19,21,27 y 28 de abril. Detectando un total de 25.45 % de las observaciones como anomalías del sistema. Como se puede observar en la Figura 4.64, es un porcentaje elevado de observaciones fuera de control, ya que el hotel solamente reconoce como atípicos los días 14,15 y 16, lo que indica que existe un 21 % de falsas alarmas.

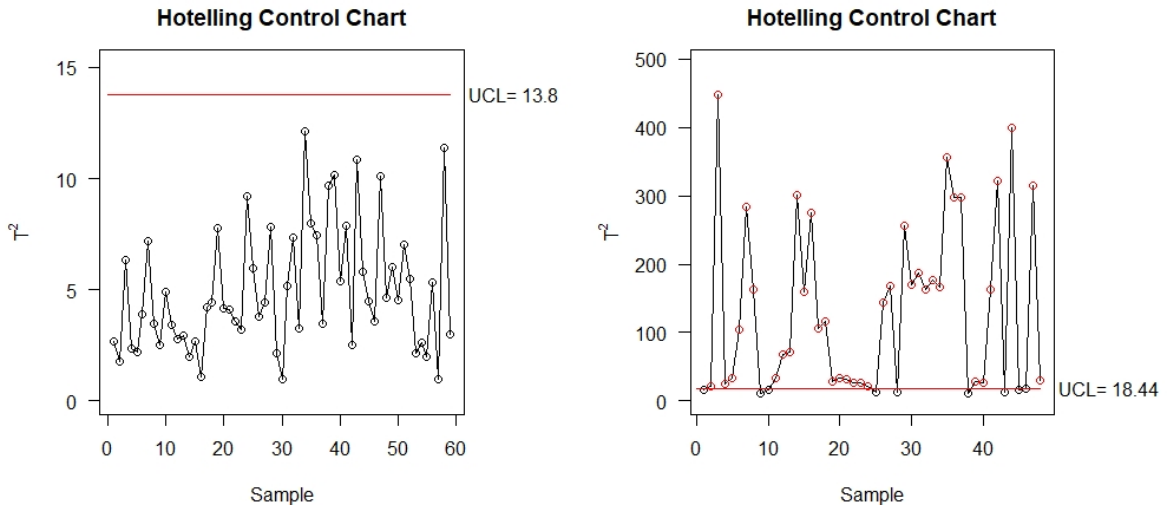


Figura 4.63: Gráfico de control multivariante T^2 de Hotelling del grupo 2, el gráfico asociado a la fase I (panel izquierdo) y fase II (panel derecho).

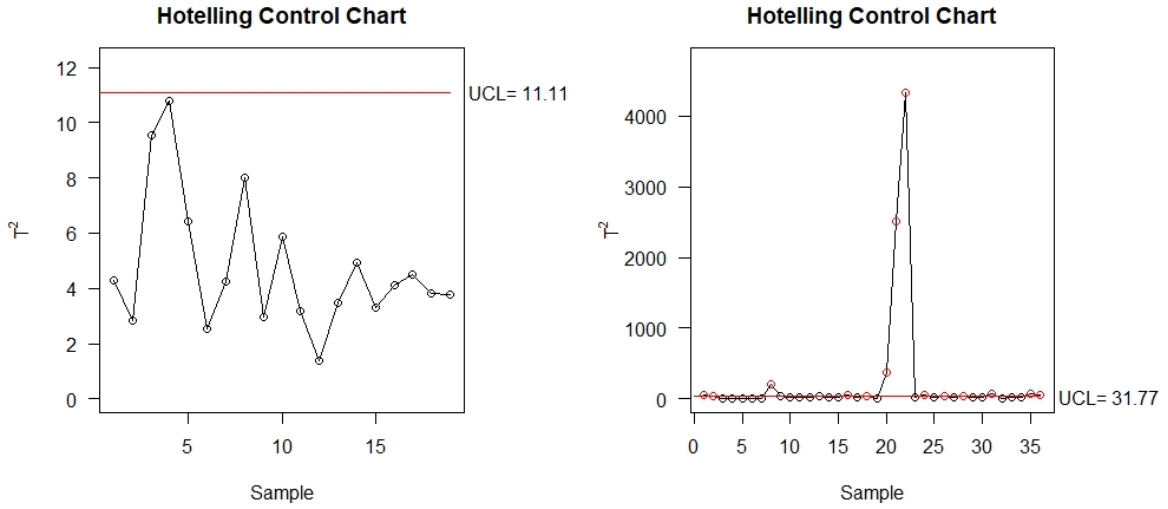


Figura 4.64: Gráfico de control multivariante T^2 de Hotelling del grupo 3, el gráfico asociado a la fase I (panel izquierdo) y fase II (panel derecho).

Por último, el cuarto grupo, compuesto por las observaciones de mayo, junio y julio, presenta un 14.2 % de observaciones fuera de control, siendo estas los días 5,7,11,12,13,14,17,2,24 y 29 de mayo, los días 4,5,8,15,28 y 29 de junio y los días 3 y 4 de julio. En este caso los días detectados por el hotel también son detectados como observaciones fuera de control por el procedimiento, el 13 de mayo y el 5 de junio. Los gráficos de control aplicados al cuarto grupo se representan en la Figura 4.65. En definitiva, este método determina que el 32.6 % de las observaciones del periodo de enero a julio son observaciones fuera de control, lo que resulta elevado, pero coincide con las anomalías registradas con el hotel, detectando el 76.5 % de las anomalías reales, y dejando sin detectar algunas fechas pertenecientes a rangos de días sí detectados, lo que se resume en que a pesar de no detectar todos los días asociados a anomalías, sí que detecta todas las causas asignables.

En este punto, podría plantearse el uso de gráficos de control no paramétricos motivado por la

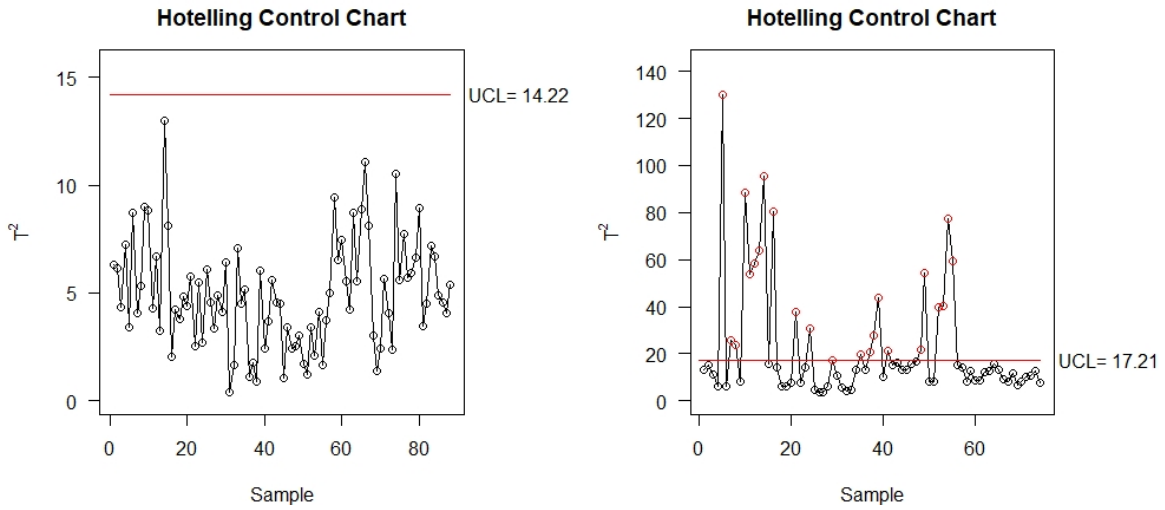


Figura 4.65: Gráfico de control multivariante T^2 de Hotelling del grupo 4, el gráfico asociado a la fase I (panel izquierdo) y fase II (panel derecho).

naturaleza de los datos. La Figura 4.66 muestra cómo estos métodos no paramétricos no son capaces de detectar observaciones fuera de control en el sistema.

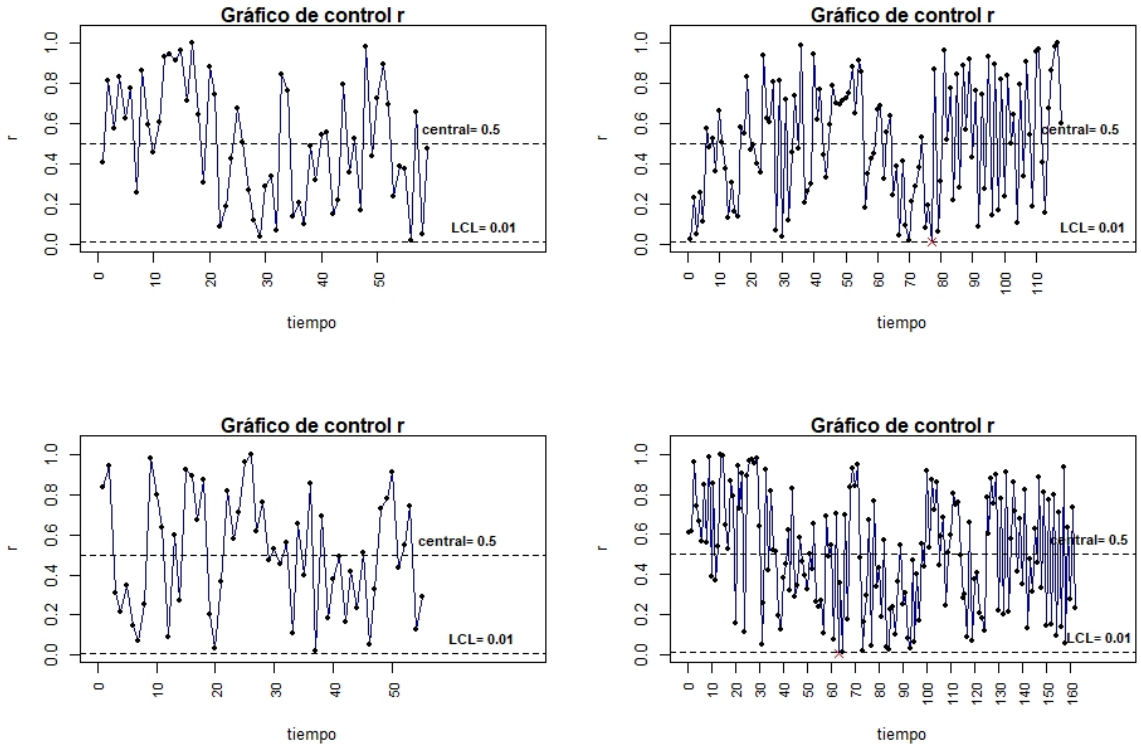


Figura 4.66: Gráfico de control multivariante no paramétrico de Liu para grupo uno (panel superior izquierdo), dos (panel superior derecho),tres (panel inferior izquierdo) y cuatro (panel inferior derecho).

En la Tabla 4.16 se puede observar tanto el porcentaje de observaciones fuera de control como el porcentaje de falsas alarmas para cada uno de los grupos empleando los gráficos de control multivariantes no paramétricos.

Tabla 4.16: Tabla resumen de los gráficos de control multivariantes con agrupación de datos

Grupo Cluster	% Fuera de control	% Falsas alarmas
Gr1 (enero)	13.56	6.45
Gr2 (febrero y marzo)	29.66	22.88
Gr3 (abril)	25.45	21
Gr4 (mayo, junio y julio)	14.20	10.86
General	19.10	12.73

En dicha tabla, se observan cómo los porcentajes de falsas alarmas son muy elevados, lo que indica que se detectan muchas observaciones fuera de control que realmente no lo son.

4.3.4. Gráficos de control multivariantes paramétricos con agrupación en función de los estados de funcionamiento

En este sistema de enfriamiento, el consumo mayoritario se produce por los chiller, que en este caso existen dos que pueden funcionar en función de las necesidades del sistema de cuatro maneras distintas, donde por lo tanto conviene proceder a diferenciar 4 estados de funcionamiento:

1. Estado 1: Chiller 1 encendido y chiller 2 encendido: En este caso, las salidas de ambos chiller rondan los 11ºC.

2. Estado 2: Chiller 1 apagado y chiller 2 apagado: En este caso, las salidas de ambos chiller rondan los 20ºC.

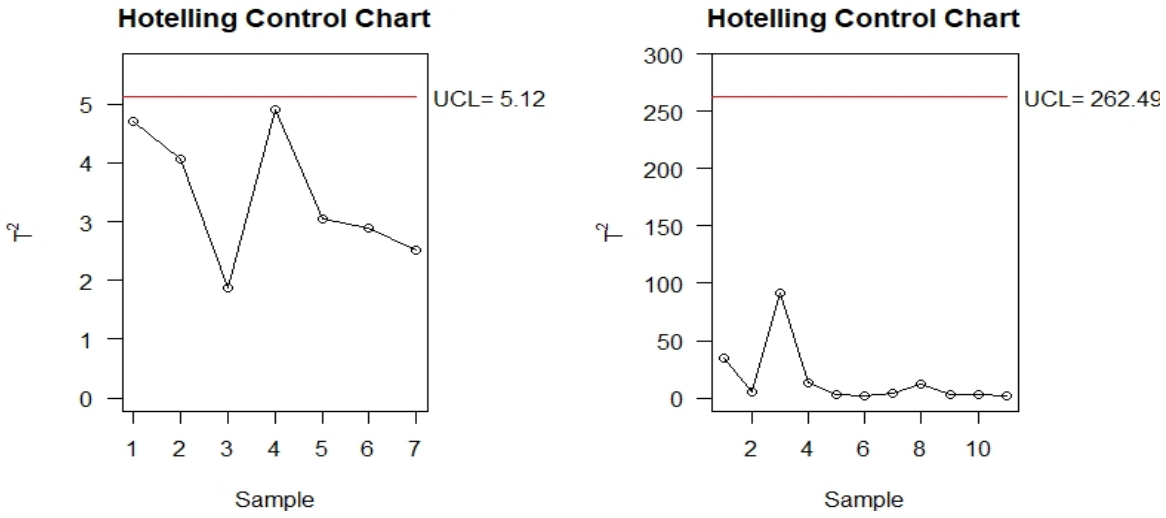
3. Estado 3: Chiller 1 encendido y chiller 2 apagado: En este caso, la temperatura del chiller encendido, el chiller 1, ronda los 12ºC, y la temperatura de salida del chiller apagado, el chiller 2, ronda los 14ºC.

4. Estado 4: Chiller 1 apagado y chiller 2 encendido: En este caso, la temperatura de salida del chiller apagado, el chiller 1,ronda los 14ºC, mientras que el chiller encendido, el chiller 2, ronda los 12ºC.

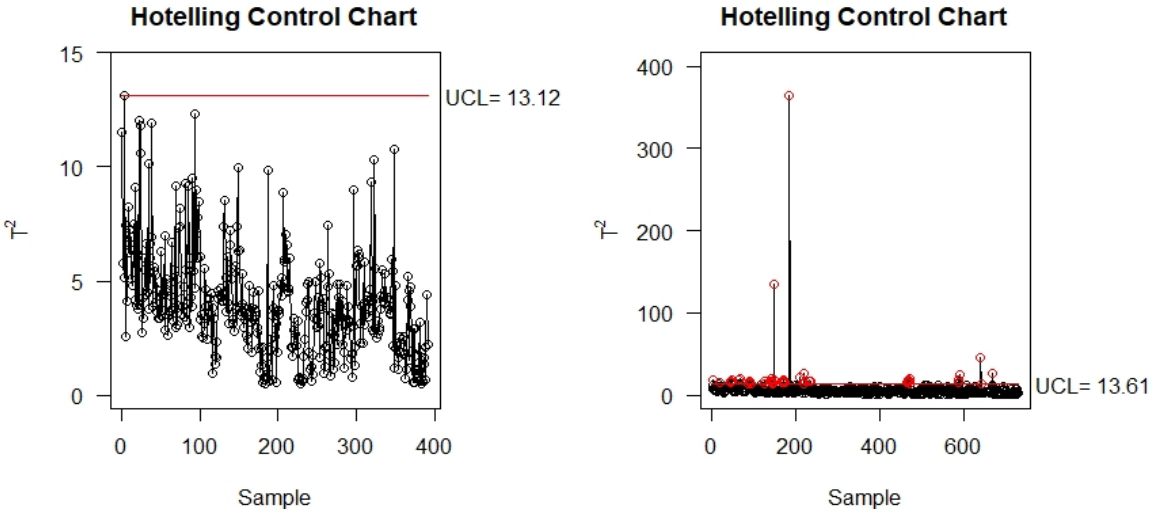
Al fin y al cabo, los gráficos de control pretenden determinar los límites de control en función de la variabilidad de los datos, por lo que es de gran interés diferenciar los cuatro estados de funcionamiento, precisamente porque la variabilidad de los valores representativos de la producción es muy diferente dependiendo del estado del equipo.

Por lo tanto, se plantea un estudio de manera que se diferencian los cuatro diferentes estados de funcionamiento de los chiller, generando el gráfico de control T^2 de Hotelling correspondiente a cada mes y estado. Nótese que en algunas ocasiones no existen suficientes datos, o directamente en algunos meses no se dan todos los estados de funcionamiento, por lo que no se contemplan dichos estudios.

Comenzando por el mes de enero, se obtienen los resultados mostrados en los gráficos de control de la Figura 4.67. En este caso, se observa que solamente se detectan observaciones fuera de control asociados



(a) Gráfico de control para estado 3 para la fase I (panel izquierdo) y fase II (panel derecho).



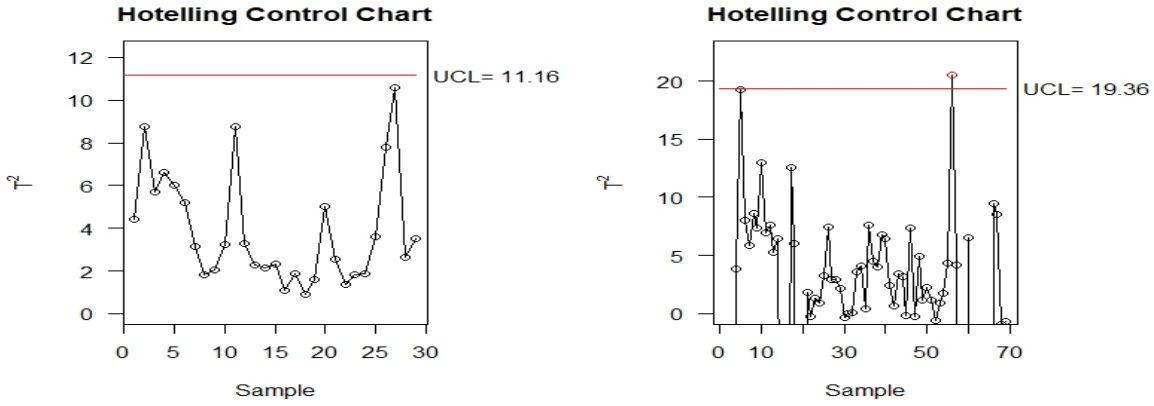
(b) Gráfico de control para estado 4 para la fase I (panel izquierdo) y fase II (panel derecho).

Figura 4.67: Gráficos de control T^2 de Hotelling para el mes de enero.

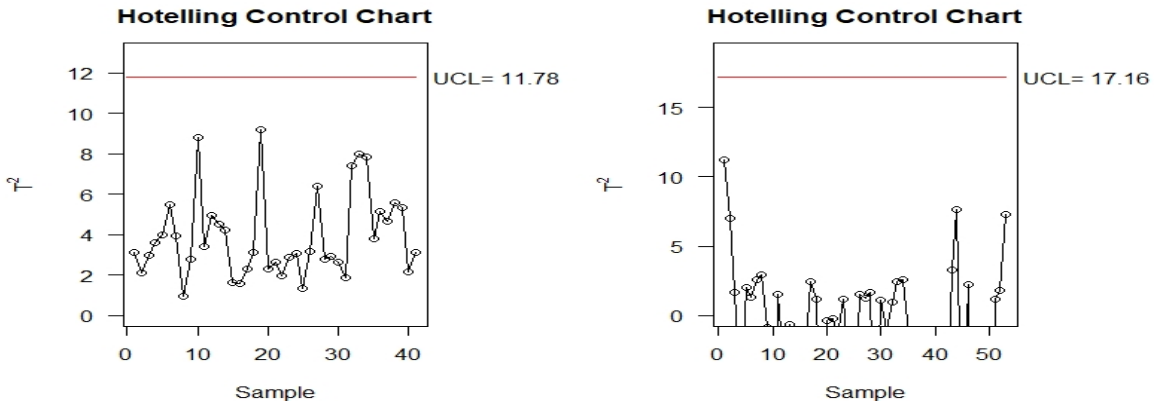
al estado cuatro del mes de enero, considerando días atípicos los días 1,2,3,4,6,7,8,10,11,21,26,28 y 29. Esto implica un total de 8.47 % de observaciones detectadas como anomalías, coincidiendo con el criterio del hotel en las fechas 7,8 y 21 de enero, y discrepando en los días 9 y 22, que el hotel sí que establece estas fechas como anomalías del sistema. Sin embargo, en este mes las anomalías responden a dos causas asignables, las del rango de los días 7,8 y 9, y por otro lado los días 21 y 22, que corresponden a cambios provocados por el personal del hotel. Ambas causas asignables quedarían detectadas por los gráficos de control, ya que a pesar de no detectar toda la franja horaria, sí que detecta algunas de ellas.

En cuanto a febrero, el porcentaje de observaciones consideradas como fuera de control se eleva notablemente al 29.61 %, y todas ellas corresponden al estado de funcionamiento cuarto, es decir, al estado donde solamente se encuentra encendido el chiller 2, como muestra la Figura 4.68. Las fechas correspondientes a las observaciones detectadas por los gráficos de control de febrero son los días que

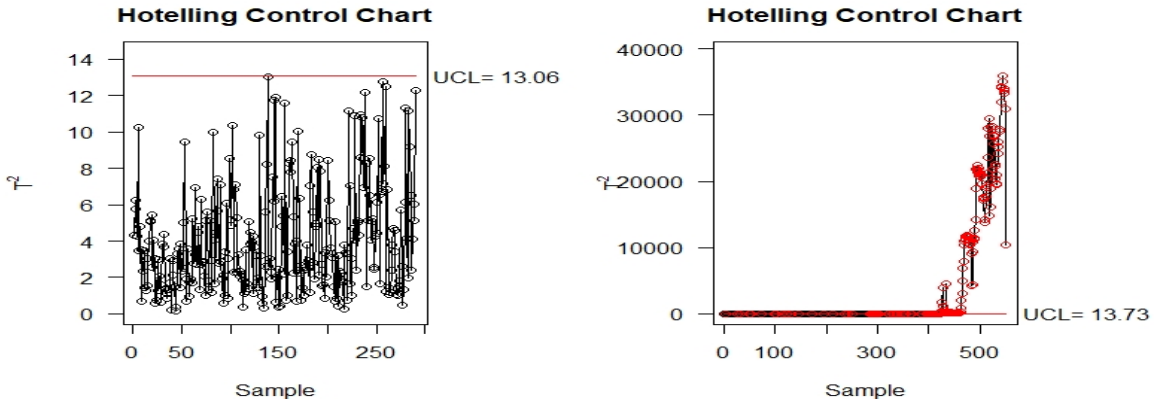
van del 1 al 19 y del 21 al 26. Coincidiendo con las anomalías registradas por el hotel en gran parte, aunque los gráficos de control no consideran el 20 de febrero como anomalía. Por lo que a marzo



(a) Gráfico de control para estado 1 para la fase I (panel izquierdo) y fase II (panel derecho).



(b) Gráfico de control para estado 3 para la fase I (panel izquierdo) y fase II (panel derecho).

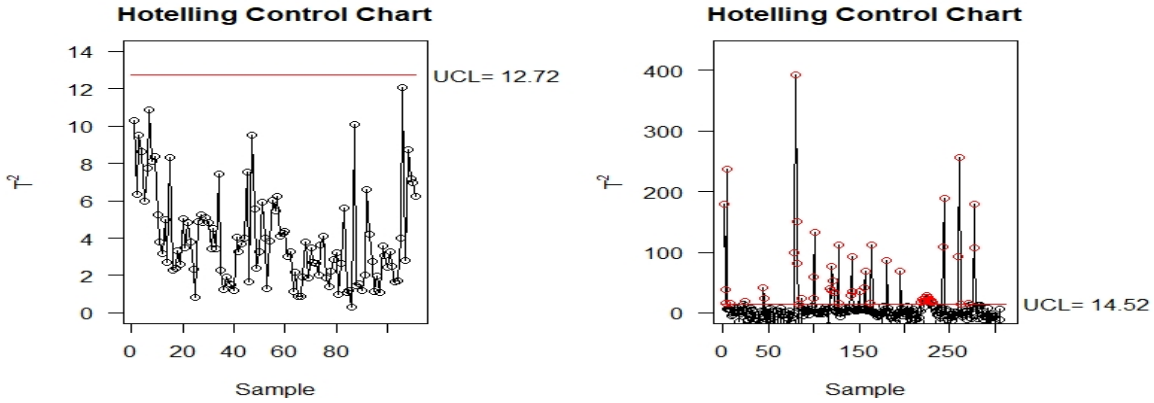


(c) Gráfico de control para estado 4 para la fase I (panel izquierdo) y fase II (panel derecho).

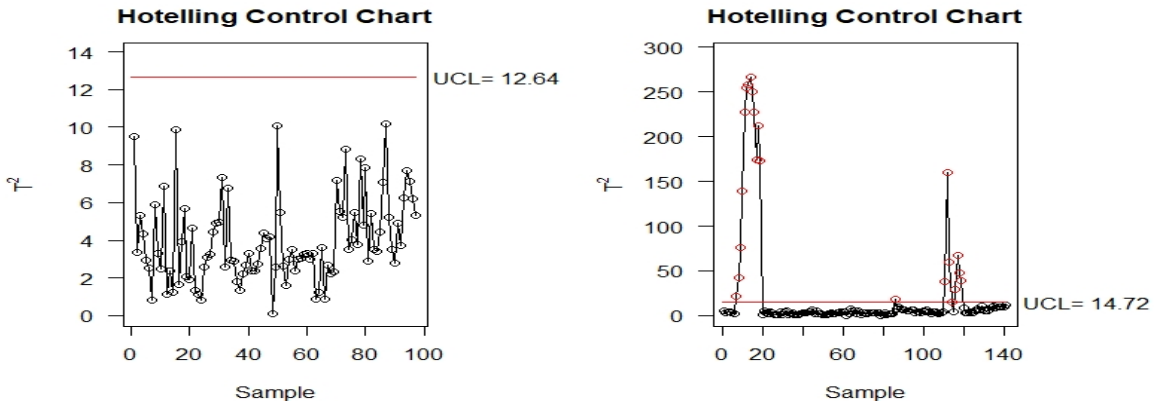
Figura 4.68: Gráficos de control T^2 de Hotelling para el mes de febrero.

respecta, el hotel no considera ninguna anomalía, pero sin embargo, como se observa en la Figura 4.69, los gráficos de control muestran que existe casi un 21 % de observaciones fuera de control. En este caso, se reparten de manera bastante uniforme en todos los estados de funcionamiento. De nuevo, esta

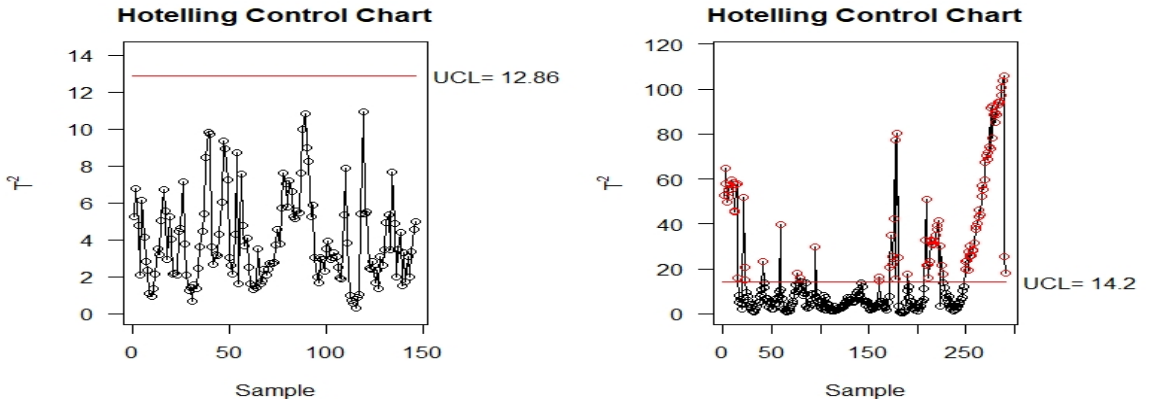
cantidad de falsas alarmas podrían no serlo, ya que los gráficos de control podrían estar detectando la anomalía del fallo del sensor de temperatura a lo largo de este mes, ya que el hotel no es conocedor del error de emisión de datos hasta el 29 de marzo. En cuanto al mes de abril, el estado uno de



(a) Gráfico de control para estado 1 para la fase I (panel izquierdo) y fase II (panel derecho).



(b) Gráfico de control para estado 3 para la fase I (panel izquierdo) y fase II (panel derecho).

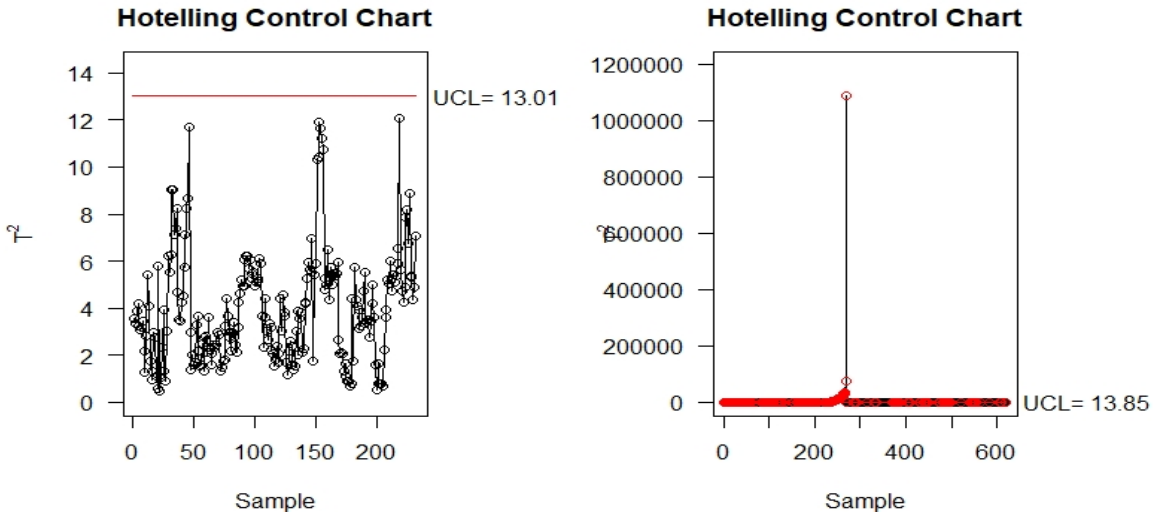


(c) Gráfico de control para estado 4 para la fase I (panel izquierdo) y fase II (panel derecho).

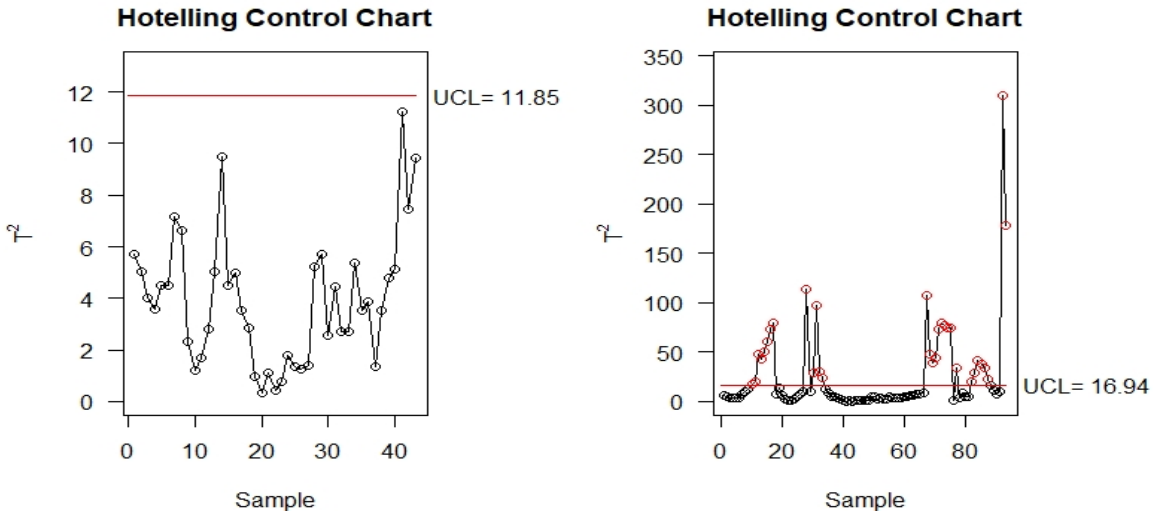
Figura 4.69: Gráficos de control T^2 de Hotelling para el mes de marzo.

funcionamiento es el que más parece salirse de control, es decir, el estado de funcionamiento donde ambos chiller funcionan de manera simultánea. El porcentaje de observaciones fuera de control es

realmente elevado, superando el 38 %. De todas formas, el gráfico de control de la Figura 4.70 detecta los días 14,15 y 16 de abril, coincidiendo con el criterio del hotel. Además, el hotel también reconoce que el sistema no vuelve a funcionar con normalidad hasta el día 8 de abril, fecha hasta la cual no se estabiliza el funcionamiento del sistema debido a la sustitución del sensor que ya provocaba descontrol en la recolección de los datos el mes anterior. En el mes de mayo, los gráficos de control detectan



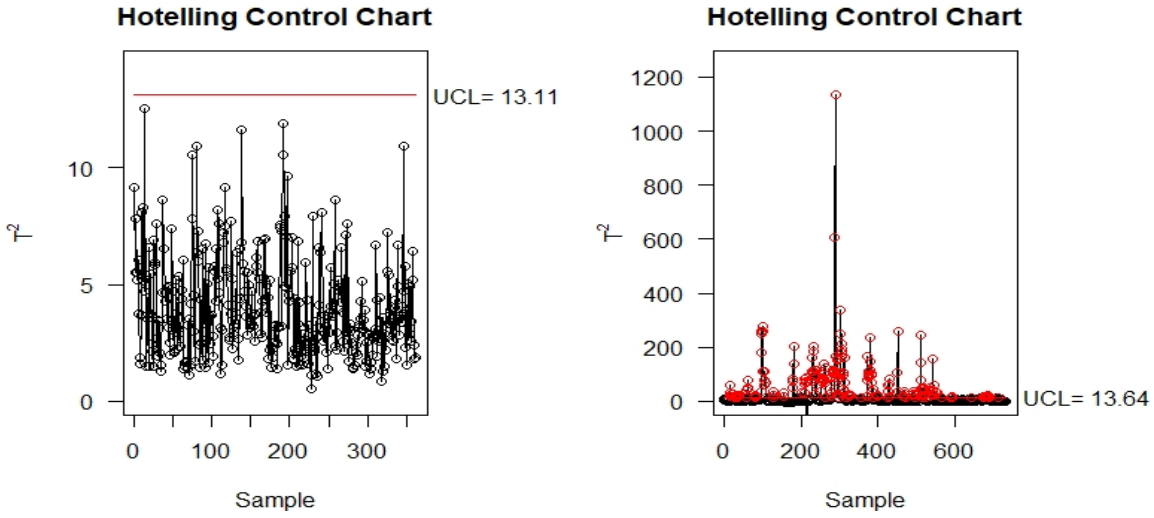
(a) Gráfico de control para estado 1 para la fase I (panel izquierdo) y fase II (panel derecho).



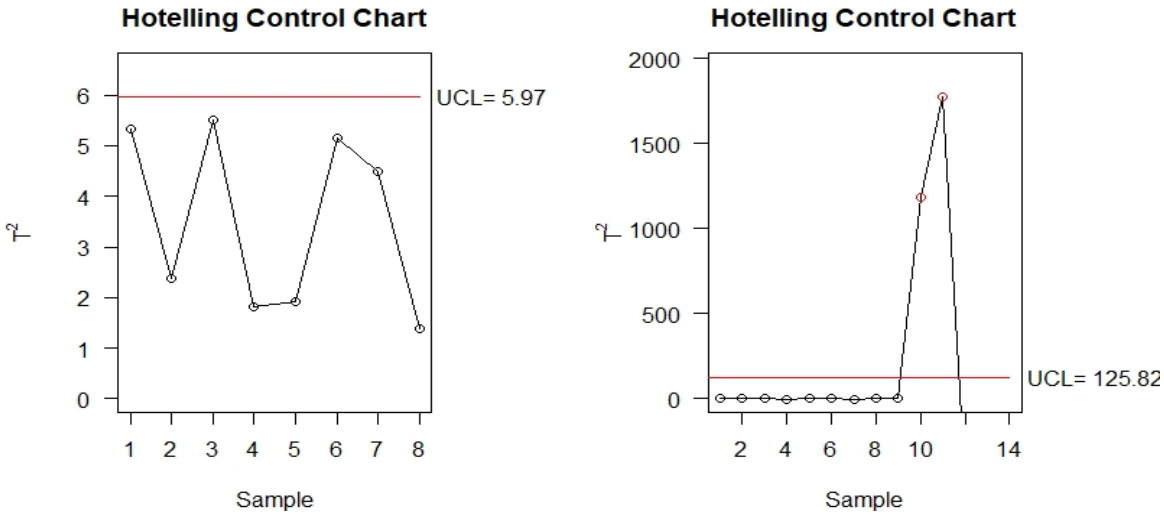
(b) Gráfico de control para estado 4 para la fase I (panel izquierdo) y fase II (panel derecho).

Figura 4.70: Gráficos de control T^2 de Hotelling para el mes de abril.

el día 13 de este mes entre muchos otros, lo que es coincidente con el criterio del hotel, ya que en esta fecha ocurre un apagado general. Sin embargo, los gráficos de control parecen dar lugar a un elevado número de falsas alarmas, con un porcentaje del 28 % de observaciones fuera de control, como muestra la Figura 4.71. El hotel no considera a priori que existan anomalías en todas ellas, sin embargo, deberían estudiarse detenidamente, ya que este mes es un mes de un gran número de días festivos, lo que implica un incremento de huéspedes y actividad general en el hotel, o lo que es lo mismo, gran variabilidad de ciertos parámetros. En el mes de junio, vuelven a darse numerosas observaciones fuera



(a) Gráfico de control para estado 1 para la fase I (panel izquierdo) y fase II (panel derecho).



(b) Gráfico de control para estado 4 para la fase I (panel izquierdo) y fase II (panel derecho).

Figura 4.71: Gráficos de control T^2 de Hotelling para el mes de mayo.

de control, véase la Figura 4.72. Nótese que en este caso solamente existen observaciones suficientes para el estado uno de funcionamiento, es decir, que ambos chiller funcionan simultáneamente en la mayor parte de este mes. El 5 de junio el hotel realiza un apagado general del sistema por temas de mantenimiento, fecha que es detectada por los gráficos de control empleados, sin embargo, en este mes las falsas alarmas alcanzan casi un 29 % de las observaciones totales. Por último, el mes de julio vuelve a registrar mayoritariamente un funcionamiento de acuerdo al estado uno, con los dos chiller funcionando a la vez. Este mes, junto con el anterior es un mes donde la ocupación es baja, y con ello también las actividades del hotel. Se encuentran observaciones puntuales fuera de control en los días 1,2,3,4,6,7,9,10,15,16,17,23,24,28 y 29, como se observa en la Figura 4.73. En este caso, como el hotel no registra anomalías es posible considerar todas las observaciones detectadas por los gráficos de control como falsas alarmas, siendo estas el 8.47 % de las observaciones de julio. Como se puede apreciar, este planteamiento detecta en gran parte todas las anomalías registradas por el hotel, de

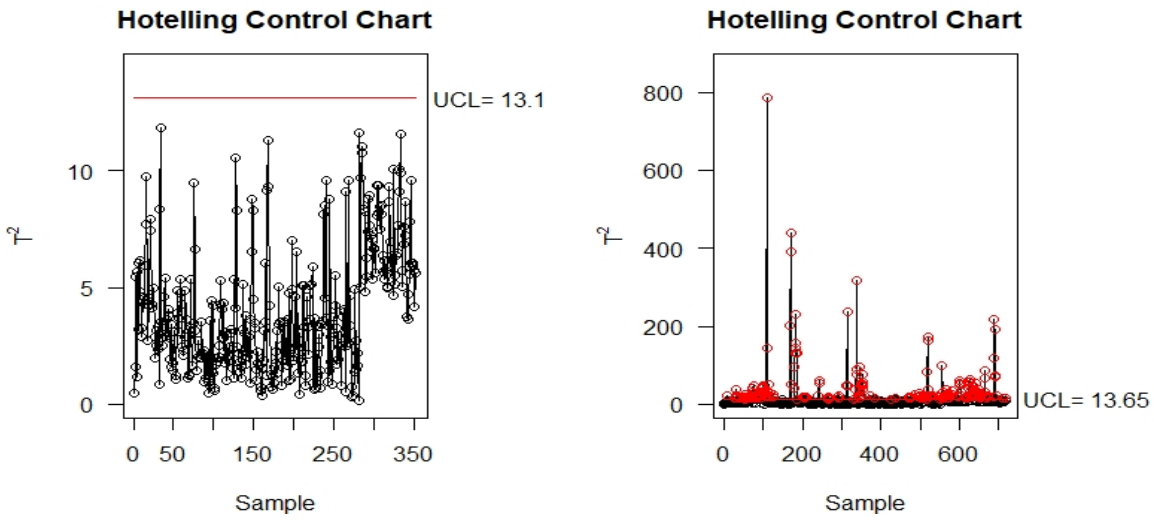


Figura 4.72: Gráficos de control T^2 de Hotelling para el mes de junio con funcionamiento según estado 1 para la fase I (panel izquierdo) y fase II (panel derecho).

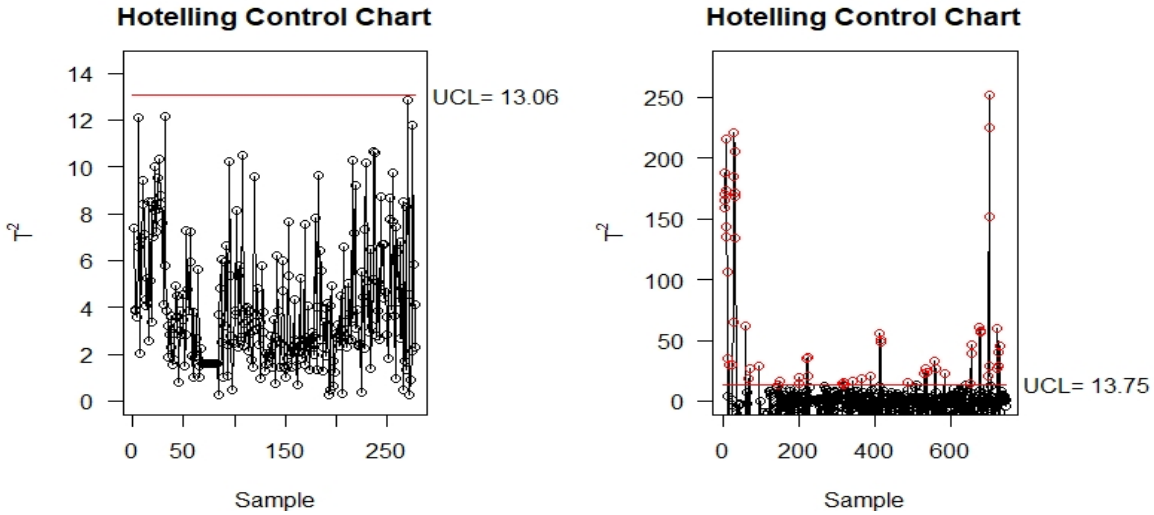


Figura 4.73: Gráficos de control T^2 de Hotelling para el mes de julio con funcionamiento según estado 1 para la fase I (panel izquierdo) y fase II (panel derecho).

hecho, coinciden en casi un 90 %, pero sin embargo, el porcentaje de observaciones detectadas como fuera de control por los gráficos de control T^2 de Hotelling es muy elevado, siendo del 23.52 % de las observaciones totales.

A continuación, en la Tabla 4.17 es posible observar el resumen por mes empleando esta diferenciación por los modos de funcionamiento de los chiller, pudiendo observar que los porcentajes de falsas alarmas son elevados, siendo en ocasiones igual al porcentaje de observaciones fuera de control detectados.

Tabla 4.17: Tabla resumen de los gráficos de control multivariantes con agrupación por estados de funcionamiento

Grupo	% Fuera de control	% Falsas alarmas
Enero	8.47	5.24
Febrero	29.61	27.82
Marzo	21	21
Abril	38.21	31.11
Mayo	27.95	25.26
Junio	29	27.08
Julio	8.47	8.47
General	23.52	20.57

4.3.5. Gráficos de control multivariantes para datos autocorrelados

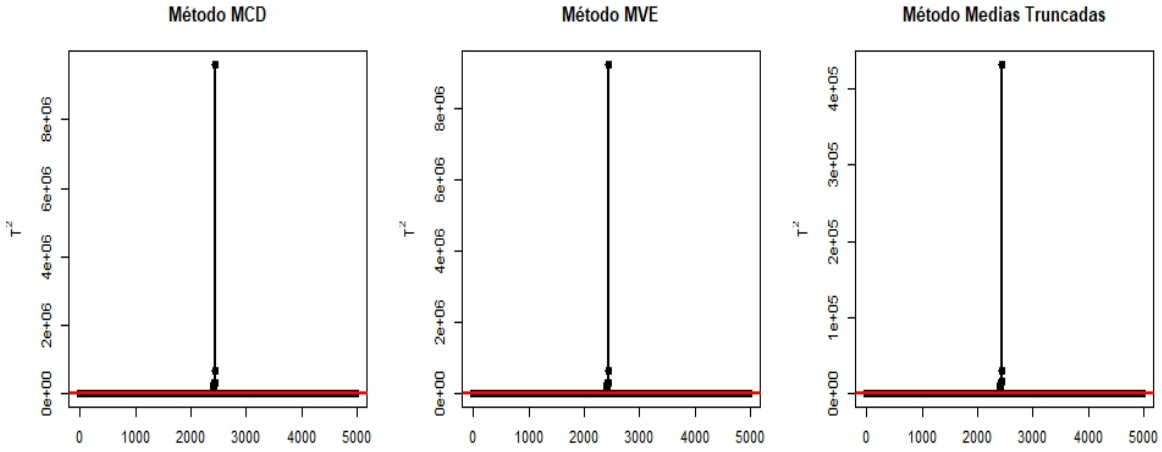
Como ya se adelantaba, los datos del estudio están altamente correlados, por lo que parece adecuado plantear un análisis de la calidad del sistema de enfriamiento mediante métodos de gráficos de control multivariantes para datos autocorrelados. Dentro de este contexto, se plantean dos enfoques al mismo tiempo.

Por un lado, se van a aplicar los métodos de estimación robustos a todos los datos en conjunto, es decir se emplearán los estimadores MCD, MVE y medias truncadas (Anexo C) al conjunto total de observaciones horarias contenidas entre las fechas 1 de enero y 31 de julio. A continuación, y teniendo en cuenta los cuatro estados de funcionamiento de los chiller, se planteará el método de medias truncadas para cada uno de los meses, diferenciando las observaciones de acuerdo al estado de funcionamiento de los chiller en dichos periodos de tiempo.

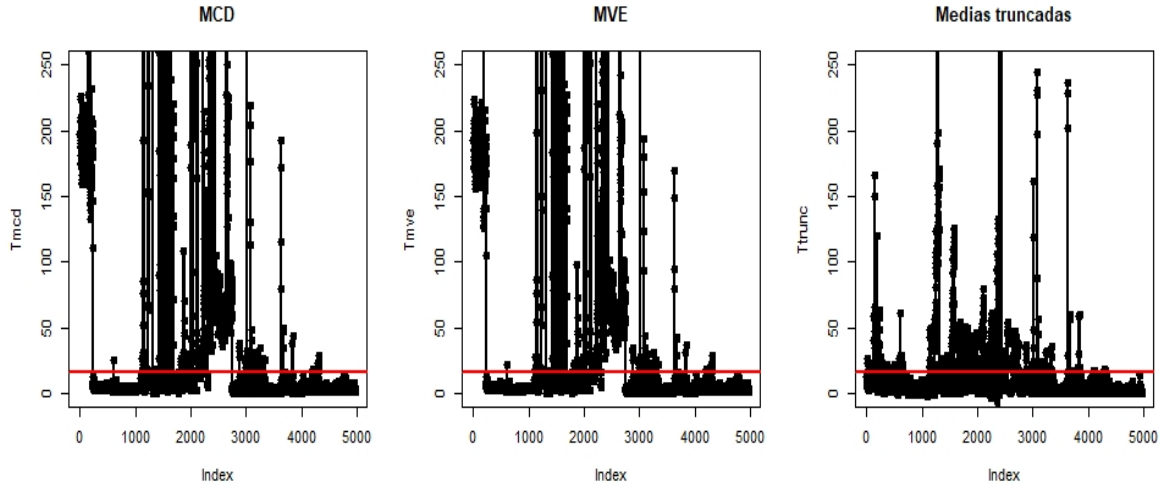
Primeramente, para el caso más general, incluyendo todas las observaciones sin diferenciar los estados de funcionamiento, se obtienen los gráficos de control representados en la Figura 4.74a. Debido a la gran magnitud del valor del estimador robusto en ciertas fechas, se reduce la estala de estos gráficos para observar mejor cómo se distribuyen respecto al límite de control obtenido, véase la Figura 4.74b. A continuación, se muestran las Tablas 4.18, 4.19 y 4.20, donde se muestran las listas de las fechas detectadas por los gráficos de control para datos autocorrelados en cada uno de los métodos robustos, MCD,MVE Y medias truncadas respectivamente. Empleando el método robusto para datos autocorrelados MCD, se obtiene un total del 27 % de observaciones fuera de control, detectando el 68.4 % de fechas asociadas a las anomalías registradas por el hotel.

En cuanto al método del estimador robusto MVE, se obtiene un total del 26 % de observaciones fuera de control, detectando el 68.4 % de fechas asociadas a las anomalías registradas por el hotel. Se observa que ambos métodos registran fechas muy similares.

Finalmente, el método de medias truncadas, presenta un porcentaje de observaciones fuera de control inferior, del 14.9 %, lo que indica que esta última variante de los estimadores robustos, conlleva a una reducción de falsas alarmas. En este caso, el porcentaje de fechas asociadas a las anomalías registradas por el hotel también se ven reducidas, pero no en gran medida, siendo del 63.2 %, lo que teniendo en cuenta ambos porcentajes, puede resultar interesante.



(a) Gráfico de control mediante estimadores robustos (MCD,MVE y medias truncadas, respectivamente).



(b) Gráfico de control mediante estimadores robustos ampliado (MCD,MVE y medias truncadas, respectivamente).

Figura 4.74: Gráficos de control T^2 de Hotelling para el mes de abril.

Como se mencionaba previamente, se va a aplicar el método de estimación robusta de medias truncadas a los datos teniendo en cuenta los modos de funcionamiento. Se escoge aplicar el de medias truncadas ya que presenta un número inferior de falsas alarmas en comparación con los otros dos métodos robustos MCD y MVE, como muestra la Tabla 4.21.

Comenzando por el mes de enero, se observa en la Figura 4.75 que empleando un truncamiento, solamente se identifican como fuera de control observaciones referentes al estado cuatro, donde sólo se encuentra trabajando el chiller 2. Aún así, el porcentaje de observaciones fuera de control es del 19.48 %, entre las que se encuentran por supuesto las anomalías registradas por el hotel, contenidas en los días 7,8,9,21 y 22.

En cuanto al mes de febrero, véase la Figura 4.76, el número de detecciones de observaciones fuera de control se ve reducido al 16.50 % de las observaciones totales del mes, distribuidas tanto para los periodos bajo funcionamiento del estado uno como del cuarto. Entre estas observaciones fuera de

Tabla 4.18: Observaciones fuera de control detectadas por el estimador robusto MCD.

Mes	**Observaciones fuera de control**
Enero	1,2,3,4,5,6,7,8,9,10,26
Febrero	16,17,18,19,20,21,22,23,24,27
Marzo	1,2,3,4,5,6,7,8,9,10,12,16,18,19,20,22,23,24,25,26,27,28,29,30,31
Abril	1,2,3,4,5,6,7,8,9,11,12,13,14,15,16,17,18,19,20,21,22,3,4,25,30
Mayo	1,4,5,6,7,8,9,10,13,16,18,20
Junio	1,3,9,10,25,29,30
Julio	-

Tabla 4.19: Observaciones fuera de control detectadas por el estimador robusto MVE.

Mes	**Observaciones fuera de control**
Enero	1,2,3,4,5,6,7,8,9,10,26
Febrero	16,17,18,19,20,21,22,23,24,27
Marzo	1,2,3,4,5,6,7,8,9,10,12,18,19,20,23,24,25,26,27,28,29,30,31
Abril	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,30
Mayo	1,4,5,6,7,8,9,10,13,16,18,20
Junio	1,3,4,9,10,25,29,30
Julio	-

control se encuentran los días 15,16,17,18,19,20 y 21, que son las que el hotel confirma como anomalías, lo que reduce en gran medida las falsas alarmas asociadas a este mes.

El mes de marzo es un mes en el que el hotel no considera que existan anomalías, y en este caso, los gráficos de control representados en la Figura 4.77 muestran solamente un 1 % de observaciones que superan el límite de control, lo que indica que dicho porcentaje es de falsas alarmas, lo que es muy reducido, y además dichas anomalías están igualmente distribuidas en los diferentes estados de funcionamiento que se dan a lo largo de este mes, siendo los días 9 y 11 en los días bajo funcionamiento con estado uno, 14 y 21 con estado tres y finalmente 11 y 27 bajo el estado cuarto.

Hasta el momento, el mes de abril es uno de los meses en los que los gráficos de control detectan un mayor numero de observaciones fuera de control, y en este caso, empleando métodos apropiados

Tabla 4.20: Observaciones fuera de control detectadas por el estimador robusto de medias truncadas.

Mes	**Observaciones fuera de control**
Enero	1,2,4,6,7,8,9,10,11,12,16,22,25,26,27
Febrero	17,18,19,21,22,23,24,26,27,28
Marzo	1,2,3,6,7,10,11,12,13,14,15,16,17,18,19,20,22,23,24,26,28,29,30,31
Abril	1,2,3,4,5,6,8,9,10,11,12,13,14,15,16,17,20,21,22,23,24,25,30
Mayo	1,4,6,7,8,9,10,13,16,18,19,20
Junio	1,3
Julio	-

Tabla 4.21: Tabla resumen de los gráficos de control multivariantes con agrupación por estimadores robustos

Método	% Fuera de control	% Falsas alarmas
MCD	27.01	23.06
MVE	26.27	21.35
Medias truncadas	14.9	12.53

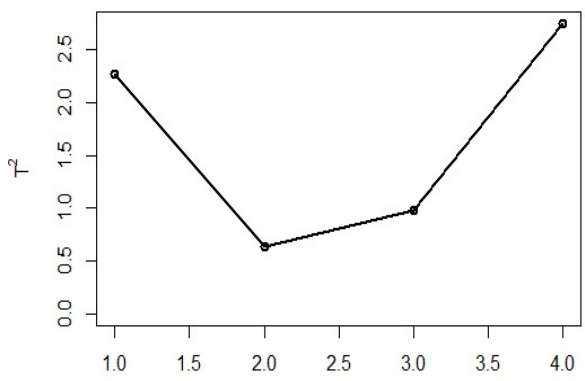
para datos autocorrelados, ese porcentaje se ve reducido a un 6.3 %, como se representa en la Figura 4.78. El estado uno parece contener la gran parte de estas observaciones anómalas, detectando los días 1,4,7,14,15,16,20,21 y 29, en los que se incluyen los días considerados anómalos por el hotel, los días 14,15 y 16.

El hotel registra en el mes de mayo el día 13 como día con presencia de datos anómalos debido a un apagado manual del sistema por temas de mantenimiento. Los gráficos de control para datos autocorrelados detectan este día también, junto con otros días como el 9,10 y 14. En este caso, véase la Figura 4.79, el porcentaje de observaciones que superan el límite de control es reducido, entorno al 1 %, lo que resulta satisfactorio ya que además de que este porcentaje es bajo, es capaz de detectar un dato atípico así considerado también por la dirección del hotel.

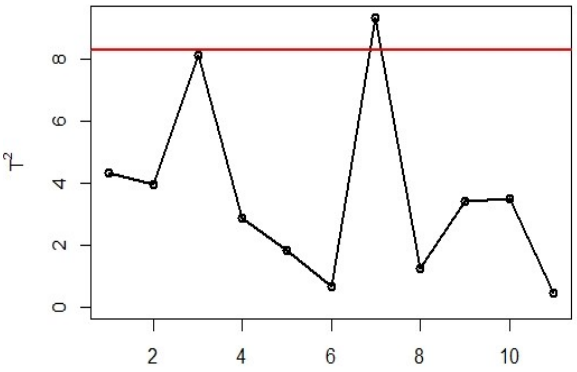
En general, por el momento todos los meses muestran un número de observaciones fuera de control reducido, comportamiento que se mantiene tanto en junio como en julio. En el caso de junio, como muestra la Figura 4.80, el número de anomalías detectadas es inferior al 1 %, pero además, entre ellas se encuentra la asociada al 5 de junio, que es la fecha en la que el hotel realiza el apagado total del sistema.

Finalmente, el mes de julio no presenta según el criterio del hotel anomalías entre sus datos. Como se representa en la Figura 4.81, todo el mes se rige bajo el estado de funcionamiento primero, con los dos chiller en funcionamiento, teniendo como falsas alarmas un 0.4 % de las observaciones totales.

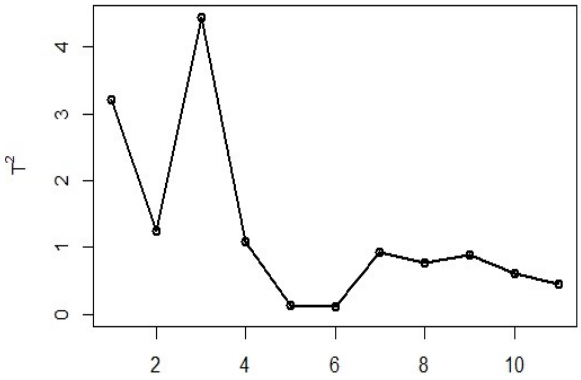
En definitiva, este último método en el que se emplean gráficos de control multivariantes para datos autocorrelados empleando la estimación robusta de medias truncadas es capaz de detectar todas



(a) Gráfico de control de medias truncadas para estado 1.



(b) Gráfico de control de medias truncadas para estado 3.



(c) Gráfico de control de medias truncadas para estado 4.

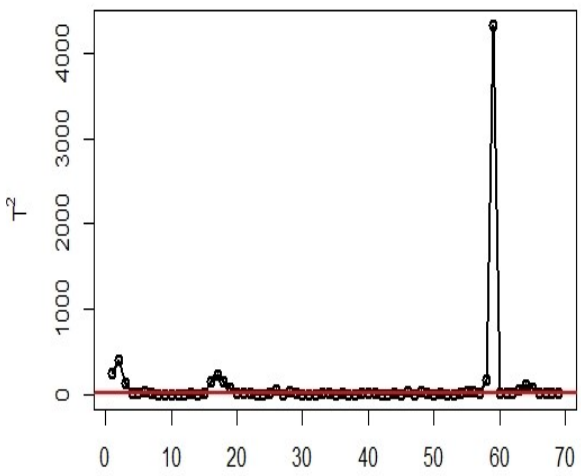
Figura 4.75: Gráficos de control T^2 de Hotelling por el método de medias truncadas para el mes de enero.

y cada unas de las fechas en las que el hotel es conocedor de que se generan modificaciones o errores en el sistema. Estos resultados pueden verse resumidos en la Tabla 4.22. Parece fundamental al mismo tiempo emplear la diferenciación de los datos en función de los estados de funcionamiento, obteniendo así un total del 6.49 % de observaciones totales fuera de control, detectando el 100 % al mismo tiempo de las anomalías registradas por el hotel.

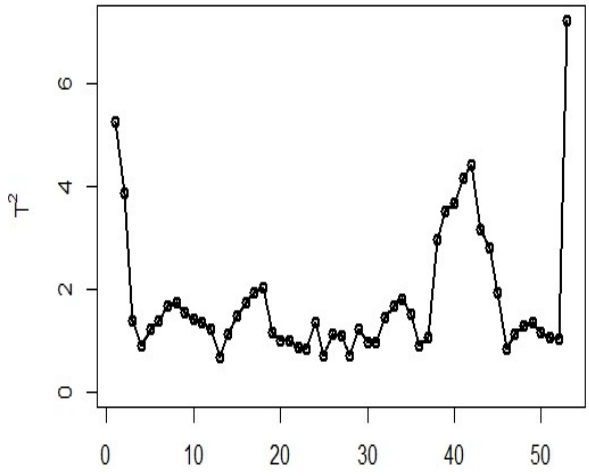
4.4. Capacidad del proceso

Una vez estimados los límites de control natural de las variables críticas para la calidad de los sistemas HVAC del hotel, el siguiente paso es evaluar el grado en el que estas instalaciones cumplen con las especificaciones marcadas por los clientes, la gerencia y/o la normativa. En esto consiste el análisis de capacidad.

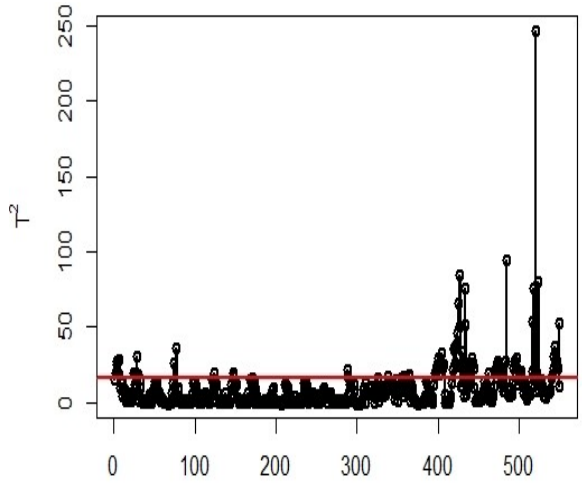
A continuación se presenta la Tabla 4.23, en ella se muestran los índices de capacidad obtenidos mediante los métodos **Taam et al. (1993)** y **Pan y Lee (2010)**, que calculan los índices de capacidad multivariantes calculados para un nivel de significación del 5 %. En referencia a la segunda y tercera columnas, se obtienen los cálculos mediante dichos métodos pero en este caso, aplicando un



(a) Gráfico de control de medias truncadas para estado 1.



(b) Gráfico de control de medias truncadas para estado 3.



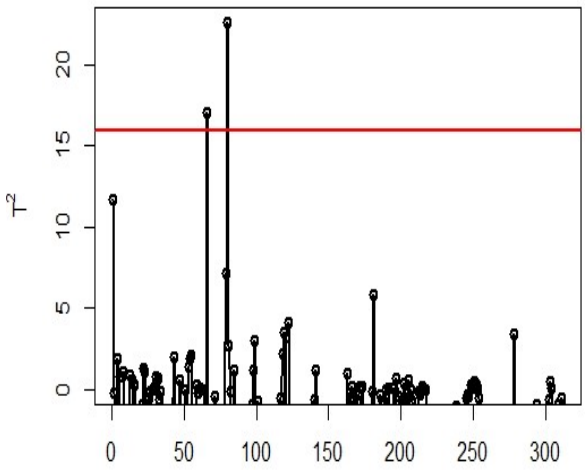
(c) Gráfico de control de medias truncadas para estado 4.

Figura 4.76: Gráficos de control T^2 de Hotelling por el método de medias truncadas para el mes de febrero.

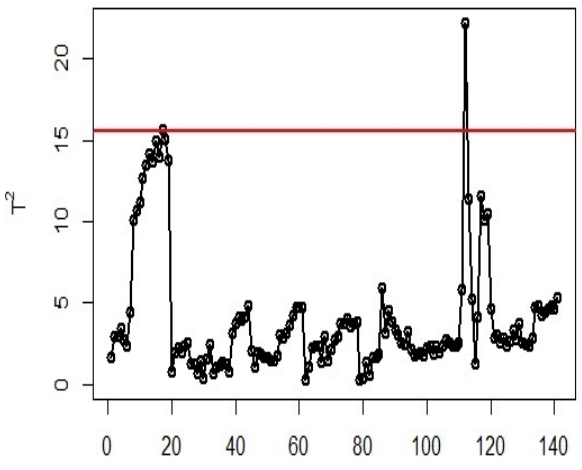
análisis de componentes principales, empleando dos y tres componentes principales respectivamente.

Para el cálculo de los diferentes índices de capacidad es necesario establecer los valores objetivo, así como la definición de los límites superiores e inferiores. En este caso, carece de sentido establecer valores objetivo y de límites para variables como temperatura exterior u ocupación, por lo que solamente se fijan valores objetivo en para las variables de temperatura interior del hotel, y las cuatro temperaturas asociadas a los procesos de entrada y salida de los chiller. En este caso, los valores objetivo para las variables son:

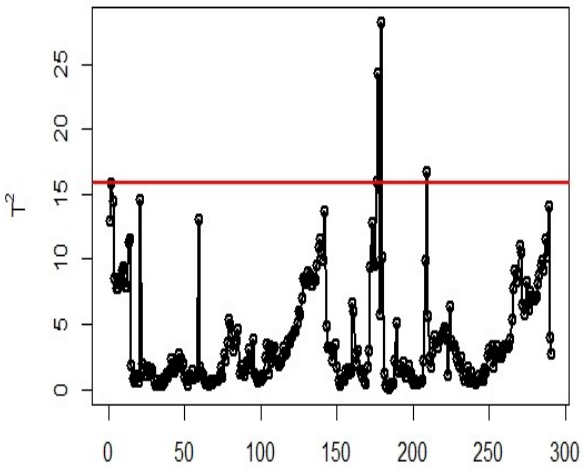
(temp_clima_on, temp_ch1_out, temp_ch1_in, temp_ch2_out, temp_ch2_in)=(22.5, 12.5, 15.5, 12.5, 15.5)



(a) Gráfico de control de medias truncadas para estado 1.



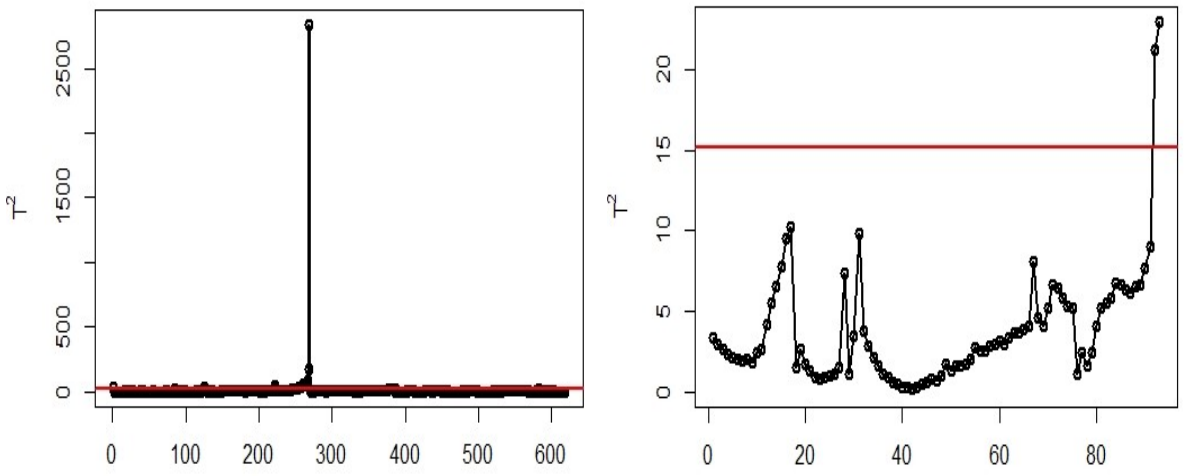
(b) Gráfico de control de medias truncadas para estado 3.



(c) Gráfico de control de medias truncadas para estado 4.

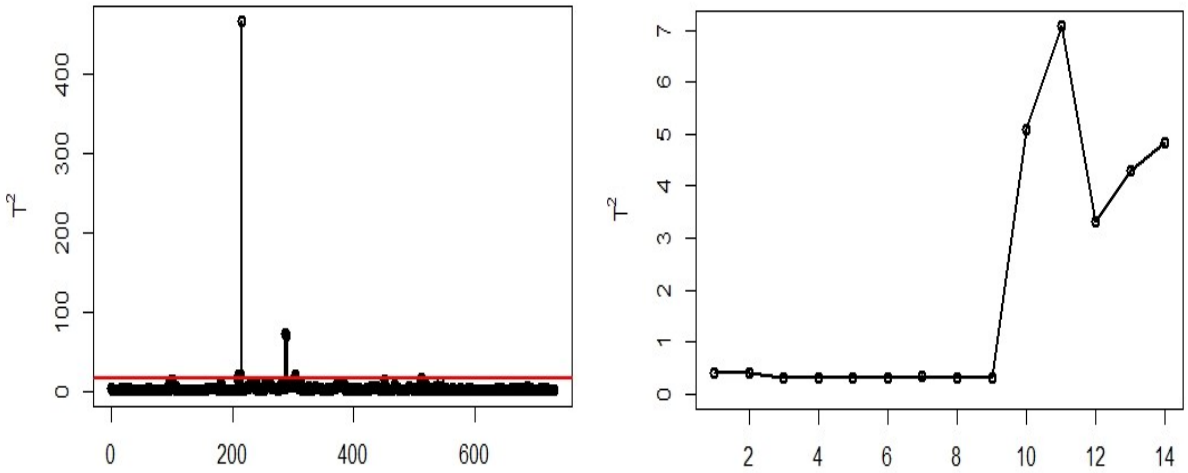
Figura 4.77: Gráficos de control T^2 de Hotelling por el método de medias truncadas para el mes de marzo.

Nótese que el valor objetivo respeta los 3ºC teóricos que se obtienen entre el proceso de entrada y salida de los chiller. De acuerdo a estos valores objetivos, o target, se calculan los valores de los límites superior sumando al valor objetivo la desviación estándar por 1.96, y restando el mismo valor para obtener el cálculo de los límites inferiores. A continuación, en la Figura 4.82 se puede observar la disposición de las regiones asociadas a los métodos anteriores cuando se emplean dos componentes principales. Como se observa en la 4.23, los índices de capacidad empleando componentes principales resultan más elevados, aún cuando las dos componentes principales resumen el 92 % de la variabilidad explicada y con las tres primeras se alcanza el 99 %. Estos índices superan el valor unitario lo que indica que el proceso es capaz. En líneas generales, el índice de capacidad establecido para que el proceso sea



(a) Gráfico de control de medias truncadas para estado 1. *(b)* Gráfico de control de medias truncadas para estado 4.

Figura 4.78: Gráficos de control T^2 de Hotelling por el método de medias truncadas para el mes de abril.

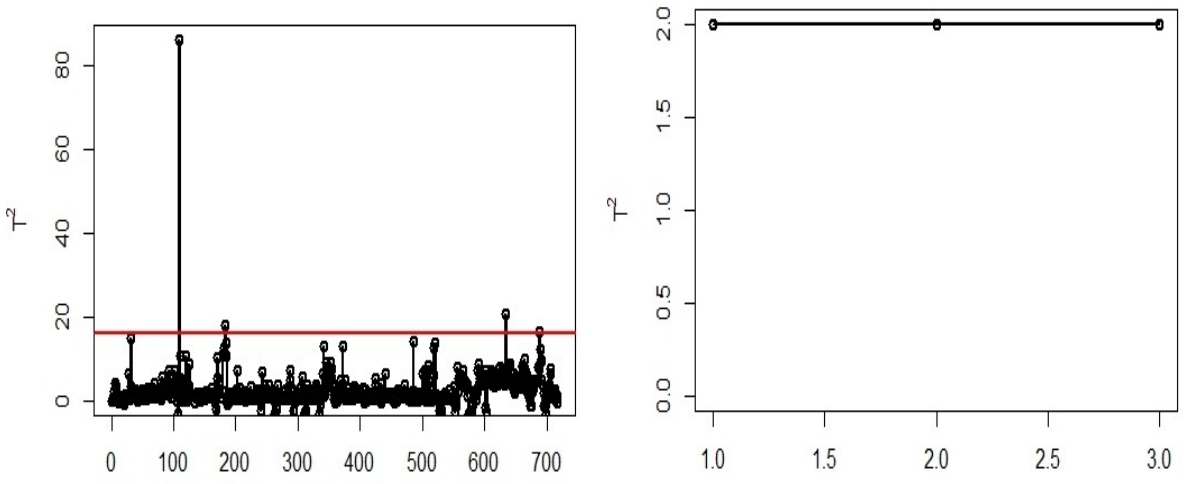


(a) Gráfico de control de medias truncadas para estado 1. *(b)* Gráfico de control de medias truncadas para estado 4.

Figura 4.79: Gráficos de control T^2 de Hotelling por el método de medias truncadas para el mes de mayo.

capaz de cumplir las especificaciones es de 1.33, donde valores cercanos a dicho valor se consideran índices comunes a procesos nuevos y a partir del valor 2 se considera que el proceso tiene calidad Six Sigma. Nótese que a partir de valores de 2.5 la literatura asegura que no existe beneficio alguno la implementación de cambios que traten de mejorarlo.

Resulta interesante calcular adicionalmente el vector de capacidad multivariante, de **Shahriari et al. (2009)** para un nivel de significación del 5 %. El vector obtenido mediante el método clásico queda de la siguiente manera: $(C_pM, PV, LI) = (0{,}6195, 0, 0)$, mientras que empleando el análisis de componentes



(a) Gráfico de control para estado 1.

(b) Gráfico de control para estado 4.

Figura 4.80: Gráficos de control T^2 de Hotelling por el método de medias truncadas para el mes de junio.

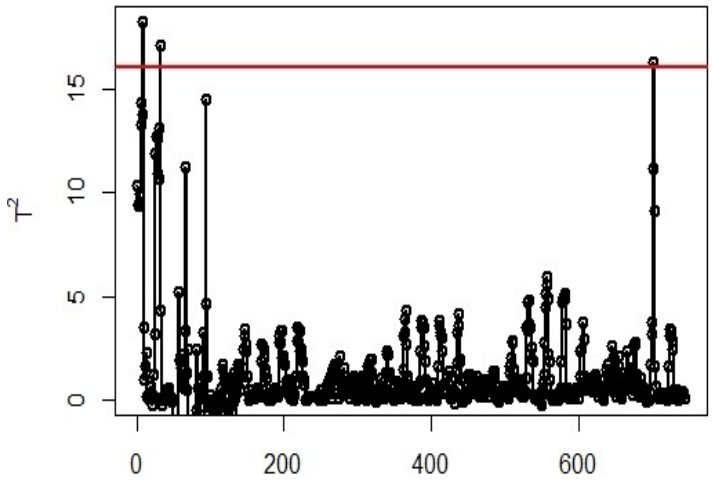


Figura 4.81: Gráficos de control T^2 de Hotelling por el método de medias truncadas para el mes de julio con funcionamiento según estado 1.

principales se obtiene el vector $(C_pM, PV, LI) = (1{,}0754, 1, 1)$ para dos componentes principales y el vector $(C_pM, PV, LI) = (1{,}0639, 1, 1)$ empleando 3.

Cabe destacar que los valores grandes de PV indican la proximidad del centro del proceso al valor objetivo preespecificado. El tercer componente del vector, que se conoce como índice de ubicación (LI), compara la ubicación de la región del proceso modificado con la tolerancia. Este índice tiene valor de uno si toda la región de proceso modificada está contenida dentro de la región de tolerancia indicando que todos los productos manufacturados cumplen con los límites de especificación y de lo contrario,

Tabla 4.22: Resumen de los gráficos de control multivariantes para datos autocorrelados empleando la estimación robusta de medias truncadas

Grupo	% Fuera de control	% Falsas alarmas
Enero	19.48	16.53
Febrero	16.50	8.77
Marzo	1.08	1.08
Abril	6.3	2.26
Mayo	1.08	0.80
Junio	0.97	0.42
Julio	0.40	0.40
General	6.49	3.38

Tabla 4.23: Tabla resumen de la capacidad.

Método	**Índice Clásico**	**Índice PCA(2)**	**Índice PCA(3)**
Taam et al. (1993)	0.1880	1.1565	1.2041
Pan y Lee (2010)	0.0549	1.1565	1.2042

tomará un valor de cero. Los valores son bajos, alcanzando un índice $C_pM = 1{,}06$ en el mejor de los casos, lo que indicaría que el proceso no es capaz. Cabe resaltar que estos índices pueden ser bajos debido a que se trata de un proceso nuevo. El índice propuesto por Taam tienen el inconveniente de la sobreestimación. Los índices de Shariari, por otro lado, tienen el inconveniente de la sobreestimación cuando las características de calidad no son independientes.

Para el caso en el que se emplean 2 componentes principales, es posible graficar las regiones, como muestra la Figura 4.83.

A continuación, en la Tabla 4.24 se muestran los índices de capacidad multivariantes calculados para los datos en cuestión. Los tres métodos que se muestran son basados en el análisis de componentes principales, donde se emplean las dos primeras ya que obtienen una variabilidad explicada suficientemente elevada para superar el 95 % de la misma. Los índices multivariantes se construyen a partir de los métodos presentados por **Wang y Chen (1998)**, **Xekalaki y Perakis (2002)** y **Wang et al. (2005)**, calculados para un nivel de significación del 5 %.

Como se observa en la Tabla 4.24 resumida de los índices multivariantes, la mayoría de los índices resultan bajos. Estos bajos índices pueden deberse a que el proceso contiene numerosas observaciones fuera de control, y que por supuesto el proceso pertenece a un hotel que acaba de implementar esta sistema de enfriamiento, lo que lo hace un proceso nuevo. Los índices inferiores a 0.67 indicaran que el proceso no es adecuado para el trabajo. Sin embargo, valores entre 0.67 y 1.33 estarían sugiriendo la implementación de mejoras en el proceso, o se refieren a procesos nuevos, según numerosos autores

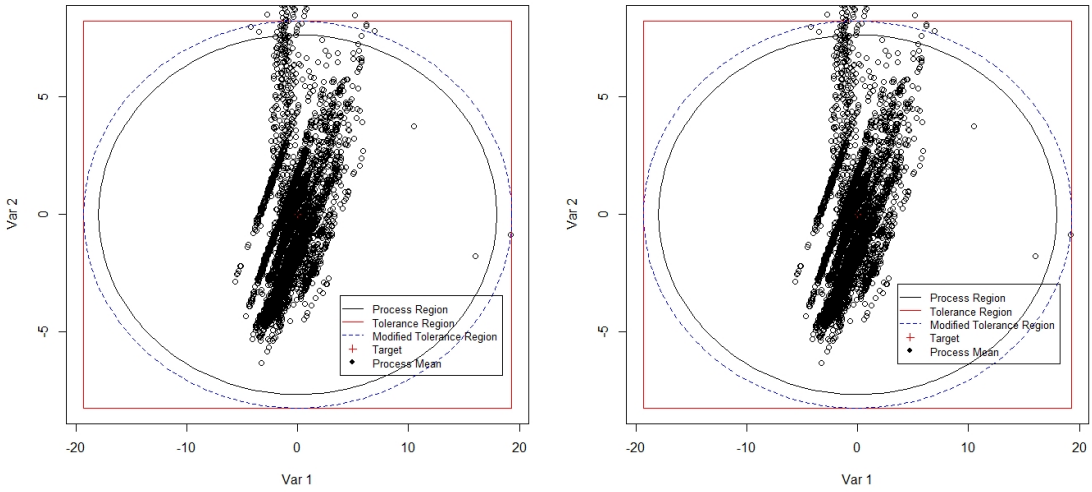


Figura 4.82: Gráfica de capacidad basado en el método de Taam y Pan con 2 componentes principales.

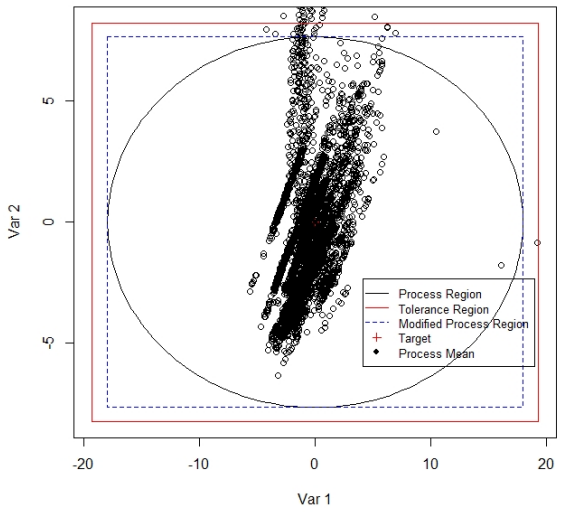


Figura 4.83: Gráfica de capacidad basado en el método de Shahriari con 2 componentes principales.

(véase **Albert P.B.**, **(2004)**). Por lo tanto, el bajo valor de los índices calculados indica que las instalaciones de climatización o HVAC del hotel no cumplen con las especificaciones u objetivos fijados por la empresa y la normativa, por lo que se debería aplicar un plan de mejora para futuros estudios.

4.5. Conclusiones de la implementación

Como se ha visto en el análisis de los datos, existen correlaciones lineales entre las variables. El consumo de la potencia activa muestra alta correlación lineal positiva con las temperaturas del exterior e interior del hotel, con un valor de 62.8 % y 55.7 % respectivamente. Es decir, a mayores temperaturas,

Tabla 4.24: Tabla de resultados de cálculo de los índices de capacidad multivariantes.

Método	**MCp**	**MCpk**	**MCpm**	**MCpmk**
Wang y Chen (1998)	0.6869	0.6365	0.6777	0.6279
Xekalaki y Perakis (2002)	0.6738	0.6386	0.6692	0.6344
Wang (2005)	0.6737	0.6386	0.6692	0.6343

mayor consumo del sistema. Por el contrario, el consumo muestra una relación negativa, con menos fuerza que las relaciones positivas, con el porcentaje de tiempo en el que el sistema se encuentra activo y la temperatura de salida del chiller 1. Es decir, que a mayor tiempo activo del sistema y a menor temperatura del chiller 1, mayor consumo. Son resultados lógicos, ya que si le tiempo de funcionamiento es menor, la potencia será mayor debido al sobreesfuerzo que sufrirá el sistema para obtener la temperatura interior del hotel deseada. Lo mismo ocurre con la temperatura de salida del chiller 1, que requiere de un mayor esfuerzo por parte el sistema para enfriar el fluido en cuestión en el chiller 1, lo que se traduce en un mayor consumo. La relación con la temperatura del chiller 2 también es una correlación lineal negativa, pero sin embargo esta ronda el 4 %, que es muy inferior en comparación con la del chiller 1, que se encuentra entorno al 26 %.

Al mismo tiempo, la temperatura interior del hotel se encuentra altamente correlacionada con las variables referentes a la temperatura exterior, la temperatura del proceso de entrada al chiller 1, la temperatura del proceso de salida del chiller 2 y la potencia del ventilador 1 de la torre de enfriamiento, con valores de la correlación lineal de 60.2 %, 38.9 % 55.9 % y 26.4 %. Como ya desvelaban los gráficos del análisis exploratorio, las temperaturas del interior del hotel eran altas cuando las temperaturas exteriores se incrementaban, y viceversa. Como consecuencia, que las temperaturas del hotel sean elevadas, conllevan lógicamente a que las temperaturas de entrada al chiller 1 también lo sean, como también ocurre con la potencia consumida por el ventilador 1 de la torre de enfriamiento, que requiere de un mayor consumo energético para poder disipar dicho calor excedente. Algo menos común ocurre con el chiller 2, que resulta que su temperatura de salida se incrementa a medida que lo hace la temperatura interior del hotel, lo que puede denotar que el chiller 2 no está funcionando con total eficiencia, ya que la temperatura del proceso de salida del chiller 1 sí que obtiene una correlación negativa con esta temperatura del hotel. También se muestra una alta correlación lineal entre la potencia de ambos ventiladores de la torre de enfriamiento, lo que denota que funcionan de manera similar y simultánea.

Para comprender cómo afecta la ocupación del hotel, calculando las correlaciones lineales de esta variable se obtienen valores de correlación elevados positivos con el porcentaje de tiempo en el que el sistema se encuentra activo, 42.1 %. Sn embargo, con la temperatura exterior del hotel la ocupación se encuentra correlacionada negativamente. En otras palabras, el hotel tiene mayor número de clientes cuando la temperatura exterior es inferior, debido al clima tropical de la zona donde se encuentra, donde las temperaturas más bajas se dan en mayo, junio y julio, que es donde mayor turismo existe en dicho país, elevando el porcentaje de ocupación del hotel. A su vez, la ocupación del hotel se incrementa al mismo tiempo que lo hace el porcentaje de tiempo activo del sistema, es decir cuanta más gente hay hospedada en el hotel, el climatizador más tiempo trabaja, pero esto no implica un aumento en el consumo general que como se ha mencionado previamente, el consumo general del sistema y la ocupación se encuentran relacionadas negativamente.

Si se atiende a los valores de correlación lineal existente en aquellas variables referentes a los chiller, se encuentra que a grandes rasgos el consumo general y el consumo de los chiller lógicamente están altamente correlacionadas. Por su lado, si el sistema funciona correctamente, cabe esperar que las temperaturas del proceso de entrada de los chiller respecto de cada una de las temperaturas de sus

procesos de salida están altamente correlacionadas, y así lo demuestran los coeficientes de correlación lineal, siendo superior al 65 % en el caso del chiller 1. Se obtiene una correlación positiva del 58.4 % entre las temperaturas de entrada al chiller 1 y las de salida del chiller 2, algo que no es del todo lógico si ambos chiller funcionan de manera adecuada. Este fenómeno implica que cuando las temperaturas de entrada de los chiller son elevadas, el chiller 1 es capaz de hacerla disminuir a su salida, lo que no ocurre en el chiller 2 con la misma eficiencia.

Por último, el estudio de correlaciones demuestra que existe una correlación lineal negativa del 33 % entre la temperatura de salida del chiller 1 y la potencia requerida por el ventilador 2 de la torre de enfriamiento, aunque también tiene esta relación negativa con el ventilador 1, con un valor del 19 %. Pero lógicamente de alguna manera tiene que disipar el calor este chiller, por lo que seguramente está dándose la situación en la que cuando las temperaturas de salida del chiller 1 son elevadas, es porque este no se encuentra en funcionamiento ya que se ha visto que es el chiller más eficiente, y por lo tanto es el chiller 2 el que lo está, que en este caso disipa en ambos ventiladores el calor excedente, ya que su correlación lineal es positiva con las potencias asociadas a los dos ventiladores de la torre de enfriamiento, rondando el 15 %.

En cuanto al uso de las diferentes estrategias empleadas para la detección de observaciones fuera de control, el método en el que se emplean gráficos de control multivariantes para datos autocorrelados empleando la estimación robusta de medias truncadas es el método que resultados más favorables ofrece para los datos en cuestión, siendo capaz de detectar el 100 % de las fechas que el hotel considera como anomalías en el sistema. Nótese que las metodologías empleadas en los gráficos de control han sido muy numerosas, empleando distintas agrupaciones, suponiendo normalidad e independencia, tomando datos en rangos horarios de alta y baja actividad etc. Sin embargo, no todas ellas son satisfactorias. Por otro lado, parece fundamental al mismo tiempo emplear la diferenciación de los datos en función de los estados de funcionamiento, obteniendo así con este método un total del 6.49 % de observaciones totales fuera de control, siendo el método que proporciona el menor porcentaje de falsas alarmas al mismo tiempo que detecta las anomalías, lo que es precisamente el objetivo de este estudio.

Por lo que respecta al análisis de capacidad, se han estimado los índices de capacidad multivariante (índices de Taam, Shahriari, Pan y Lee, Xekalaki y Perakis y Wang y Chen), obteniéndose en todos los casos que las instalaciones de climatización del hotel no son capaces de cumplir especificaciones de la empresa y de la normativa, dados los bajos valores de los índices (<1), por lo que se recomienda implementar un plan de mejora.

Capítulo 5

Conclusiones

En este trabajo se ha abordado el caso de estudio real de la descripción, control y detección de anomalías, así como el análisis de capacidad, de las instalaciones de climatización (HVAC) de un hotel en Latinoamérica. Para ello, primeramente se han aplicado herramientas del análisis exploratorio de datos como son el estudio de correlación, la clasificación supervisada y la reducción de dimensión mediante el análisis de componentes principales. Estas técnicas permiten definir la estructura de dependencia, identificar las variables más relevantes y detectar los diferentes regímenes de funcionamiento o poblaciones, entre otras acciones. Una vez realizado el estudio descriptivo, se han aplicado diferentes gráficos de control para datos multivariantes, ya sean paramétricos (asumiendo distribución normal) como no paramétricos, como son los gráficos r, Q y S. Además, asumiendo que existe autocorrelación entre las observaciones de cada variable, se han aplicado alternativas de gráficos de control que asumen este hecho como son los denominados métodos robustos (MCD, MVE y medias truncadas). Finalmente, se ha realizado un análisis de capacidad de las instalaciones HVAC para cumplir las especificaciones de la empresa y de la normativa.

En el desarrollo de este proyecto se ha presentado un estudio comparativo entre diferentes tipos de gráficos de control a través de la implementación a datos reales. Los datos en cuestión no presentaban normalidad ni independencia, además de contener variables altamente autocorreladas. Es importante tener en cuenta que se trata de una base de datos referente a un proceso nuevo, del cual no se tienen antecedentes ni datos para un calibrado más exhaustivo, lo que complica el estudio considerablemente. A su vez, los datos abarcan un periodo de tiempo en el que se incluyen meses de verano y meses de invierno, así como épocas de temporada alta y temporada baja. Todos estos parámetros hacen que el proceso tenga grandes variaciones, complicando el objetivo del estudio: obtener un método gráfico de control de la calidad apropiado capaz de detectar anomalías con un porcentaje de falsas alarmas reducido.

No obstante, a lo largo del proyecto se han desarrollado ciertos métodos que sí que son capaces de detectar aquellas anomalías registradas por el hotel. Muchos de estos métodos se han visto que pueden no cumplir todas las hipótesis necesarias, pero que sí que muestran resultados satisfactorios, detectando un gran porcentaje de las anomalías. Entre todos ellos, cabe resaltar que el método que contempla la autocorrelación de los datos es aquel que responde mejor a los mismos, detectando el 100 % de las anomalías consideradas por el hotel con un porcentaje de falsas alarmas relativamente bajo. También resulta importante realizar la agrupación correspondientes a los diferentes modo de funcionamiento de los chiller, ya que de este modo, se tiene en cuenta la variabilidad de cada uno de ellos, obteniendo así resultados más favorables que sin emplear agrupaciones o incluso empleando las agrupaciones clúster. Finalmente, cabe resaltar que los gráficos de control no paramétricos no responden adecuadamente a los datos en cuestión, obteniendo un número de observaciones fuera de control realmente reducido, lo que puede ser resultado de una sobresuavización. Estos gráficos podrían ser una alternativa, pero habría que hacer un estudio más completo de búsqueda de la ventana óptima que podría ser objeto de un estudio futuro.

Partiendo de una base de datos donde las variables son numerosas, resulta lógico aplicar un análisis de componentes principales, sobre las cuales aplicar los métodos gráficos de control de calidad. Como se ha visto a lo largo del trabajo, esta metodología no resulta muy interesante ya que tiende a que el número de detecciones sea muy reducida, a pesar de que las componentes principales cuenten con un porcentaje de variabilidad explicada de los datos originales elevada.

En el contexto de datos multivariantes, el análisis clúster o las técnicas de agrupación, como pueden ser los métodos jerárquicos, pueden ser una herramienta adecuada. De esta manera, se trata de obtener grupos más homogéneos sobre los cuales poder aplicar los gráficos de control. La ventaja de este planteamiento radica en que al tener grupos precisamente más homogéneos, la variabilidad dentro de los mismos es inferior al conjunto de datos completo original, llevando a que los gráficos de control sean más capaces de detectar observaciones fuera de control, ya que estos van ligados a la variabilidad de las observaciones respecto del conjunto total de datos.

Más allá del número de variables, el número de observaciones es un parámetro a tener muy en cuenta. Esto se debe a que la construcción por fases es un enfoque muy recomendado, donde se realiza un primer análisis de calibrado con un conjunto de observaciones bajo control para posteriormente en la segunda de las fases proceder a la monitorización, donde se incluyen las nuevas observaciones que se quieren examinar. Si el número de datos es reducido, esta construcción por fases no será posible, o en el mejor de los casos, se estará reduciendo la variabilidad de la primera fase de calibrado, lo que conlleva a que posiblemente la variabilidad se vea reducida al mismo tiempo que los límites de control. Esto puede llevar a que el análisis estadístico por gráficos de control tenga un número de observaciones fuera de control muy elevado aún cuando no son del todo ciertas, incrementando notablemente el número de falsas alarmas. En la mayoría de los procesos reales, el número de observaciones fuera de control es desconocido a priori, por lo que tomar ciertas pruebas piloto donde sí se tenga conocimiento de las observaciones fuera de control puede resultar muy útil. En estas primeras pruebas, el objetivo es estudiar qué tipo de gráfico de control funciona mejor para el tipo de datos en cuestión, analizando tanto el número de observaciones fuera de control detectadas como el número de falsas alarmas detectadas por cada uno de los métodos aplicados.

Se debe tener en cuenta que no todos los métodos empleados son satisfactorios, bien porque no detectan las anomalías registradas por el hotel o bien porque las detectan pero asumen un número elevado de falsas alarmas, lo que tampoco es interesante. Hay que tener en cuenta que, a pesar de que los datos empleados en la implementación a datos reales del proyecto cuentan con numerosas variables, el comportamiento del sistema de enfriamiento depende de muchas otras variables, dependiendo incluso de otras variables difíciles de medir. No obstante, gracias a la implementación a datos reales realizada en el presente proyecto, se observa que es posible obtener resultados satisfactorios en términos de porcentaje de observaciones fuera de control frente al porcentaje de falsas alarmas incluso con cinco de las doce variables iniciales. Es decir, la selección de las variables a tener en cuenta en el análisis resulta fundamental. De todas formas, este tema del control estadístico de la calidad aplicado a datos multivariantes es todavía un tema en desarrollo, sobre el que se está estudiando la aplicación de nuevos métodos y procedimientos.

En cuanto a líneas futuras, será posible emplear los datos actuales como datos para la calibración, diferenciando los cuatro diferentes estados de funcionamiento de los chiller. Esto llevará a un análisis más exhaustivo de los próximos meses. Como bien muestra el análisis de capacidad de los datos, el proceso no es capaz, lo que invita a realizar ciertas mejoras en el proceso para futuros análisis. Será posible emplear un año completo para la primera de las fases del estudio, la fase de calibrado, teniendo así bajo control un año completo incluyendo toda la variabilidad de los datos de todos los meses. Por lo que respecta a la implementación de los datos realizada hasta el momento, el código empleado para los gráficos de control desarrollado en R se puede ver en el Apéndice C.

Resulta evidente que el control estadístico de procesos está experimentando un resurgimiento con la aparición de nuevos datos y problemas en el contexto de la Industria 4.0, por lo que la propuesta de nuevas metodologías adaptadas a este contexto, como los gráficos para datos funcionales o la mejora de los no paramétricos tienen un futuro muy prometedor.

Anexos

Anexo A

Algunas herramientas básicas

A.1. Pruebas de Repetibilidad y Reproducibilidad (R&R)

Las pruebas de Repetibilidad y Reproducibilidad, o pruebas R&R, son pruebas para medir el desempeño de un sistema de medida.

- Repetibilidad:

  La repetibilidad hace referencia a la variabilidad del sistema de medición. Dicho en otras palabras, es la variación de las mediciones obtenidas con un sistema de medición cuando se usa varias veces por un usuario, midiendo la misma característica y sobre la misma pieza. Se trata de la variación o habilidad inherente del equipo mismo, siendo una variación de causa común (error aleatorio) de intentos sucesivos y bajo condiciones definidas de medición.

- Reproducibilidad:

  Se trata de la variación en el promedio de las mediciones hechas por diferentes evaluadores usando el mismo sistema de medición cuando se mide la misma característica y sobre la misma pieza. Esto a menudo es importante para instrumentos manuales influenciados por la habilidad del operador. Sin embargo, para procesos de medición donde el operador no es una fuente principal de variación los errores de reproducibilidad son pequeños o incluso resultan inexistentes. Es decir, hace referencia a la variabilidad en las medidas debidas al operador.

El modelo de medición para un estudio R&R es el siguiente:

$$y = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \epsilon_{ijk} \tag{A.1}$$

donde α_i identifica al efecto de las piezas, β_j al efecto operador y ϵ_{ijk} es el error aleatorio del proceso,

$$\alpha_i \sim N(0, \sigma_\alpha) \rightarrow Efecto \quad piezas \tag{A.2}$$

$$\beta_j \sim N(0, \sigma_\beta) \rightarrow Efecto \quad operador \tag{A.3}$$

$$\alpha\beta_{ij} \sim N(0, \sigma_{\alpha\beta}) \rightarrow Interacción \quad piezas - operador \tag{A.4}$$

$$\epsilon_{ijk} \sim N(0, \sigma) \rightarrow Error \tag{A.5}$$

Nótese que en dicho modelo, los efectos y el término de error se consideran independientes. En cuanto a la variabilidad del sistema de medición, viene dada por:

$$\sigma^2_{Repetibilidad} = \sigma^2_{Error} \tag{A.6}$$

Por otro lado, la variabilidad debida al operador:

$$\sigma^2_{Reproducibilidad} = \sigma^2_{Operador} + \sigma^2_{OperadorxPieza} \tag{A.7}$$

Se define por tanto variabilidad total del sistema de medición, como:

$$\sigma^2_{Total} = \sigma^2_{Pieza} + \sigma^2_{Operador} + \sigma^2_{OperadorxPieza} + \sigma^2_{Error} \tag{A.8}$$

Finalmente, la variabilidad de la prueba R&R viene dada por la suma de los componentes de la varianza repetibilidad y reproducibilidad de la prueba:

$$\sigma^2_{R\&aR} = \sigma^2_{Repetibilidad} + \sigma^2_{Reproducibilidad} \tag{A.9}$$

A.2. ANOVA

El análisis de la varianza (ANOVA) es un método estadístico que permite estudiar la relación entre una variable dependiente cuantitativa y una o varias variables independientes cualitativas. Al fin y al cabo, el objetivo es comprobar si varias muestras (k) de una misma variable proceden de la misma población o no, es decir, si provienen de poblaciones distintas. En definitiva se estaría contrastando la homogeneidad de la muestra.

Primeramente, conviene definir ciertos conceptos:

- Factores: cada una de las variables independientes, causas de la posible heterogeneidad de las muestras.
- Niveles del factor: cada uno de los valores posibles del factor.
- Efectos: medida de la influencia de los factores en la variable dependiente.
- Error muestral: debido a la aleatoriedad en la selección de las muestras.

El análisis de la varianza puede darse en un entorno de efectos fijos o efectos aleatorios. A continuación se definen ambos conceptos:

- Por un lado los efectos fijos, que son aquellos niveles del factor se seleccionan de modo específico por el experimentador.
- Por otro lado, los efectos aleatorios, siendo éstos los niveles de un factor son una muestra aleatoria de una población mayor de tratamientos.

El modelo probabilístico que se plantea en un ANOVA, en este caso de efectos fijos, requiere de ciertas hipótesis:

- Existen k niveles o grupos del factor A, por lo que a cada grupo i ($i = 1, 2, \cdots, k$) le corresponde un valor de la variable respuesta Y, que se denotará por Y_i.
- Homocedasticidad:
$$Var(Y_i) = \sigma^2 \qquad \forall i = 1, 2, \cdots, k \tag{A.10}$$
- Distribución normal:
$$\varepsilon_i \to N(0, \sigma) \quad ; \quad Y_i \to N(\mu_i, \sigma) \tag{A.11}$$
- El modelo puede representarse de las dos siguientes formas:
$$\mu_i = \mu + \varepsilon_i \quad ; \quad i = 1, 2, \cdots, k \tag{A.12}$$

$$\mu_i = \mu + \delta_i + \varepsilon_i \quad ; \quad i = 1, 2, \cdots, k \quad ; \quad \sum_{i=1}^{k} \delta_i = 0 \quad ; \quad \delta_i \to N(0, \sigma_b) \tag{A.13}$$

Como ya se ha mencionado, el Análisis de la Varianza tiene como propósito discernir si los valores de la variable cuantitativa Y son diferentes según los distintos valores o niveles de la variable cualitativa A.

Para ello, se compara la dispersión de los valores de la variable Y dentro de cada uno de los niveles del factor A con respecto a la dispersión de los valores de Y entre niveles. Es decir, se pretende comparar si las diferencias intragrupos es grande en comparación con la varianza entre grupos. Como consecuencia, si la variación entre niveles es lo suficientemente elevada, las observaciones dentro de niveles diferentes, se considerarán diferentes y existirá dependencia entre la variable cuantitativa y la variable cualitativa. En sí, este método no es más que una generalización del contraste de medias para más de dos poblaciones.

De manera más generalizada, siendo A y B los factores principales, se plantea el modelo probabilístico correspondiente al ANOVA de dos factores, en este caso con efectos fijos.

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk} \tag{A.14}$$

donde μ es la media global, A_i es el efecto del nivel i-ésimo del primer factor, factor A, y B_j es el efecto del nivel j-ésimo del segundo factor, factor B. Finalmente, ε_{ijk} son las desviaciones aleatorias alrededor de las medias, que también se debe asumir que se distribuyen normalmente, que son independientes y con media 0 y varianza σ^2.

Además de las condiciones de muestreo aleatorio, normalidad e independencia, en el ANOVA de dos factores se debe contemplar el efecto de $(AB)_{ij}$, que corresponde al efecto de la interacción entre los factores A y B, que representa el hecho de que el efecto de un determinado nivel de uno de los factores sea distinto para cada nivel del otro factor.

A continuación se muestra la correspondiente Tabla A.1, la tabla ANOVA de dos factores con interacción:

Tabla A.1: Tabla ANOVA de dos factores con interacción.

Fuente	Suma de cuadrados	Grados de libertad	Varianza	Test F
Factor A	scA	a-1	$scmA = \frac{scA}{a-1}$	$F_A = \frac{scmA}{scmR}$
Factor B	scB	b-1	$scmB = \frac{scB}{b-1}$	$F_B = \frac{scmB}{scmR}$
Interacción AB	scAB	(a-1)(b-1)	$scmAB = \frac{scAB}{(a-1)(b-1)}$	$F_AB = \frac{scmAB}{scmR}$
Residual	scR	ab(r-1)	$scmR = \frac{scR}{ab(r-1)}$	
Global	scG	abr-1		

Como se puede ver en las siguientes líneas, se procede a la descomposición de la variabilidad.

$$scA = br \sum_{i=1}^{a} (\bar{y}_{i..} + \bar{y}_{...})^2 \tag{A.15}$$

$$scB = ar \sum_{j=1}^{b} (\bar{y}_{.j.} + \bar{y}_{...})^2 \tag{A.16}$$

$$scAB = r \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \tag{A.17}$$

$$scR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{ij.})^2 \tag{A.18}$$

$$scG = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{...})^2 \tag{A.19}$$

Como nota, se debe tener en cuenta que cuando los tamaños muestrales no son iguales, las variables aleatorias SC(A),SC(B) y SC(AB) no son independientes, lo que conlleva a que $scA + scB + scAB + scR \neq scG$. Esta última igualdad se cumplirá, por lo tanto, solamente cuando se den situaciones donde los tamaños muestrales sean iguales.

En primer lugar, conviene realizar el contraste de hipótesis de la interacción, donde se comparará un modelo reducido sin interacción frente a la hipótesis alternativa del modelo completo.

$$\begin{aligned} H_0: &\quad (AB)_{ij} - (AB)_{i.} - (AB)_{.j} + (AB)_{..} = 0 \quad \forall\ i,j \\ H_1: &\quad \exists ij/(AB)_{ij} - (AB)_{i.} - (AB)_{.j} + (AB)_{..} \neq 0 \end{aligned} \tag{A.20}$$

y equivalentemente:

$$H_0: \quad Y_{ijk} = \mu + A_i + B_j + \varepsilon_{ijk} \quad \forall ijk \quad (Modelo\ Reducido) \tag{A.21}$$

$$H_1: \quad Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk} \quad \forall ijk \quad (Modelo\ Completo) \tag{A.22}$$

Una vez realizado dicho contraste, si la interacción resulta no significativa, se puede proceder a realizar el contraste de efectos principales, donde por ejemplo, para el efecto A se debe plantear el modelo reducido sin el efecto A frente al modelo completo.

$$H_0: \quad A_1 = A_2 = \cdots = A_a \tag{A.23}$$

$$H_1: \quad \exists ij/A_i \neq A_j \tag{A.24}$$

y equivalentemente:

$$H_0: \quad Y_{ijk} = \mu + B_j + \varepsilon_{ijk} \quad \forall ijk \quad (Modelo\ Reducido) \tag{A.25}$$

$$H_1: \quad Y_{ijk} = \mu + A_i + B_j + \varepsilon_{ijk} \quad \forall ijk \quad (Modelo\ Completo) \tag{A.26}$$

De manera análoga se deberá plantear el contraste para el efecto principal B, contrastando un modelo reducido sin el efecto de B frente al modelo completo.

A.3. La Curva Característica de Operación (OC)

La Curva Característica de Operación muestra la probabilidad que tiene un plan de muestreo de aceptar un lote con una calidad dada. Si el lote es muy bueno debe tener una probabilidad alta de ser aceptado y si es muy malo debe tener una probabilidad pequeña de ser aceptado.

En el caso más sencillo, suponiendo que el tamaño N del lote es muy grande, la distribución del número de unidades defectuosas en la muestra de tamaño n, es Binomial con parámetros n y p, donde p es la fracción de unidades defectuosas en el lote.

De esta manera, la probabilidad de tener t unidades defectuosas en la muestra es:

$$P(t\ defectuosos) = \frac{\binom{k}{t}\binom{N-k}{n-t}}{\binom{N}{n}}, \quad donde\ 0 \leq t \leq min(n,k) \tag{A.27}$$

Y por lo tanto, la probabilidad de aceptación del lote es la probabilidad de tener c o menos unidades defectuosas en la muestra:

$$P_{aceptación} = P(t \leq c) = \sum_{t=0}^{c} P(t\ defectuosos) \tag{A.28}$$

Por lo tanto, la probabilidad de tener t unidades defectuosas en la muestra es:

$$P(t\ defectuosos) = \frac{n!}{t!(n-t)!}p^t(1-p)^{n-t} \tag{A.29}$$

Entonces, la probabilidad de aceptación del lote es la probabilidad de tener c o menos unidades defectuosas en la muestra:

$$P_{aceptación} = P(t \leq c) = \sum_{t=0}^{c} \frac{n!}{t!(n-t)!}p^t(1-p)^{n-t} \tag{A.30}$$

Una curva OC ideal se muestra en la Figura A.1, la cual acepta todos los lotes con proporción de defectuosos menor o igual al 1 %, y con probabilidad cero no acepta ningún lote con proporción de defectuosos mayor a dicho 1 %. En la realidad no existen planes de muestreo con este tipo de curvas OC, que haga discriminación perfecta entre lotes buenos y malos. No obstante, cuando más se aleje la curva real de esta ideal teórica peor proceso se tendrá en cuanto a riesgo se refiere.

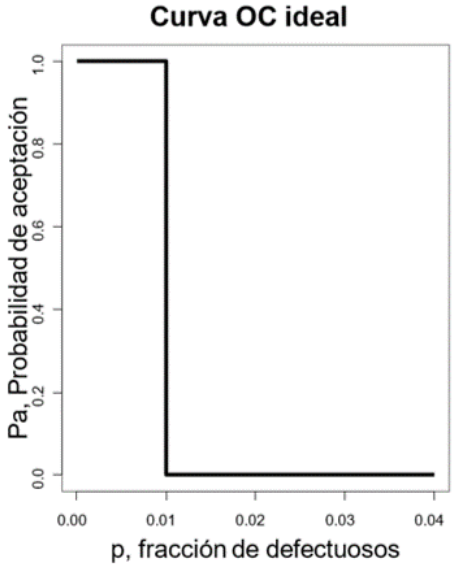


Figura A.1: Curva Característica de Operación Ideal.

Empleando las expresiones anteriores y un poco de lógica, resulta sencillo ver los diferentes efectos que tienen en las curvas OC las variaciones en los parámetros n y c. Por un lado, la precisión con que un plan de muestreo discrimina entre lotes buenos y malos aumenta con el tamaño de la muestra. A tamaños muestrales más grandes, mayor será también el poder de discriminación. Por otro lado,

cuanto más pequeño sea el número de aceptación c, la curva OC se verá desplazada al lado izquierdo, indicando mayor discriminación a niveles menores de la fracción de defectuosos en el lote.

Como ya se ha definido previamente, la curva OC no es más que la representación gráfica del riesgo β en función de la magnitud del cambio que se pretende estudiar. En el caso en el que el interés radique en calcular el error de tipo II al pasar de un valor nominal μ a un valor $\mu_1 = \mu_0 + k\sigma$, suponiendo que la característica de calidad medible se distribuye según una normal $N(\mu, \sigma)$, con media $\bar{x} = N(\mu, \frac{\sigma}{\sqrt{n}})$, límite superior $UCL = \mu_0 + d\frac{\sigma}{\sqrt{n}}$ y límite inferior $LCL = \mu_0 - d\frac{\sigma}{\sqrt{n}}$:

$$\beta = P\left(LCL \leq \bar{x} \leq \frac{UCL}{\mu = \mu_1 = \mu_0 + k\sigma}\right) \tag{A.31}$$

$$\beta = \Phi\left[\frac{UCL - (\mu_0 + k\sigma)}{\frac{\sigma}{\sqrt{n}}}\right] - \Phi\left[\frac{LCL - (\mu_0 + k\sigma)}{\frac{\sigma}{\sqrt{n}}}\right] \tag{A.32}$$

$$\beta = \Phi\left[d - k\sqrt{n}\right] - \Phi\left[-d - k\sqrt{n}\right] \tag{A.33}$$

Para la construcción de la respectiva curva OC, se debe graficar en el eje de ordenadas la probabilidad del error de tipo II o riesgo β y en el eje de abscisas, los distintos valores de k, según la expresión $\beta = \Phi\left[d - k\sqrt{n}\right] - \Phi\left[-d - k\sqrt{n}\right]$. Finalmente, dados los puntos en la curva característica $(0, \alpha)$ y (μ_1, β), el tamaño muestral vendrá determinado por:

$$n = \left(\frac{Z_{\alpha/2} + Z_\beta}{d}\right)^2 \tag{A.34}$$

siendo d el descentrado relativo, que se calcula como:

$$d = \frac{|\mu - \mu_0|}{\sigma_0} \tag{A.35}$$

Entonces, la probabilidad de rechazar la hipótesis nula $H_0(\mu = \mu_0)$ frente a la alternativa $H_1(\mu \neq \mu_0)$ es:

$$2 - \left[\Phi(Z_{\alpha/2} - d\sqrt{n}) + \Phi(Z_{\alpha/2} + d\sqrt{n})\right] \tag{A.36}$$

La curva característica o probabilidad de aceptar H_0 en función del descentrado d, será:

$$OC(d) = \Phi(Z_{\alpha/2} - d\sqrt{n}) + \Phi(Z_{\alpha/2} + d\sqrt{n}) - 1 \tag{A.37}$$

En el supuesto de un contraste unilateral con hipótesis nula $H_0(\mu = \mu_0)$ frente a la alternativa $H_1(\mu > \mu_0)$, la curva será:

$$OC(d) = \Phi(Z_{\alpha/2} - d\sqrt{n}),\ d \geq 0. \tag{A.38}$$

En la práctica, un enfoque usual de diseño de planes de muestreo de aceptación, requiere que la curva OC pase por dos puntos designados. En otras palabras, se desea que la probabilidad de aceptación sea $1 - \alpha$ para lotes con una fracción de defectuosos p_1, y que la probabilidad de aceptación sea β para lotes con una fracción de defectuosos p_2.

De manera que, suponiendo que el muestreo Binomial es apropiado, con tamaño muestral n y número de aceptación c, se deberá proceder a la solución del sistema de ecuaciones:

$$\begin{cases} 1 - \alpha = \sum_{k=0}^{c} \frac{n!}{k!(n-k)!} p_1^k (1 - p_1)^{n-k} \\ \beta = \sum_{k=0}^{c} \frac{n!}{k!(n-k)!} p_2^k (1 - p_2)^{n-k} \end{cases} \tag{A.39}$$

No obstante, se trata de un sistema de ecuaciones no lineales muy difícil de resolver, por lo que suele emplearse el siguiente nomograma de la Figura A.2 para resolver estas ecuaciones. Para la resolución,

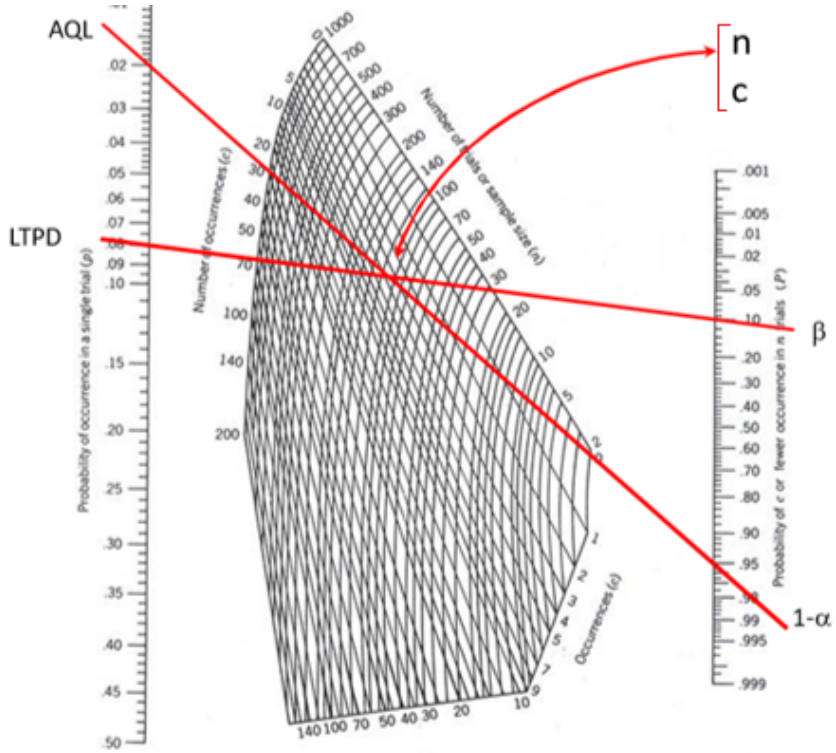


Figura A.2: Nomograma binomial.

bastaría con unir como se muestra en la Figura las líneas correspondientes a los valores prefijados de AQL, LTPD, $1-\alpha$ y β, y obteniendo así el n y c necesarios.

Finalmente y a modo de resumen, se listan algunas de las propiedades de las Curvas Característica de Operación:

- No existe un plan de muestreo que tenga la curva OC ideal.
- Al aumentar el tamaño de muestra con el número de aceptación se obtienen curvas OC más cercanas a la ideal.
- El criterio de tamaño de muestra igual a un porcentaje del tamaño del lote no es un buen criterio.
- Al disminuir el número de aceptación, c, la curva OC cae más rápido lo que conlleva a que los planes sean más estrictos.
- Los planes con $c = 0$ no siempre son los más apropiados.
- La influencia del tamaño del lote en el diseño de planes de muestreo adecuados es menor de lo que comúnmente se cree.

Anexo B

Agrupaciones

La base de datos empleada en el presente trabajo, incluye siete meses en los que se estudian diferentes parámetros de temperaturas, consumos, porcentajes de ocupación etc. En estos siete meses, existen variaciones en algunas variaciones como la estación del año, temporada alta o baja, festividades y muchas otras variables que pueden llevar a que la variabilidad dentro del periodo de enero a julio sea elevada para muchos de los parámetros a estudiar. Por ello, puede ser interesante tratar de aplicar ciertas agrupaciones de los datos, pudiendo seguir diversos criterios, con el objetivo de obtener grupos más homogéneos, con menor variabilidad dentro de cada uno de ellos.

B.1. Técnicas de formación de grupos

Las técnicas de formación de grupos están formadas por una amplia y variada gama de técnicas y procedimientos. El aspecto común a todos ellos es el objetivo de formar grupos que todavía no están definidos, como es el caso del análisis discriminante, donde los grupos están perfectamente definidos a priori y el propósito es clasificar a un individuo en alguno de los grupos. Sin embargo, las técnicas de formación de grupos pretenden formar grupos, reconociendo patrones o estructuras dentro de la población general, en función de las observaciones de la muestra.

A continuación se presentan los métodos más empleados en la actualidad para la formación de grupos, los métodos jerárquicos y los de particionamiento. Dichos métodos no establecen suposiciones sobre el modelo estadístico que genera los datos, por lo que se pueden interpretar como métodos descriptivos que ayudan a la comprensión sobre las propiedades que presenta un conjunto de datos multivariantes.

B.1.1. Métodos jerárquicos

Estos métodos son conocidos también como métodos de taxonomía numérica, ya que tienen mucha semejanza con la taxonomía de los seres vivos en Biología. Los métodos jerárquicos parten de una matriz de distancia entre individuos, y en base a ella pretenden un agrupamiento de los individuos a distintos niveles. En el nivel más bajo cada grupo estaría formado por individuos, mientras que los grupos a niveles superiores serán el resultado de agregar grupos de niveles inferiores.

Los algoritmos que se emplean para crear esta jerarquía pueden ser de dos tipos:

- Aglomerativos:

  Partiendo de los individuos, se construyen grupos formados por individuos, para después construir grupos, mediante la agregación de los grupos ya formados en etapas anteriores.

- Divisivos:

Partiendo del grupo total formado por todos los individuos, se genera una división en subgrupos (generalmente en dos), que más adelante vuelven a ser subdivididos.

Los métodos aglomerativos son los más empleados en la actualidad, y el algoritmo que lo caracteriza es el siguiente:

1. Se definen los grupos o clústers $C_1, C_2, \cdots, C_n$, que está formados cada uno por un individuo.
2. Se buscan los dos grupos C_i y C_j que están más próximos, se juntan y consecuentemente se reduce el número de grupos.
3. Se recalculan las distancias de todos los demás grupos al nuevo grupo, formado al juntar C_i y C_j.
4. Si el número de grupos es uno, se detiene el algoritmo. En otro caso, se vuelve al paso 2.

En el paso tercero del algoritmo a su vez existen diferentes metodologías a emplear:

- Método del mínimo: Consiste en definir $d(C_k, C_i \cap C_j) = min\{d(C_k, C_i), d(C_k, C_j)\}$, donde $d(C_r, C_s)$ denota la distancia del grupo C_r al grupo C_s. Esto es equivalente a definir la distancia entre dos grupos de la siguiente manera:

$$d(C_r, C_s) = \min_{i \in C_r, j \in C_s} d_{ij} \tag{B.1}$$

  Siendo d_{ij} la distancia entre los individuos i y j, que en este caso pertenecen a los grupos C_r y C_s respectivamente.

- Método del máximo: Consiste en definir $d(C_k, C_i \cap C_j) = max\{d(C_k, C_i), d(C_k, C_j)\}$. En este caso es equivalente a definir la distancia entre dos grupos como la mayor distancia entre sus individuos:

$$d(C_r, C_s) = \max_{i \in C_r, j \in C_s} d_{ij} \tag{B.2}$$

- Método del promedio: Consiste en definir $d(C_k, C_s) = \frac{n_i}{n_i+n_j} d(C_k, C_i) + \frac{n_j}{n_i+n_j} d(C_k, C_j)$, siendo n_i y n_j el número de individuos en los grupos C_i y C_j, respectivamente. Es equivalente a definir la distancia entre dos grupos como el promedio de las distancias entre sus individuos:

$$d(C_r, C_s) = \frac{1}{n_r n_s} \sum_{i \in C_r, j \in C_s} d_{ij} \tag{B.3}$$

B.1.2. Métodos de particionamiento

Los métodos de particionamiento, según ya se dijo en la introducción, pretenden una partición de los individuos de la muestra en k grupos, en base a los valores de las variables observadas en cada individuo. Por supuesto, los grupos se formarán por proximidad en el espacio d-dimensional, siendo d el número de variables. Un criterio natural para la formación de los grupos, consistiría en elegir la partición en grupos que aga mínima la variabilidad dentro de cada grupo, medida por la suma de cuadrados intra-grupo. El criterio parece sencillo pero el problema será la imposibilidad de recorrer todas las particiones posibles de n individuos en k grupos, ya que con valores de n y k grandes el numero de particiones crece de manera desorbitada.

Ante esta situación, se requiere de un algoritmo que precisamente permita, partiendo de una solución razonable, sea capaz de llegar a una solución 0mejor de acuerdo con el criterio, mediante sucesivos pasos que aporten mejoras sucesivas. El algoritmo más conocido es el algoritmo K-means. Este algoritmo propone el siguiente procedimiento:

1. Crear una partición inicial en k grupos.

2. Asignar cada individuo al grupo cuyo centro, que denominaremos centroide, le quede más próximo. Recalcular los centroides en base a los grupos modificados.

3. Repetir el paso 2 hasta que no haya más reasignaciones.

El primer paso se puede obtener proporcionando k centroides iniciales escogidos adecuadamente a la vista de la muestra. De hecho, pueden ser individuos de la muestra. En cuanto al segundo paso,este admite muchas variantes dependiendo del algoritmo concreto. Planteado en términos generales, la idea es considerar posibles cambios de algunos individuos de un grupo a otro. Serían candidatos a cambiar de grupo los individuos que ocupan posiciones fronterizas. Para cada posible cambio, se evalúan las consecuencias que tendría sobre el criterio (por ejemplo, suma de cuadrados intra-grupo), y se efectúa el cambio que más contribuye a mejorar el criterio. Se detiene el algoritmo cuando no hay cambios que redunden en mejoras apreciables del criterio.

Los algoritmos de las k-medias presentan ciertas limitaciones, como puede ser la dependencia de la partición inicial que se escoja, sí como la tendencia a formar grupos esféricos, ya que la distancia que se considera es la distancia euclídea usual.

B.2. Técnicas de agrupación a datos reales

A continuación se presenta la implementación a los datos reales del sistema de enfriamiento del hotel sudamericano empleando los métodos de agrupación mencionados anteriormente.

Método manual

B.1

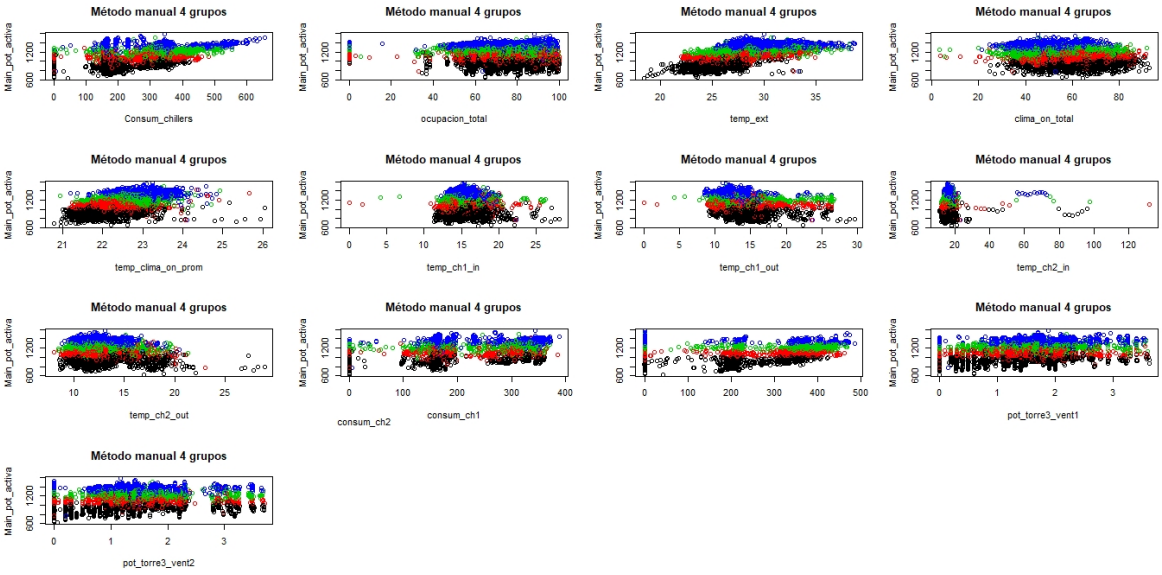


Figura B.1: Gráficos de la agrupación por el método manual en función de las diferentes variables.

B.2

De hecho, puede resultar interesante conocer si los grupos clúster son significativamente distintos o no. El t-test es un test estadístico paramétrico que permite contrastar la hipótesis nula de que las

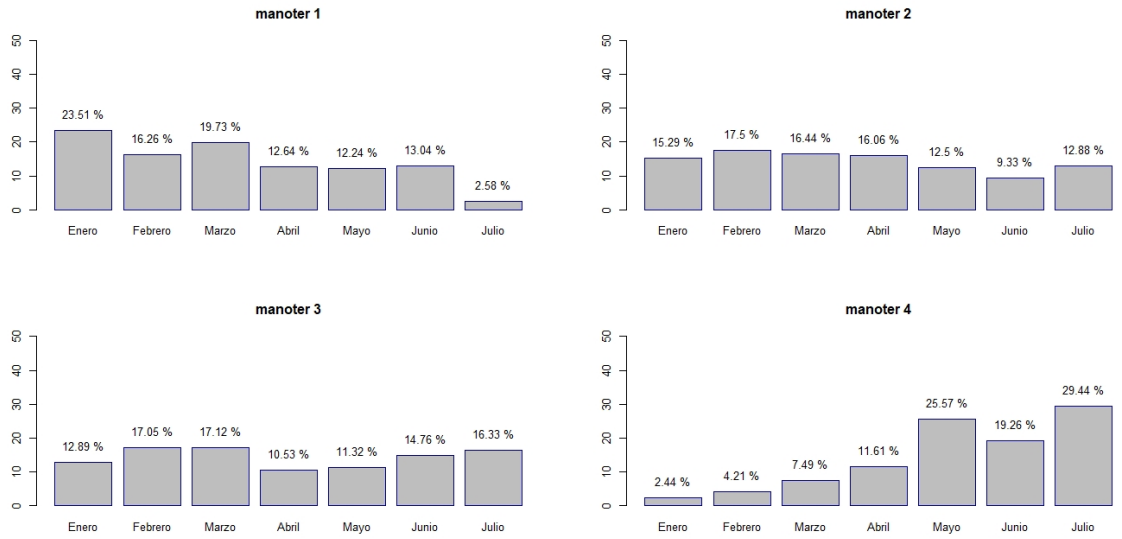


Figura B.2: Gráficos de los porcentajes de las observaciones en cada mes de la agrupación por el método manual.

medias de dos poblaciones son iguales, frente a la hipótesis alternativa, que implica que no lo son.

$$\begin{aligned} H_0: &\quad \mu_A = \mu_B \\ H_1: &\quad \mu_A \neq \mu_B \end{aligned} \tag{B.4}$$

Otra forma equivalente de definir estas hipótesis es:

$$\begin{aligned} H_0: &\quad \mu_A - \mu_B = 0 \\ H_1: &\quad \mu_A - \mu_B \neq 0 \end{aligned} \tag{B.5}$$

A pesar de la sencillez y utilidad del t-test, para que sus resultados sean válidos es necesario que se cumplan una serie de condiciones, entre las que se encuentran:

- Independencia: Las observaciones tienen que ser independientes las unas de las otras.
- Normalidad: Las poblaciones que se comparan tienen que seguir una distribución normal.
- Igualdad de varianza (homocedasticidad): la varianza de las poblaciones comparadas debe de ser igual.

El test de Mann−Whitney−Wilcoxon (WMW), también conocido como Wilcoxon rank-sum test o u-test, es un test no paramétrico que contrasta si dos muestras proceden de poblaciones equidistribuidas.

La idea fundamental de este test es que si las dos muestras comparadas proceden de la misma población, al juntar todas las observaciones y ordenarlas de menor a mayor, cabría esperar que las observaciones de una y otra muestra estuviesen intercaladas aleatoriamente. Por lo contrario, si una de las muestras pertenece a una población con valores mayores o menores que la otra población, al ordenar las observaciones, estas tenderán a agruparse de modo que las de una muestra queden por encima de las de la otra.

Acorde a esta idea, el test de Mann−Whitney−Wilcoxon contrasta que la probabilidad de que una observación de la población X supere a una observación de la población Y es igual a la probabilidad de que una observación de la población Y supere a una de la población X. Es decir, que los valores de una población no tienden a ser mayores que los de otra.

$$H_0: \quad P(X \geq Y) = P(Y \geq X)$$

$$H_0: \quad P(X \geq Y) = 0.\ 5\ H_\alpha: \quad P(X \geq Y) \neq P(Y \geq X) \tag{B.6}$$

$$H_\alpha: \quad P(X \geq Y) \neq 0.5$$

Es común encontrar mencionado que el test de Mann−Whitney−Wilcoxon compara medianas, sin embargo, esto solo es cierto cuando las poblaciones comparadas difieren únicamente es su localización, pero el resto de características (dispersión, asimetría−) son iguales.

Por otro lado, las condiciones necesarias del test de Mann−Whitney−Wilcoxon:

- Los datos tienen que ser independientes.
- Los datos tienen que ser ordinales o bien se tienen que poder ordenar de menor a mayor.
- No es necesario asumir que las muestras se distribuyen de forma normal o que proceden de poblaciones normales. Pero, para que el test compare medianas, ambas han de tener el mismo tipo de distribución (varianza, asimetría−).
- Igualdad de varianza entre grupos (homocedasticidad).

Aplicando el contraste de hipótesis de Mann−Whitney−Wilcoxon a los datos y los 4 grupos identificados, se obtienen que los cuatro grupos muestran evidencias suficientes para un nivel de significación del 95 % para rechazar la hipótesis nula de que los clúster son diferentes por lo que a la variable main_pot_activa se refiere, lógico conociendo que esta agrupación se ha realizado en función de los valores de dicha variable. Los grupos se muestran distintos también en cuanto a la variable clima_on_total . Sin embargo, no ocurre lo mismo para las variables consum_chillers y ocupacion_total, en los que si que se observan ciertos pares de clúster que contienen a la línea discontinua del cero, lo que implica que esos pares de clúster para sus respectivas variables no muestran evidencias suficientes para el rechazo de la H_0 de igualdad de los grupos, o lo que es lo mismo, que para ellas el test resulta estadísticamente no significativo. En el caso de consum_chillers los grupos 3 y 4, que son precisamente los grupos que se muestran distintos también para la variable ocupacion_total, como puede apreciarse en la Figura B.3. En cuanto a las variables temp_clima_on_prom, temp_ext, temp_ch1_out y temp_ch1_in, se representa en la Figura B.4 los resultados correspondientes. En dicha Figura se observa rápidamente cómo en las variables temp_clima_on_prom y temp_ext ni siquiera se observa la línea discontinua vertical correspondiente al cero, lo que indica que las diferencias existentes en las agrupaciones dos a dos son realmente grandes, ya que los intervalos de confianza ni siquiera se acercan al valor cero, que implicaráa que no existen evidencias suficientes para un nivel de significación del 95 % para rechazar la hipótesis nula como resultado de un contraste estadísticamente significativo. Sin embargo, en cuanto a las variables asociadas a la entrada y la salida del chiller 1 sí que existen comparaciones dos a dos que no son estadísticamente significativas, lo que implica que no existen evidencias suficientes como para decir que los grupos son diferentes. Esto ocurre para la variable temp_ch1_out con el par de grupos 2 y 3. Para la variable temp_ch1_in prácticamente todas las variables muestran un resultado estadísticamente no significativo, es decir, que solo existen evidencias para afirmar que los grupos son diferentes en el caso de la comparación del grupo 2 con el grupo 4. A continuación, en la Figura B.5 se aplica el mismo contraste, pero en este caso para las variables temp_ch2_out, temp_ch2_in, consum_ch1 y consum_ch2. En este caso, la variable consum_ch1 ni siquiera deja ver la línea vertical referente al cero, valor a contener en caso de que no existan evidencias de que los grupos comparados son distintos, por lo que

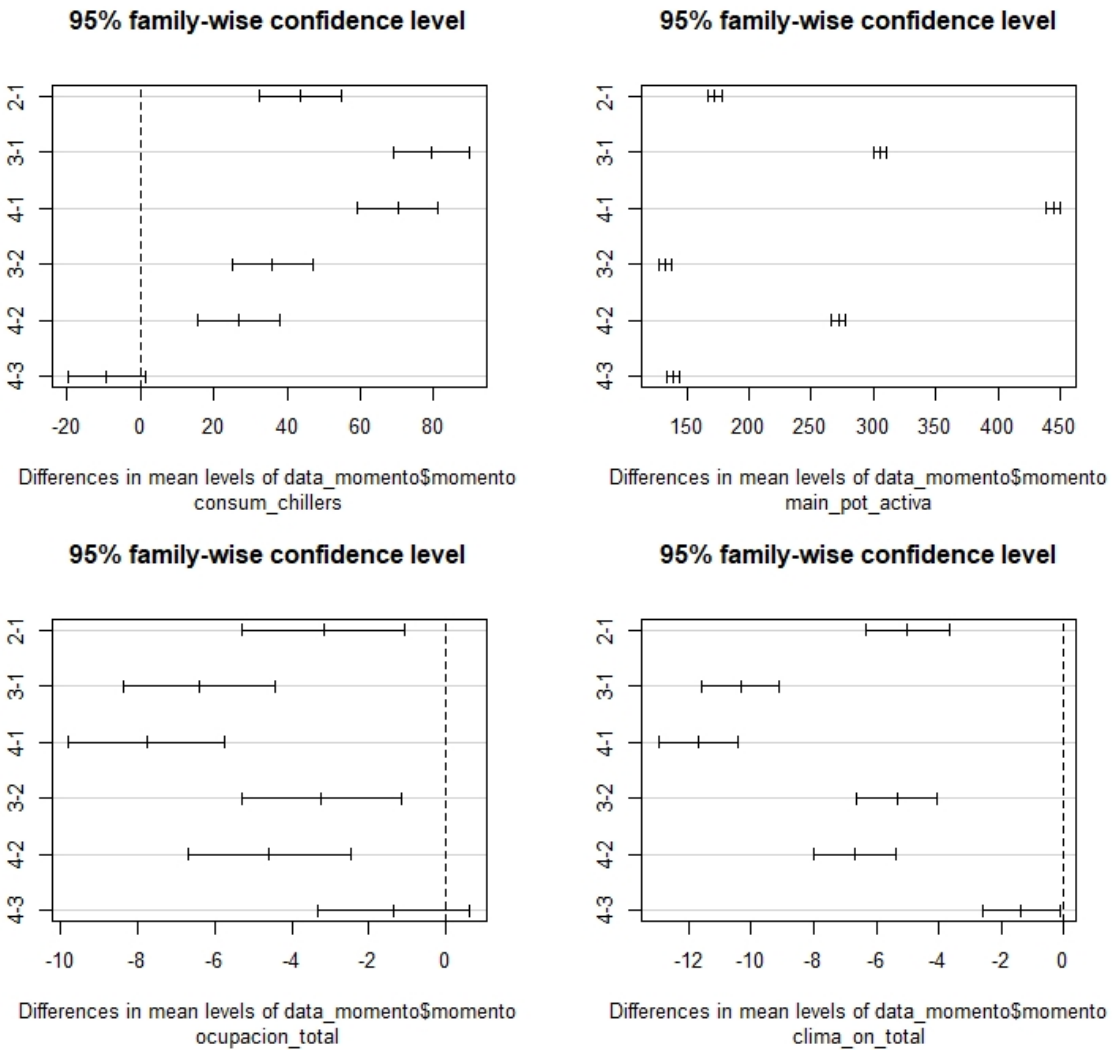


Figura B.3: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos en función de la potencia activa de forma manual.

esto denota directamente que los pares comparados en referencia a dicha variable son estadísticamente muy diferentes.

En términos de la variable temp_ch2_in, las comparaciones entre los grupos 1-2, 1-3, 1-4 y 2-3 no muestran evidencias de que sean diferentes, ya que como muestra la Figura, el intervalo de confianza al 95 % contiene al valor cero, correspondiente a la línea vertical discontinua. Para estos dos últimos grupos, 1-4 y 2-3, tampoco muestran evidencias de que sean diferentes en términos de las variables temp_ch2_out y el par 2-3 concretamente tampoco lo hace para la variable consum_ch2. Por último se muestra la Figura B.6, la cual se encarga de mostrar los resultados del test no paramétrico de Mann−Whitney−Wilcoxon aplicado a los cuatro grupos y las variables correspondientes a los ventiladores de la torre de enfriamiento, las variables pot_torre3_vent1 y pot_torre3_vent2. En este caso, la comparación entre los grupos 2 y 3 concluyen que los dos grupos no muestran evidencias suficientes como para rechazar la igualdad entre dichos grupos. Lo mismo ocurre para la comparación entre el grupo 1 y el grupo 2 en cuanto a la variable pot_torre3_vent2 se refiere. Por lo tanto, se concluye que en gran parte, al 95 % los cuatro grupos obtenidos de la partición manual en función de los valores

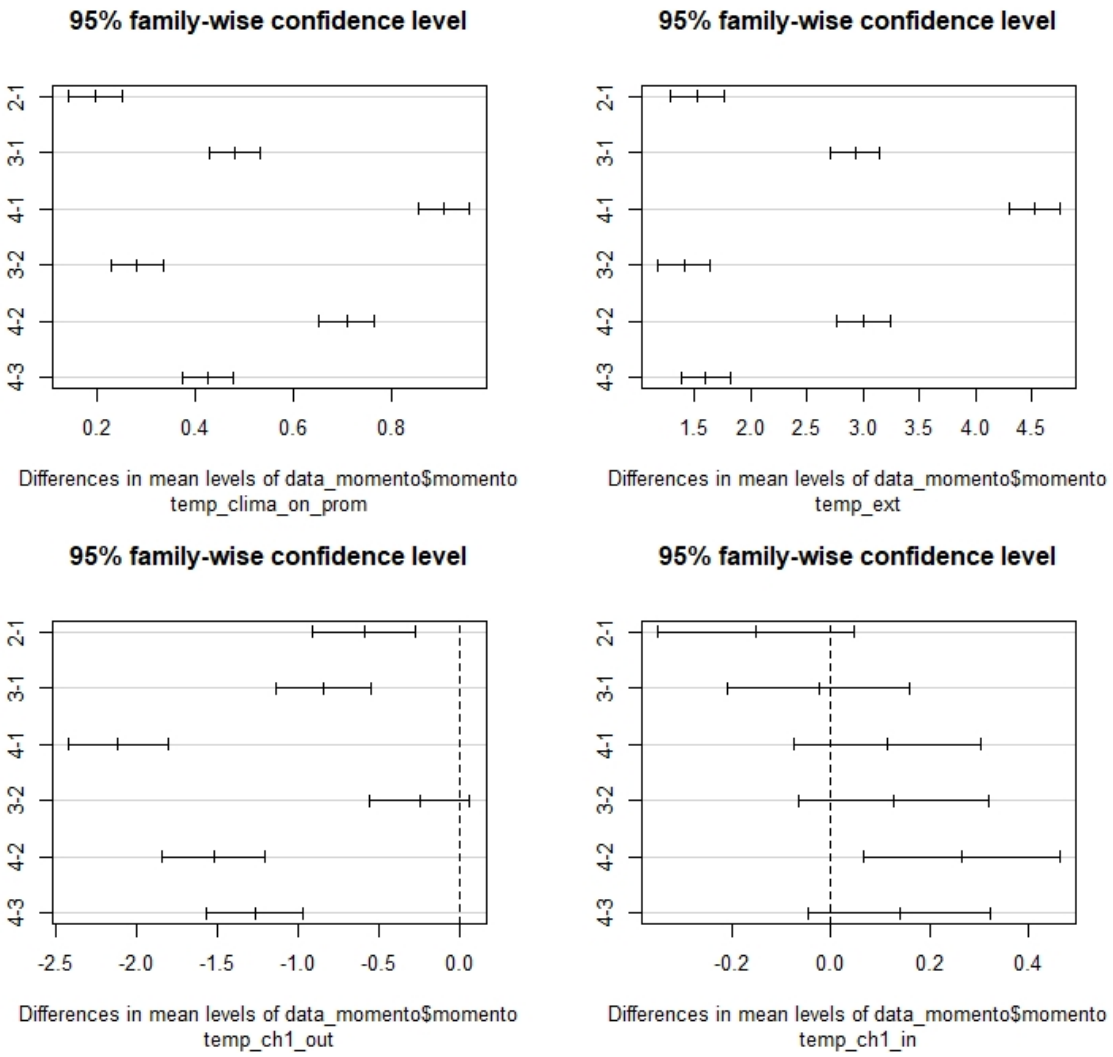


Figura B.4: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos en función de la potencia activa de forma manual.

de la potencia activa son significativamente diferentes en general, aunque los grupos 2 y 3 en gran medida parecen algo más similares.

B.2.1. Método análisis discriminante

Aplicando el contraste de hipótesis de Mann−Whitney−Wilcoxon a los datos y los 4 grupos identificados, se obtienen que los cuatro grupos muestran evidencias suficientes para un nivel de significación del 95 % para rechazar la hipótesis nula de que los clúster son diferentes por lo que a la variable consum_chilers y main_pot_activa se refiere, suceso lógico para la última de estas variables ya que el análisis discriminante se realiza a partir de la agrupación realizada previamente de forma manual en función de los valores de esta. Sin embargo, para las variables ocupacion_total y clima_on_total sí que se observan ciertos pares de clúster que contienen a la línea discontinua del cero, lo que implica que esos pares de clúster para sus respectivas variables no muestran evidencias suficientes para el rechazo de la H_0 de igualdad de los grupos, o lo que es lo mismo, que para ellas el test resulta estadísticamente no

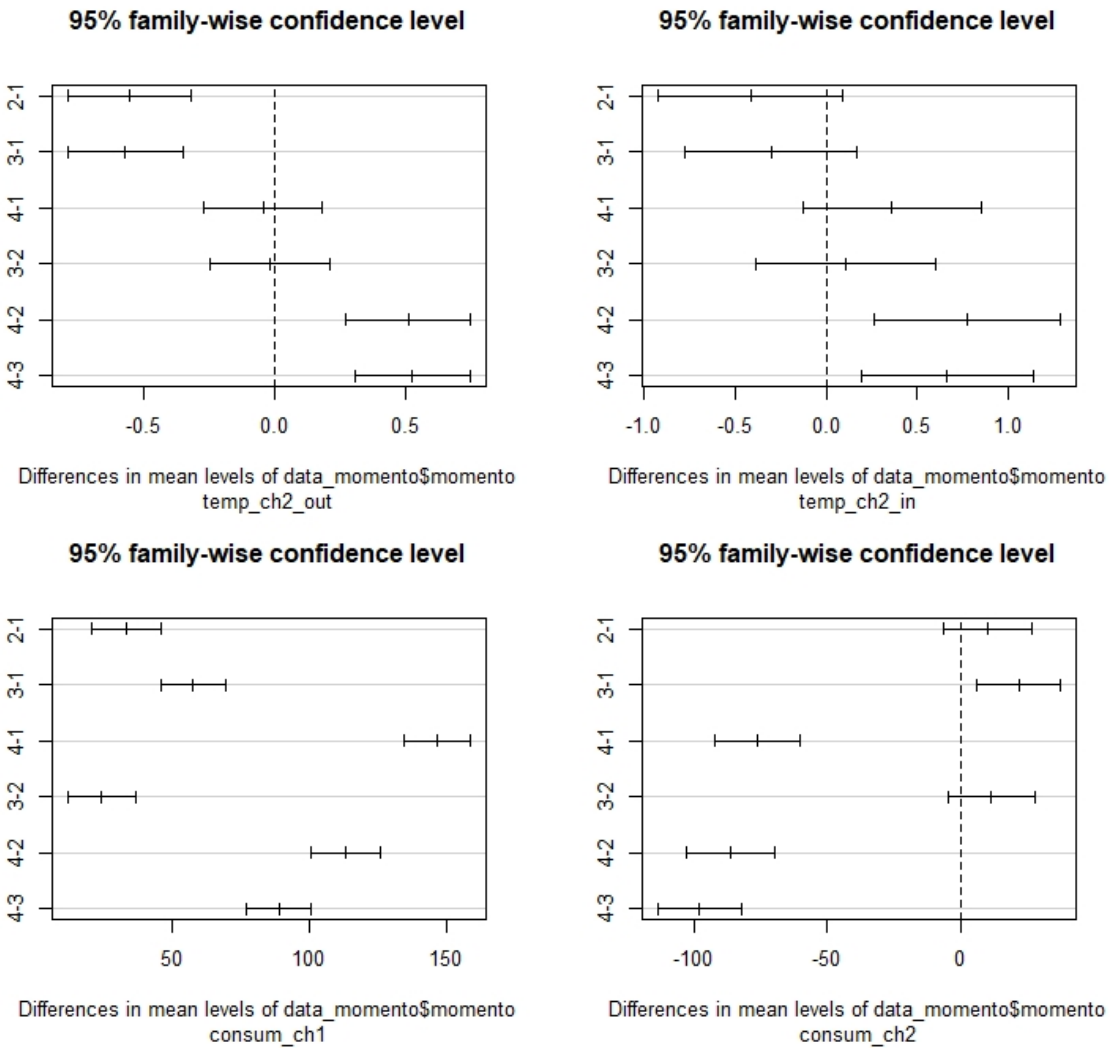


Figura B.5: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos en función de la potencia activa de forma manual.

significativo. En ambos casos, para la variable ocupacion_total y la variable clima_on_total los grupos 3 y 4 son los que para un nivel de significación del 95 % no se encuentran diferencias significativas. Esta información se recoge en la Figura B.9. En cuanto a las variables temp_clima_on_prom, temp_ext, temp_ch1_out y temp_ch1_in, se representa en la Figura B.4 los resultados correspondientes. En dicha Figura se observa rápidamente cómo en las variables temp_clima_on_prom y temp_ext ni siquiera se observa la línea discontinua vertical correspondiente al cero, lo que indica que las diferencias existentes en las agrupaciones dos a dos son realmente grandes, ya que los intervalos de confianza ni siquiera se acercan al valor cero, que implicaráa que no existen evidencias suficientes para un nivel de significación del 95 % para rechazar la hipótesis nula como resultado de un contraste estadísticamente significativo. Sin embargo, en cuanto a las variables asociadas a la entrada y la salida del chiller 1 sí que existen comparaciones dos a dos que no son estadísticamente significativas, lo que implica que no existen evidencias suficientes como para decir que los grupos son diferentes. Esto ocurre para la variable temp_ch1_out con el par de grupos 2 y 3, y para la variable temp_ch1_in con el par de grupos 1-3, 1-4 y 3-4. A continuación, en la Figura B.11 se aplica el mismo contraste, pero en este caso para las

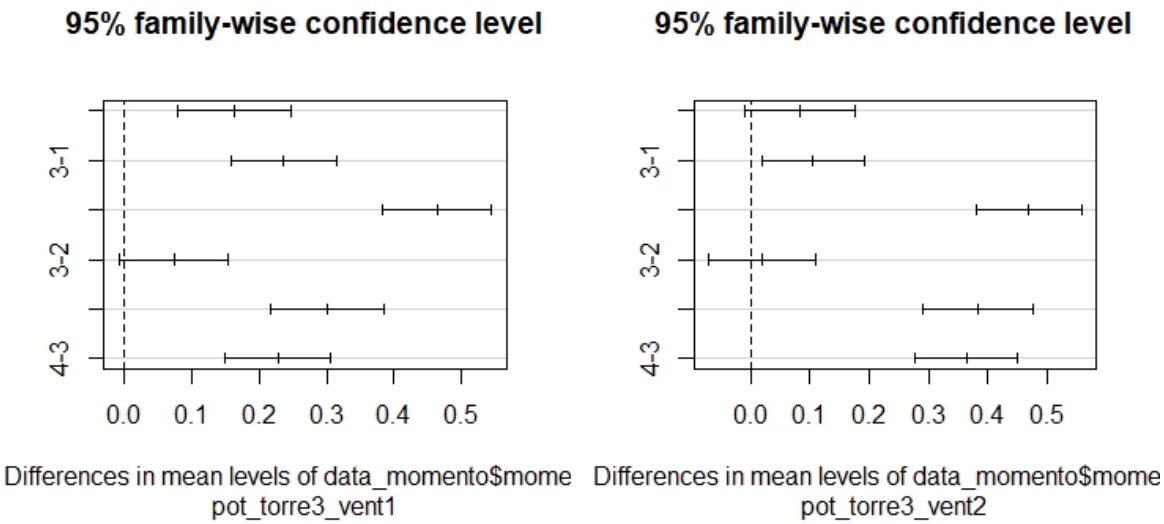


Figura B.6: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos en función de la potencia activa de forma manual.

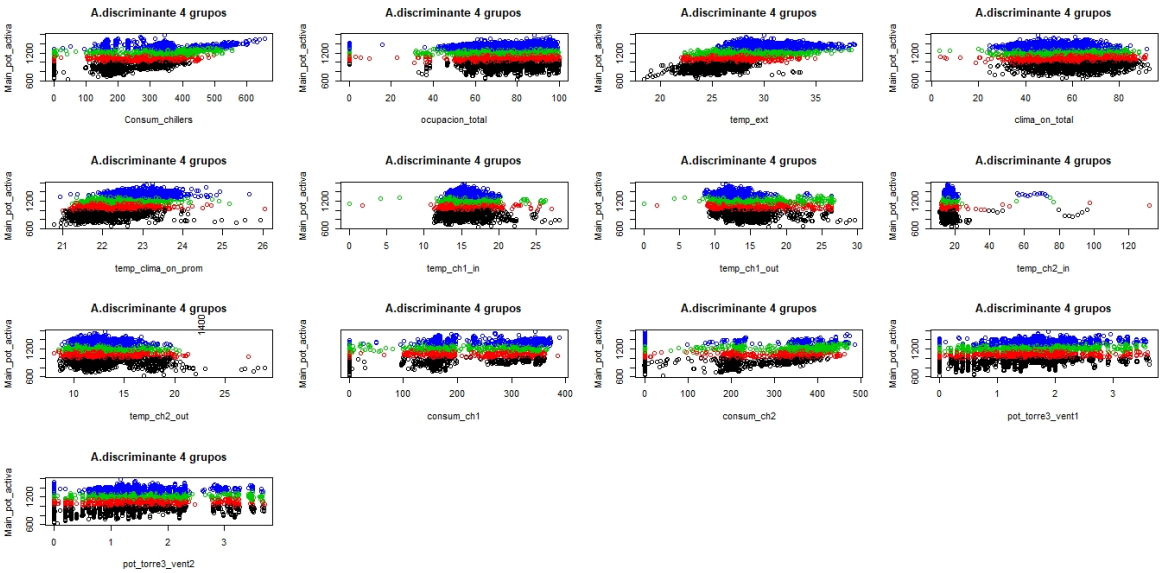


Figura B.7: Gráficos de la agrupación por el análisis discriminante en función de las diferentes variables.

variables temp_ch2_out, temp_ch2_in, consum_ch1 y consum_ch2. En este caso, la comparación entre los grupos 2 y 3 no muestran diferencias significativas en ninguna de las tres primeras variables. Por otro lado, en el caso de la variable temp_ch2_out contiene al cero la comparación realizada a partir de los grupos 1 y 4, cosa que también ocurre para la variable temp_ch2_in. Además, esta variable asociada al proceso de entrada del chiller 2 solo muestran diferencias significativas los pares de grupos 2-4 y 3-4. Por último se muestra la Figura B.12, la cual se encarga de mostrar los resultados del test no paramétrico de Mann−Whitney−Wilcoxon aplicado a los cuatro grupos y las variables correspondientes a los ventiladores de la torre de enfriamiento, las variables pot_torre3_vent1 y pot_torre3_vent2. Para estas dos variables, la comparación entre los grupos 2 y 3 contienen al cero para un nivel de confianza

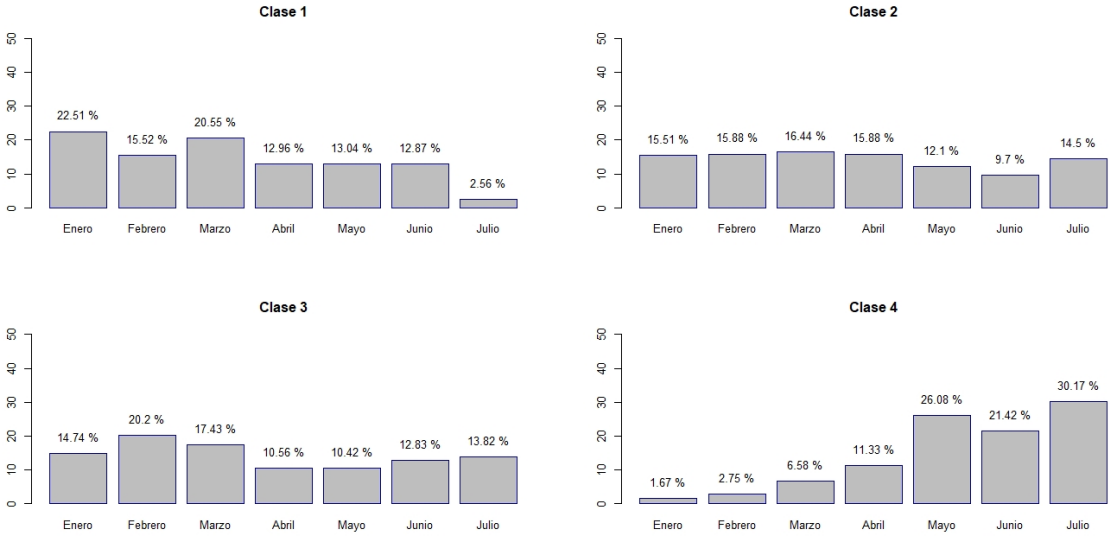


Figura B.8: Gráficos de los porcentajes de las observaciones en cada mes de la agrupación por el método discriminante.

del 95 %. En el caso del segundo ventilador, la comparación del grupo 1 con el 2 y con el 3 tampoco muestran evidencias suficientes para rechazar la hipótesis nula de igualdad de los grupos. Por lo tanto, se concluye que en gran parte, los resultados son muy similares a los obtenidos mediante el método anterior, lógicamente. Y de nuevo ocurre que los grupos 2 y 3 son aquellos que muestra una mayor similitud.

B.2.2. Método aglomerativo clúster

Como ya se introdujo previamente, los métodos jerárquicos parten de una matriz de distancia entre individuos, en este caso se emplea la distancia euclídea, y en base a ella pretenden un agrupamiento de los individuos a distintos niveles. En el nivel más bajo cada grupo estaría formado por individuos, mientras que los grupos a niveles superiores serían aquellos provenientes del resultado de agregar grupos de niveles inferiores. Los algoritmos que se emplean para crear esta jerarquéa pueden ser de dos tipos, aglomerativo o divisivo. En este caso, se emplea el método aglomerativo, que es el más empleado usualmente.

Estos métodos aglomerativos Partiendo de los individuos, se construyen grupos formados por individuos, para después construir grupos, mediante la agregación de los grupos ya formados en etapas anteriores. En este caso, en el paso tercero del algoritmo, es decir, cuando se recalculan las distancias de todos los demás grupos al nuevo grupo, formado al juntar C_i y C_j, se emplea el método del promedio.

Este método del promedio consiste en definir $d(C_k, C_s) = \frac{n_i}{n_i+n_j} d(C_k, C_i) + \frac{n_j}{n_i+n_j} d(C_k, C_j)$, siendo n_i y n_j el número de individuos en los grupos C_i y C_j, respectivamente. Lo que es equivalente a definir la distancia entre dos grupos como el promedio de las distancias entre sus individuos:

$$d(C_r, C_s) = \frac{1}{n_r n_s} \sum_{i \in C_r, j \in C_s} d_{ij} \tag{B.7}$$

Siguiendo estos procedimientos, se obtiene el siguiente dendograma, representado en la Figura B.13.

Se observa en el dendograma la partición correspondiente a realizar una partición en cuatro grupos, dicha partición puede verse en la Figura, representando el valor de corte correspondiente a la línea

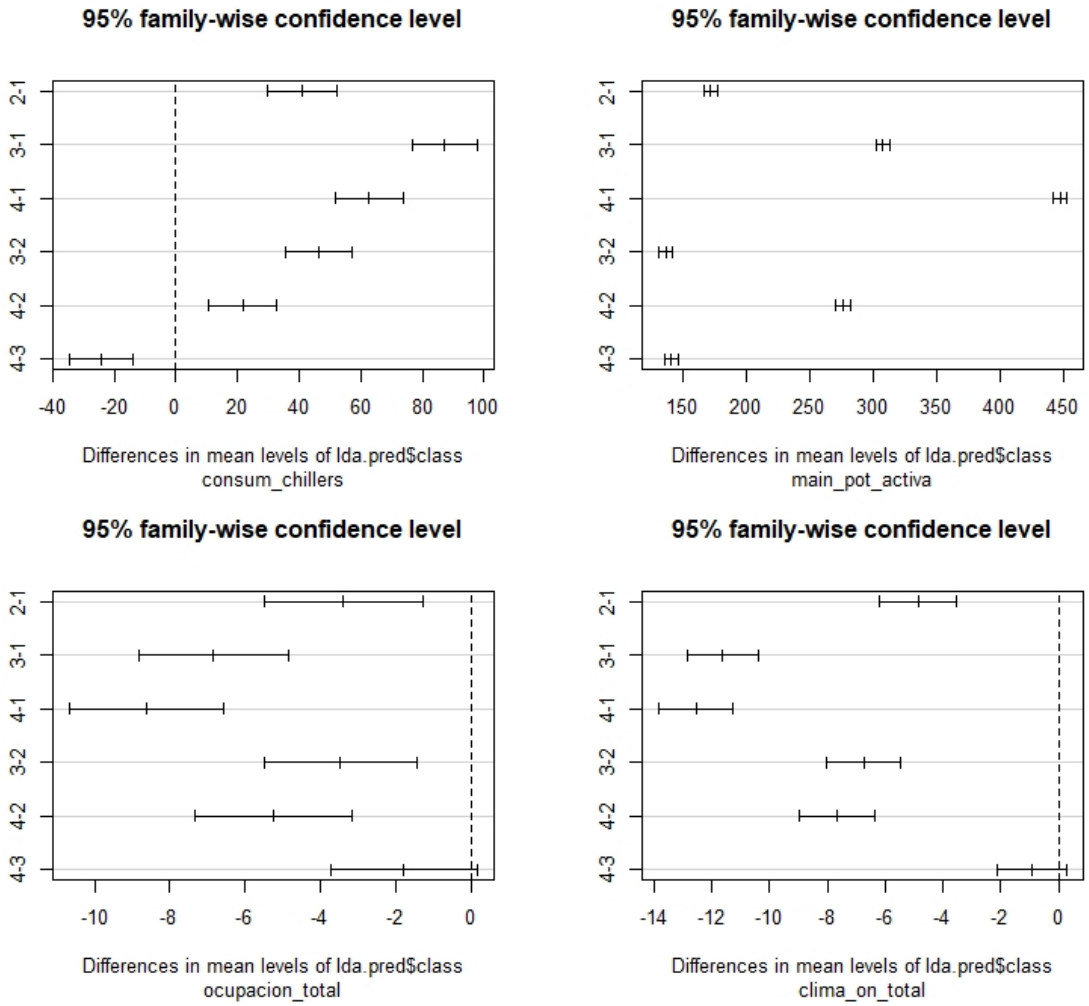


Figura B.9: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos mediante el análisis discriminante.

horizontal de color azul. Se observa que existe un grupo muy grande, y otro muy pequeño, mientras que los otros dos restantes, situados a la izquierda del dendograma son más similares.

Con el objeto de comprender la distribución de las observaciones pertenecientes a cada uno de los grupos, se muestra la Tabla B.1, donde se revela el número de observaciones perteneciente a cada uno de los clústers, así como el porcentaje en cada grupo respecto del total.

Tabla B.1: Tabla resumen del número de observaciones en cada grupo.

	Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**
Número de obs.	930	2442	219	1386
% del total	18.69 %	49.07 %	4.40 %	27.85 %

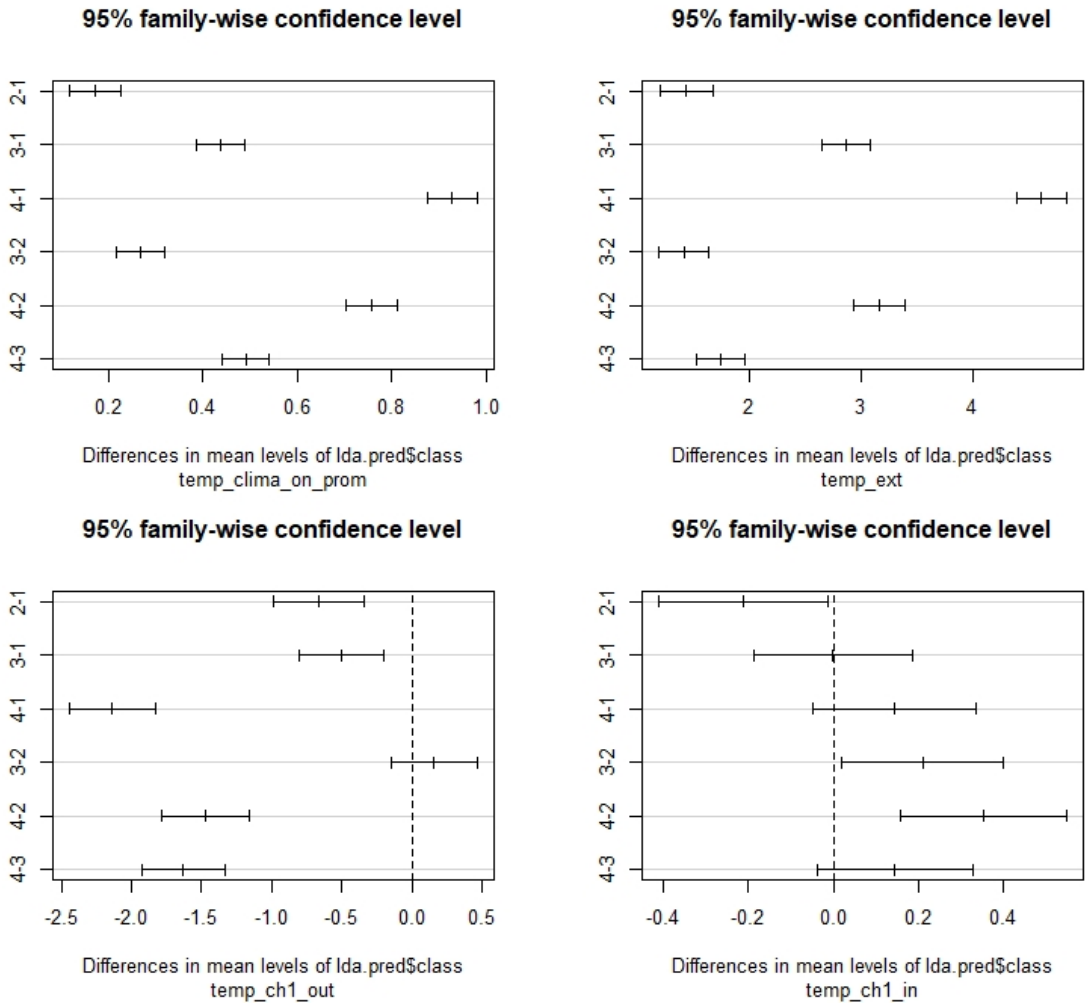


Figura B.10: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos mediante el análisis discriminante.

Es posible observar de manera más exacta lo que el dendograma dejaba ver, que el grupo 3 es realmente un grupo pequeño en comparación con el resto y que existe un grupo que contiene caso el 50 % del total de observaciones, mientras que los dos grupos restantes son similares en lo que a tamaño respecta.

Para visualizar el agrupamiento propuesto en función de los valores de las diferentes variables, se procede a graficar en la Figura B.14 para cada variable de estudio un gráfico de dispersión diferenciando el grupo por colores, con la variable main_pot_activa fija en el eje y y el resto de variables en el eje x.

Se observa cómo los grupos prácticamente se distribuyen en función de los valores de a variable main_pot_activa, es decir, se diferencian dos grupos que principalmente se posicionan en valores elevados de dicha variable y otras dos que se sitúan en valores bajos de la misma. Al mismo tiempo dentro de estos dos grupos parece existir una diferenciación entre ls grupos en función del valor del consumo de los chiller.

Finalmente, para observar de qué meses provienen los datos de cada uno de los grupos, se realiza un gráfico de barras para cada una de las agrupaciones, donde se muestran también los porcentajes del número de observaciones provenientes de cada mes en cada uno de los grupos.

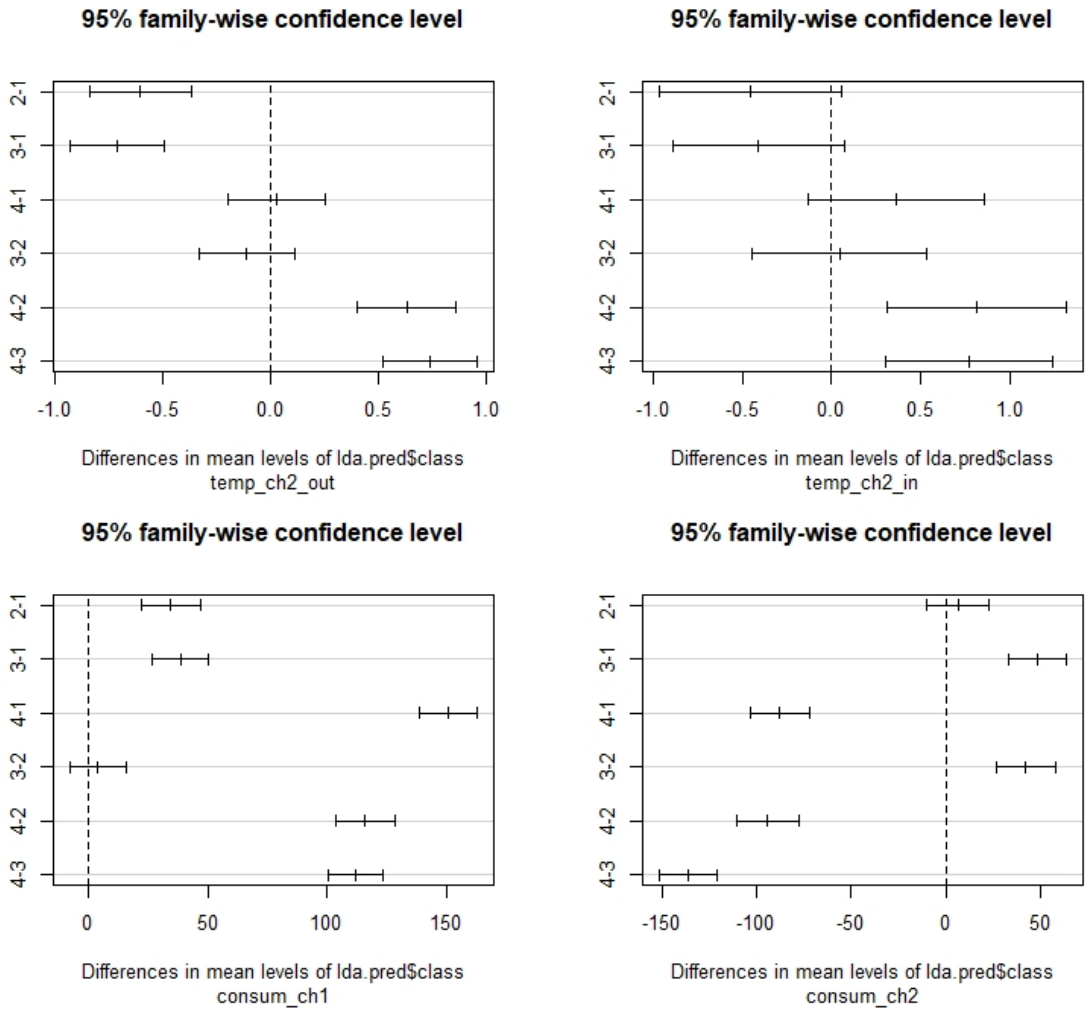


Figura B.11: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos mediante el análisis discriminante.

Como muestra la Figura B.15, se observa cómo el grupo 1 recoge prácticamente datos de los primeros meses enero-marzo y por el contrario, el grupo cuatro recoge los de los últimos meses, es decir, observaciones de mayo-julio. Sin embargo los grupos dos y tres, son algo más confusos. En el caso del grupo grande, el grupo 2 recoge datos de manera indiscriminada de todos los meses de manera bastante equitativa, siendo quizás marzo el mes que más observaciones de este grupo contiene. El grupo de observaciones minoritario, el grupo tercero, recoge muchas observaciones de enero y marzo, pero también lo hace del resto de meses, y como es un grupo pequeño, estas diferencias e el número de observaciones de cada mes no serán muy significativas.

Aplicando el contraste de hipótesis de Mann−Whitney−Wilcoxon a los datos y los 4 grupos identificados por el método aglomerativo, se obtienen que los cuatro grupo muestran evidencias suficientes para un nivel de significación del 95 % para rechazar la hipótesis nula de que los grupo son diferentes por lo que a las variables main_pot_activa y consum_chillers se refiere. No ocurre lo mismo para las variables ocupacion_total y clima_on_total, en los que si que se observan ciertos pares de grupo que contienen a la línea discontinua del cero, lo que implica que esos pares de grupo para sus respectivas variables no muestran evidencias suficientes para el rechazo de la H_0 de igualdad de los grupos, o

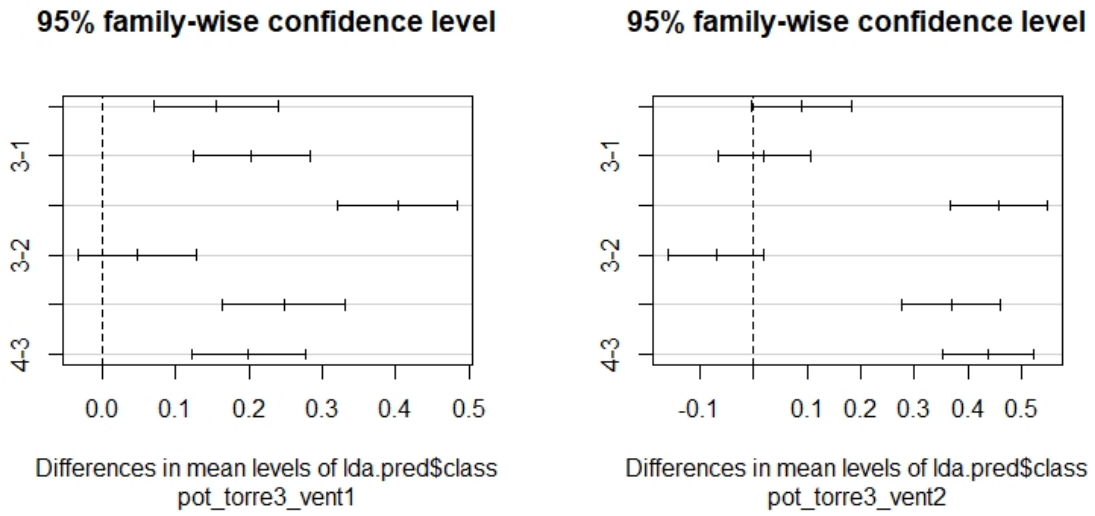


Figura B.12: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos mediante el análisis discriminante.

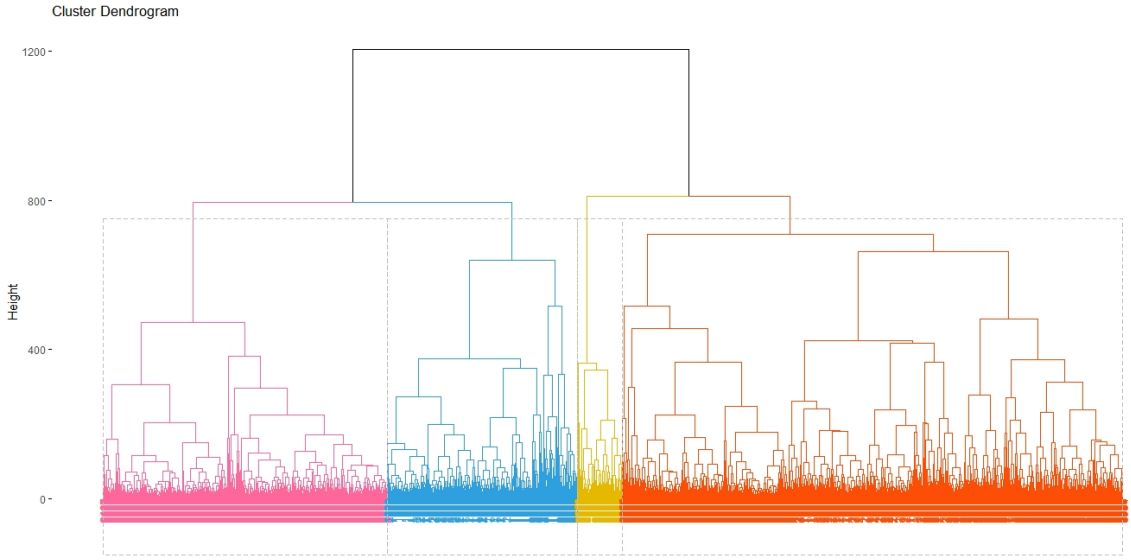


Figura B.13: Dendograma de las variables basado en la distancia euclídea.

lo que es lo mismo, que para ellas el test resulta estadísticamente no significativo. En el caso de la variable ocupacion_total los grupo 1-3 y 2-3, y en la variable clima_on_total los grupo 2 y 3 son los que para un nivel de significación del 95 % no se encuentran diferencias significativas. Esta información se recoge en la Figura B.16. En cuanto a las variables temp_clima_on_prom, temp_ext, temp_ch1_out y temp_ch1_in, se representa en la Figura B.17 los resultados correspondientes. En dicha Figura se observa rápidamente cómo para un nivel de significación del 95 % las variables temp_clima_on_prom y temp_ext muestran rechazo de la igualdad entre grupos en todos los casos. No ocurre lo mismo para las variables temp_ch1_out y temp_ch1_in, en los que si que se observan ciertos pares de grupo que contienen a la línea discontinua del cero, lo que implica que esos pares de grupo para sus respectivas

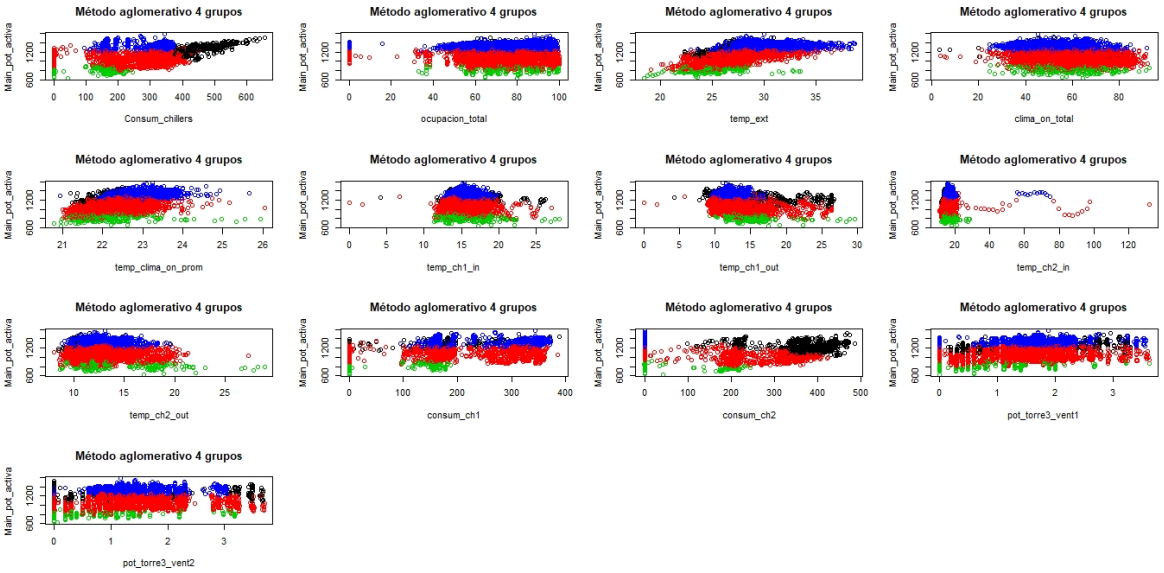


Figura B.14: Gráficos de la agrupación por el método aglomerativo en función de las diferentes variables.

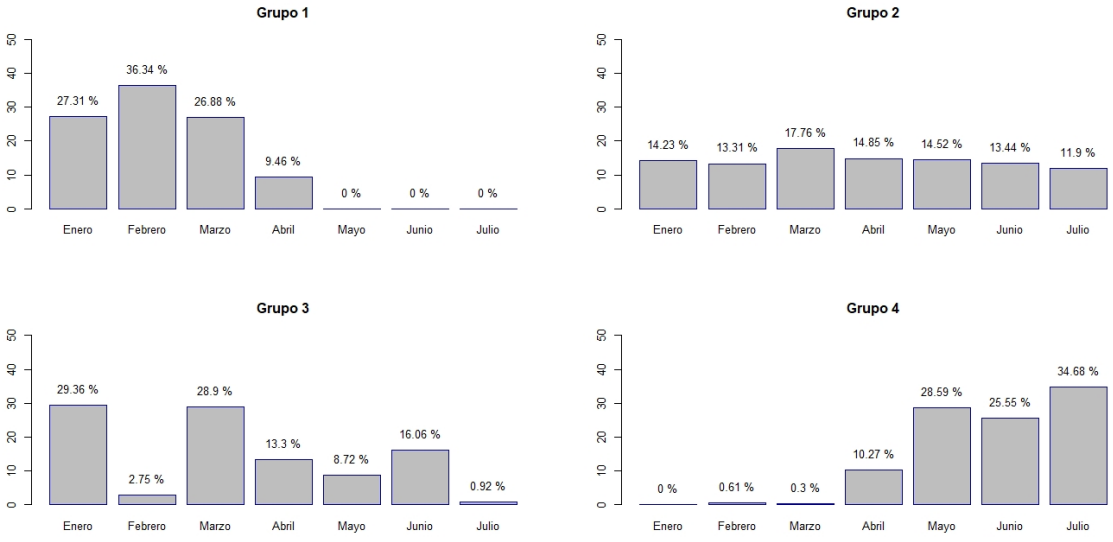


Figura B.15: Gráficos de los porcentajes de las observaciones en cada mes de la agrupación por el método aglomerativo.

variables no muestran evidencias suficientes para el rechazo de la H_0 de igualdad de los grupos, o lo que es lo mismo, que para ellas el test resulta estadísticamente no significativo. En el caso de la variable temp_ch1_out muestran igualdad los grupo 1 y 3 y en la variable temp_ch1_in los grupo 1 y 4 son los que para un nivel de significación del 95 % no se encuentran diferencias significativas. A continuación, en la Figura B.18 se aplica el mismo contraste, pero en este caso para las variables temp_ch2_out, temp_ch2_in, consum_ch1 y consum_ch2. En este caso, la gran parte de las comparaciones dos a dos son estadísticamente significativas y rechazan la igualdad entre clúster. Cosa que no ocurre para el caso particular de la comparación dos a dos entre el clúster 2-3 y 3-4 en términos de la variable temp_ch2_in,

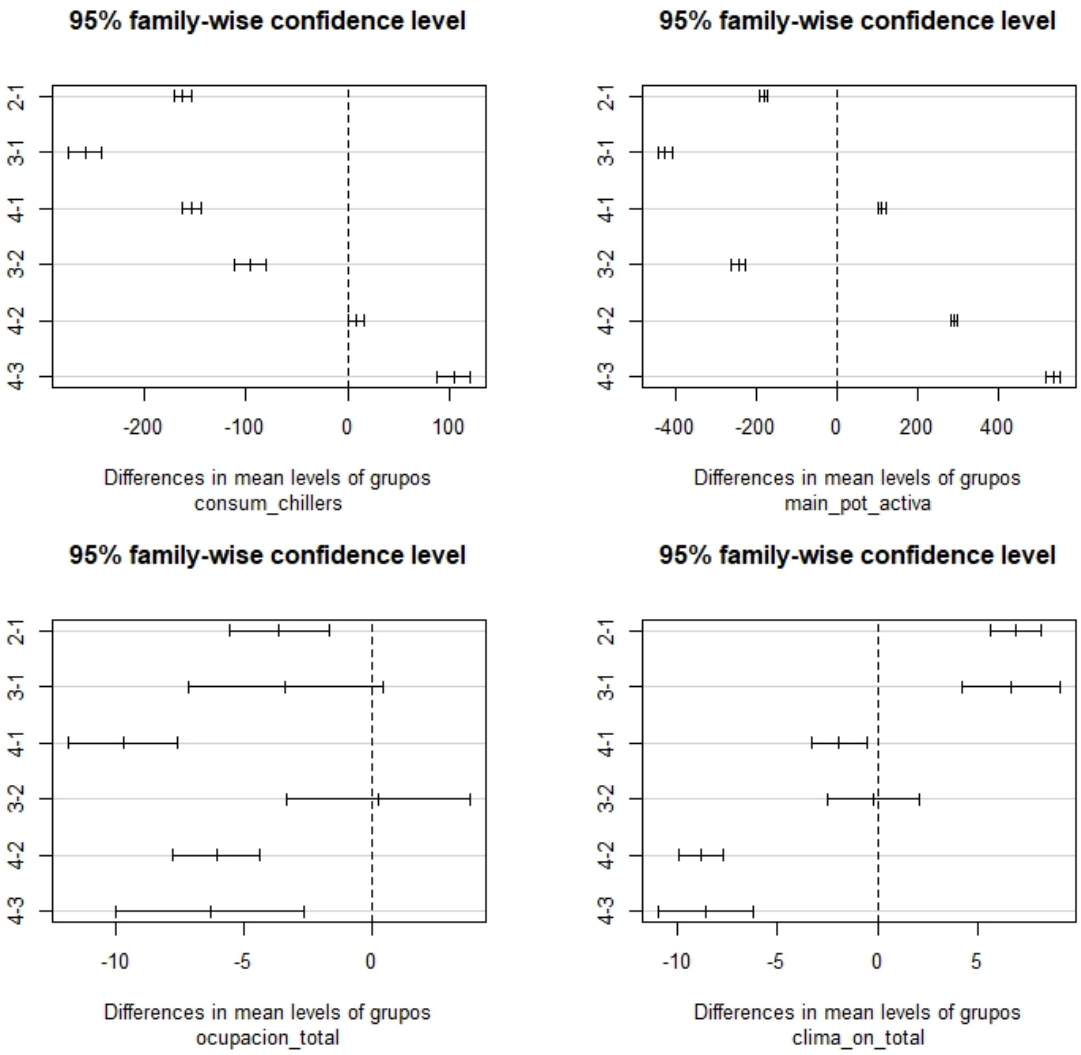


Figura B.16: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos por el método aglomerativo.

que no muestran evidencias de que sean diferentes, ya que como muestra la Figura, el intervalo de confianza al 95 % contiene al valor cero, correspondiente a la línea vertical discontinua. Nótese cómo en el gráfico correspondiente a la variable consum_ch2 ni siquiera se observa la línea discontinua vertical correspondiente al valor 0, lo que significa que los grupos son muy diferentes en lo que a esta variable se refiere, incluso podría decirse que esta variable es decisiva a la hora de definir las agrupaciones. Finalmente, se muestra la Figura B.19, la cual se encarga de mostrar los resultados del test no paramétrico de Mann−Whitney−Wilcoxon aplicado a los cuatro clúster y las variables correspondientes a los ventiladores de la torre de enfriamiento, las variables pot_torre3_vent1 y pot_torre3_vent2. En este caso, ninguna de las comparaciones dos a dos parecen concluir que los dos grupos obtenidos por el método aglomerativo sean iguales. Por lo tanto, se concluye que en gran parte, al 95 % los cuatro grupos grupo obtenidos del análisis son significativamente diferentes, pudiendo encontrar algunas similitudes entre los grupo 2 y 3 en mayor medida.

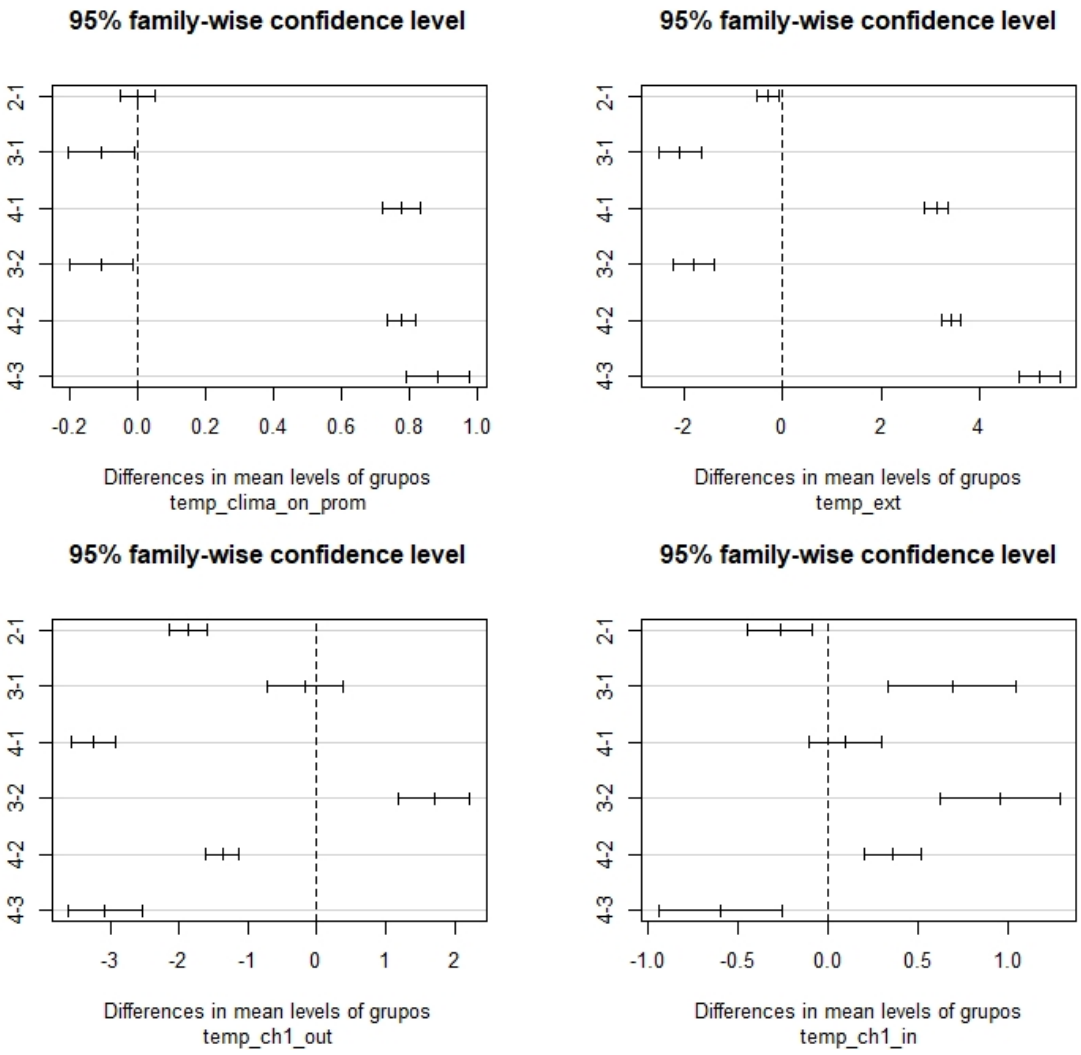


Figura B.17: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos por el método aglomerativo.

B.2.3. Método kmeans clúster

El análisis de conglomerados, o como es popularmente conocido el análisis clúster, es un método estadístico multivariante que tiene por objetivo la identificación de grupos de entre una gran cantidad de observaciones, en función de sus características. A diferencia del análisis discriminante, en este caso los grupos no se conocen a priori, por lo que tampoco el número de grupos. Cada observación responde a la forma $x_i = (x_1, x_2, \cdots, x_p), i \in (1, n)$, que en este caso cada observación se puede denotar por $x_i = (consum\_chillers, main\_pot\_activa, ocupacion\_total, clima\_on,_t otal, temp\_clima\_on\_prom, temp\_ext, temp\_ch1\_out, temp\_ch1\_in, temp\_ch2\_out, temp\_ch2\_in, consum\_ch1, consum\_ch2, 4pot\_torre3\_vent1, pot\_torre3\_vent2)$.

Para la determinación del número de grupos, se suele emplear un criterio basado en la variabilidad explicada obtenida en función del número de grupos seleccionados. En este caso, y como muestra la Figura B.20, con dos grupos ($k = 2$) se obtiene una variabilidad explicada de casi el 50 %, con tres grupos ($k = 3$) se obtiene una variabilidad explicada del 65 % y con cuatro grupos ($k = 4$) se obtiene

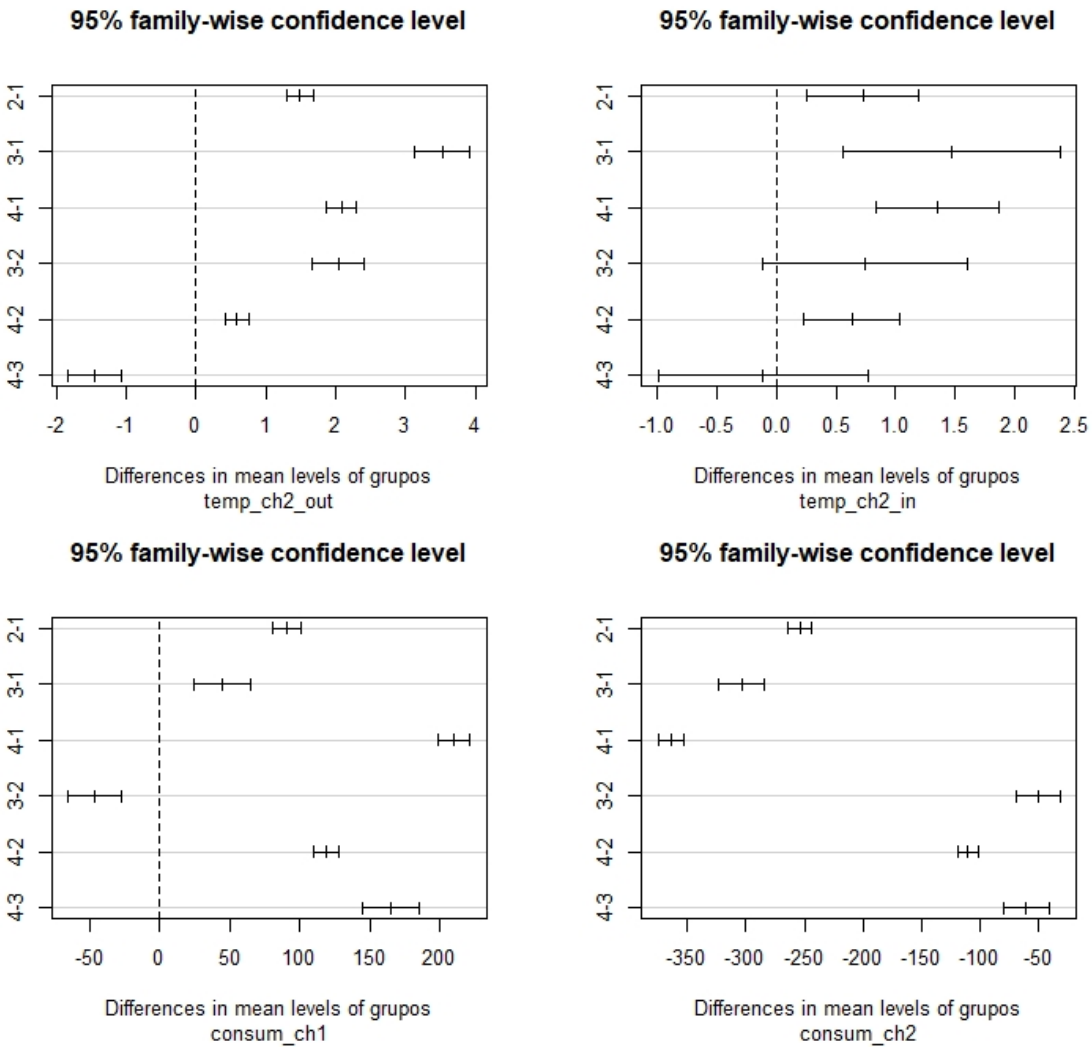


Figura B.18: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos por el método aglomerativo.

una variabilidad explicada superior al 70 %. A partir de este número de grupos ($k \geq 4$) el incremento obtenido en la variabilidad explicada por aíadir un grupo más resulta poco significante, de hecho con k superiores, como $k = 8$ se obtiene una disminución en la variabilidad explicada por aíadir ese octavo grupo adicional. Por lo tanto, se considerarán 4 grupos para el análisis clúster, ya que estos consiguen una variabilidad aceptable, superior al 70 %.

Una vez aplicado el análisis clúster, se obtienen los cuatro grupos, y para observar la forma de agrupación empleada en función de las distintas variables, se muestra la Figura B.21, que representa por colores los cuatro clústers en gráficos donde la variable del eje de ordenadas es fija, la variable main_pot_activa, ya que es una de las más relevantes, y en el eje de abscisas, se propone el resto de variables de estudio.

En esta Figura, se observa cómo los grupos prácticamente se distribuyen en función de los valores de a variable main_pot_activa, es decir, se diferencian dos grupos que principalmente se posicionan en valores elevados de dicha variable y otras dos que se sitúan en valores bajos de la misma. Al mismo tiempo dentro de estos dos grupos parece existir una diferenciación entre ls grupos en función del valor

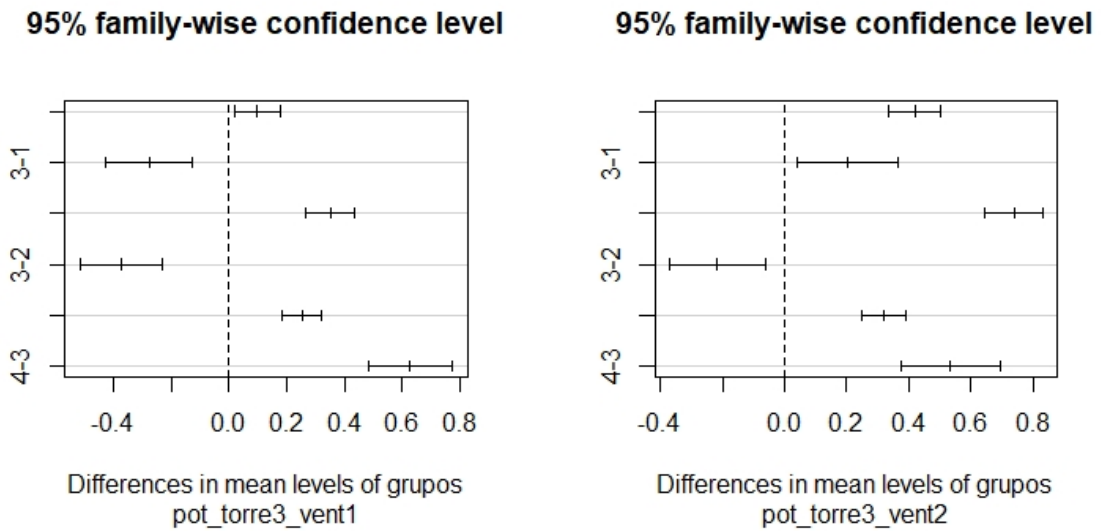


Figura B.19: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos obtenidos por el método aglomerativo.

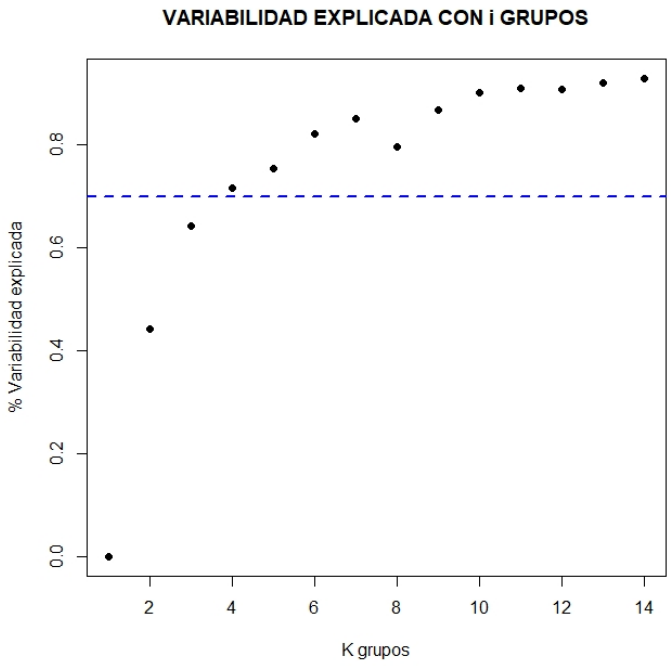


Figura B.20: Gráfico de la variabilidad explicada obtenida en función del número de grupos K.

del consumo de los chiller. Estos son resultados algo lógicos, ya que era de esperar que las variables más relevantes fueran importantes a la hora de obtener los grupos.

Finalmente, para comprender la distribución de las observaciones pertenecientes a cada uno de los grupos, se muestra la Tabla B.2, donde se revela el número de observaciones perteneciente a cada uno de los clústers, así como el porcentaje en cada grupo respecto del total.

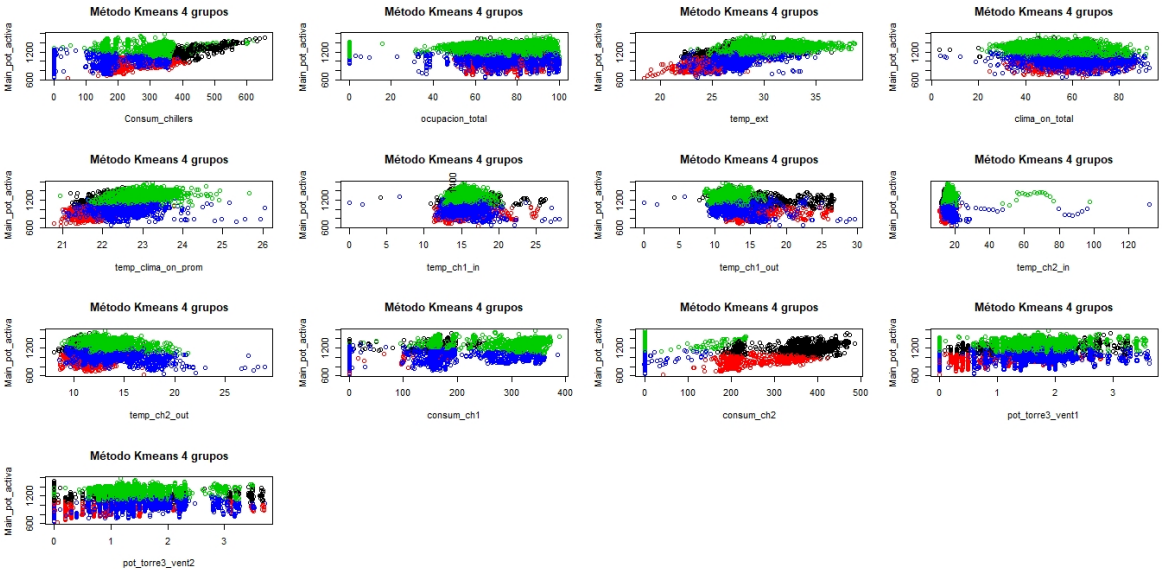


Figura B.21: Gráficos de la agrupación por el método clúster kmeans en función de las diferentes variables.

Tabla B.2: Tabla resumen del número de observaciones en cada clúster.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Número de obs.	1219	959	1510	1289
% del total	24.49 %	19.27 %	30.34 %	25.90 %

Se observa que son bastante semejantes en tamaño, aunque el clúster 2 se observa que se compone de un menor número de observaciones en comparación con el resto. Para seguir observando el origen de las observaciones de esta agrupación, se muestra la Figura B.22, donde se muestran los porcentajes de observaciones de clúster provenientes de la toma de datos de cada uno de los meses.

Se observa cómo parece que cada clúster toma observaciones de ciertos meses, bien diferenciados. El clúster 1 se centra en observaciones de enero a marzo, el clúster 2 en observaciones de mayo julio, el clúster 3 de abril-junio y finalmente el clúster 4 de enero a marzo.

Al mismo tiempo, esto denota que sí que existen diferencias en función de los meses, donde también se ve una clara diferenciación en el comportamiento del sistema en los dos periodos enero-abril y abril-julio. Se concluye por tanto también, que en abril se produce un cambio significativo en a distribución de los datos observados.

Aplicando el contraste de hipótesis de Mann−Whitney−Wilcoxon a los datos y los 4 clústers identificados, se obtienen que los cuatro clúster muestran evidencias suficientes para un nivel de significación del 95 % para rechazar la hipótesis nula de que los clúster son diferentes por lo que a la variable consum_chilers se refiere. No ocurre lo mismo para las variables main_pot_activa, ocupacion_total y clima_on_total, en los que si que se observan ciertos pares de clúster que contienen a la línea discontinua del cero, lo que implica que esos pares de clúster para sus respectivas variables no muestran evidencias suficientes para el rechazo de la H_0 de igualdad de los grupos, o lo que es lo mismo, que para ellas el test resulta estadísticamente no significativo. En el caso de main_pot_activa los clúster 2 y 3, en la variable ocupacion_total los clúster 3 y 4 y en la variable clima_on_total los clúster 2 y 3 son los

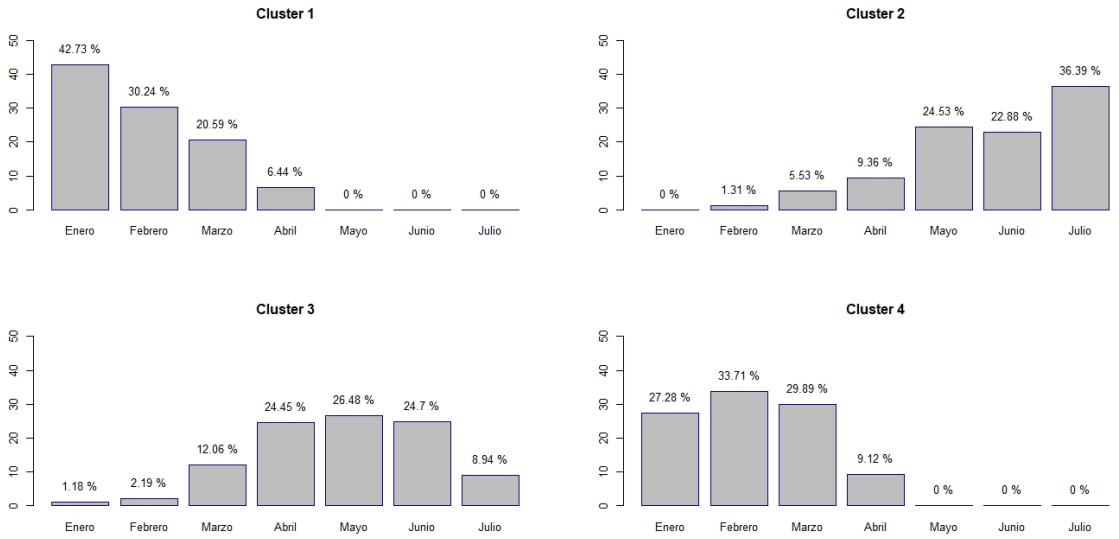


Figura B.22: Gráficos de los porcentajes de las observaciones en cada mes de la agrupación por el método kmeans.

que para un nivel de significación del 95 % no se encuentran diferencias significativas. Esta información se recoge en la Figura B.23. En cuanto a las variables temp_clima_on_prom, temp_ext, temp_ch1_out y temp_ch1_in, se representa en la Figura B.24 los resultados correspondientes. En dicha Figura se observa rápidamente cómo solamente se encuentran igualdad entre clúster en la variable temp_ch1_in, donde los clúster 1-3, 1-4 y 3-4 no muestran evidencias suficientes para un nivel de significación del 95 % para rechazar la igualdad entre grupos. A continuación, en la Figura B.25 se aplica el mismo contraste, pero en este caso para las variables temp_ch2_out, temp_ch2_in, consum_ch1 y consum_ch2. En este caso, la gran parte de las comparaciones dos a dos son estadísticamente significativas y rechazan la igualdad entre clúster. Cosa que no ocurre para el caso particular de la comparación dos a dos entre el clúster 3 y 4 en términos de la variable temp_ch2_in, que no muestran evidencias de que sean diferentes, ya que como muestra la Figura, el intervalo de confianza al 95 % contiene al valor cero, correspondiente a la línea vertical discontinua. Por último se muestra la Figura B.26, la cual se encarga de mostrar los resultados del test no paramétrico de Mann−Whitney−Wilcoxon aplicado a los cuatro clúster y las variables correspondientes a los ventiladores de la torre de enfriamiento, las variables pot_torre3_vent1 y pot_torre3_vent2. En este caso, ninguna de las comparaciones dos a dos parecen concluir que los dos clúster comparados sean iguales. Por lo tanto, se concluye que en gran parte, al 95 % los cuatro grupos clúster obtenidos del análisis son significativamente diferentes, pudiendo encontrar algunas similitudes en cuanto a las temperaturas de entrada a los chiller sobre todo, y entre los clúster 3 y 4 en mayor medida.

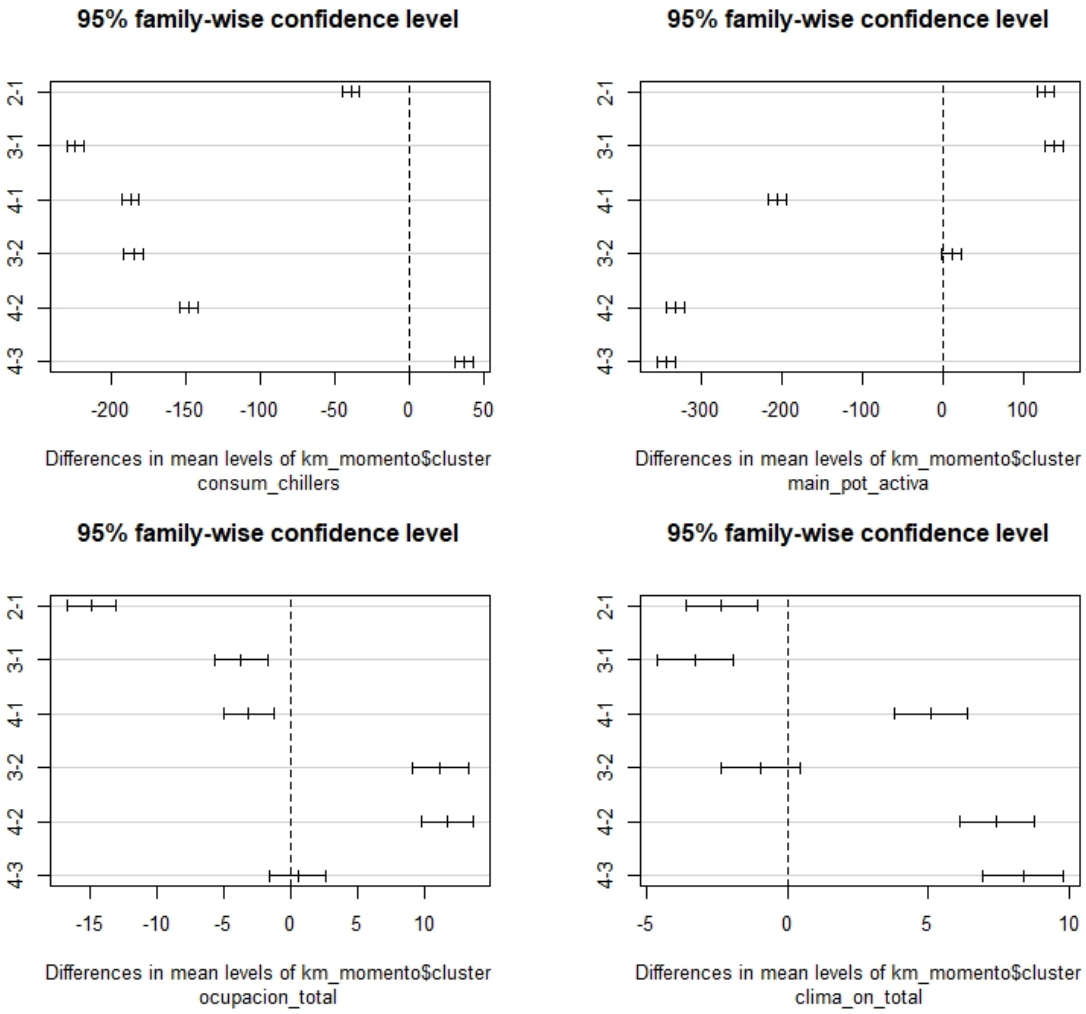


Figura B.23: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos clúster.

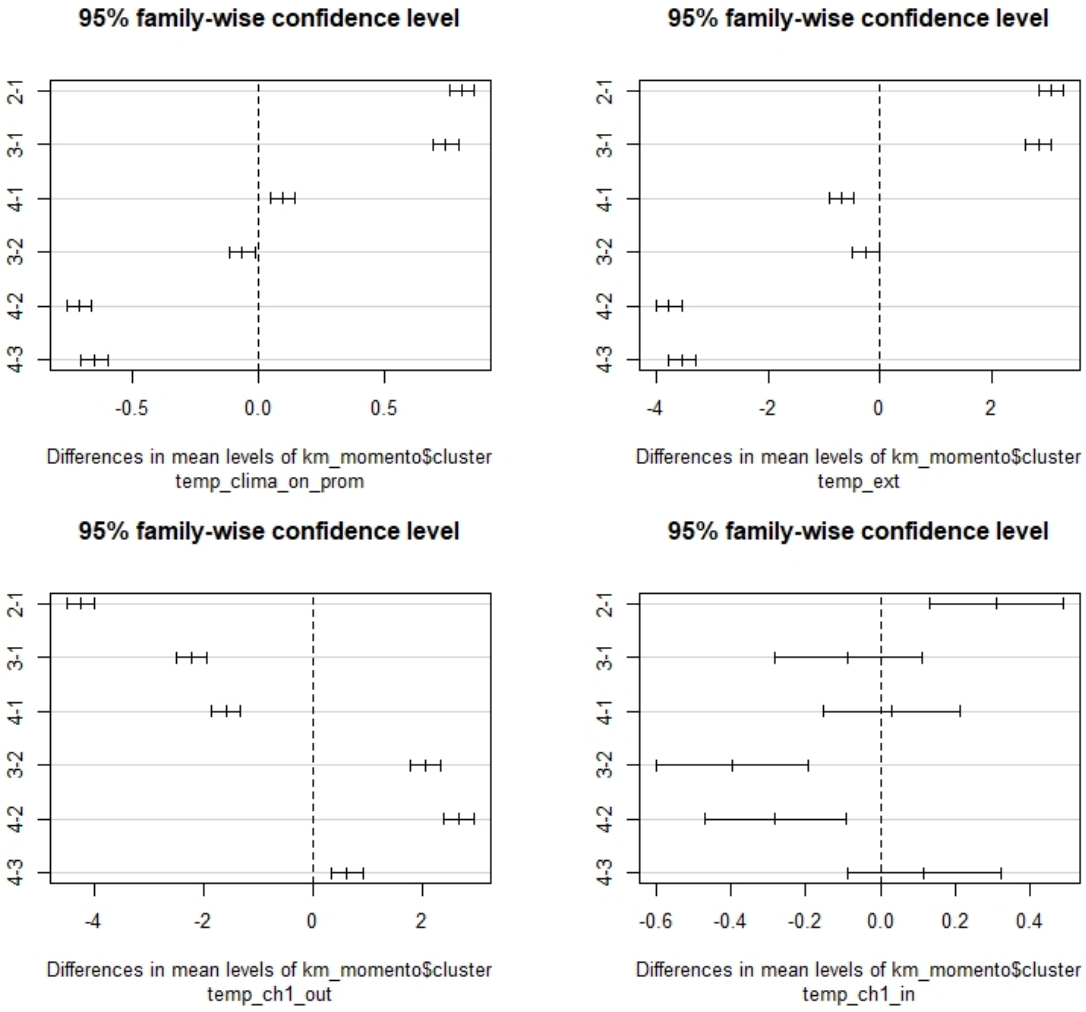


Figura B.24: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos clúster.

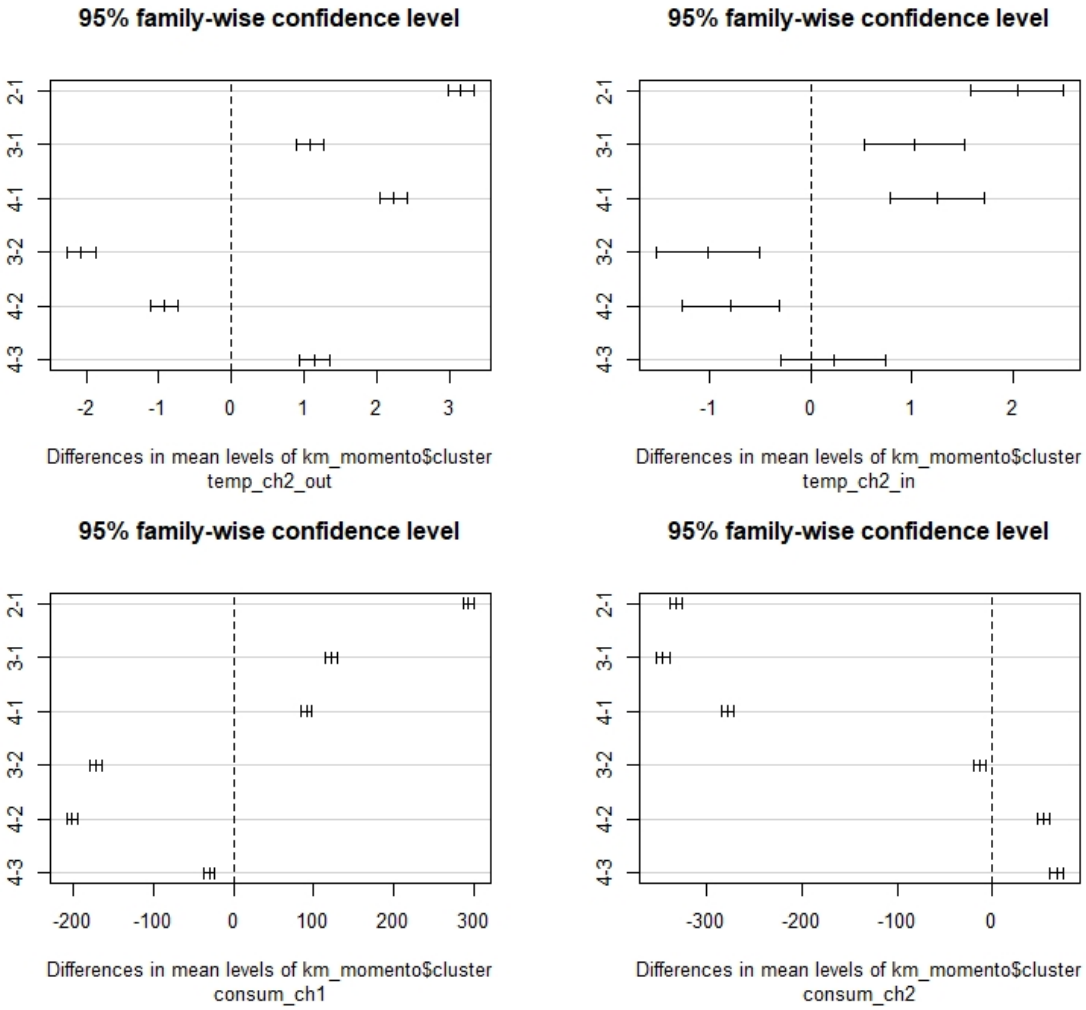


Figura B.25: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos clúster.

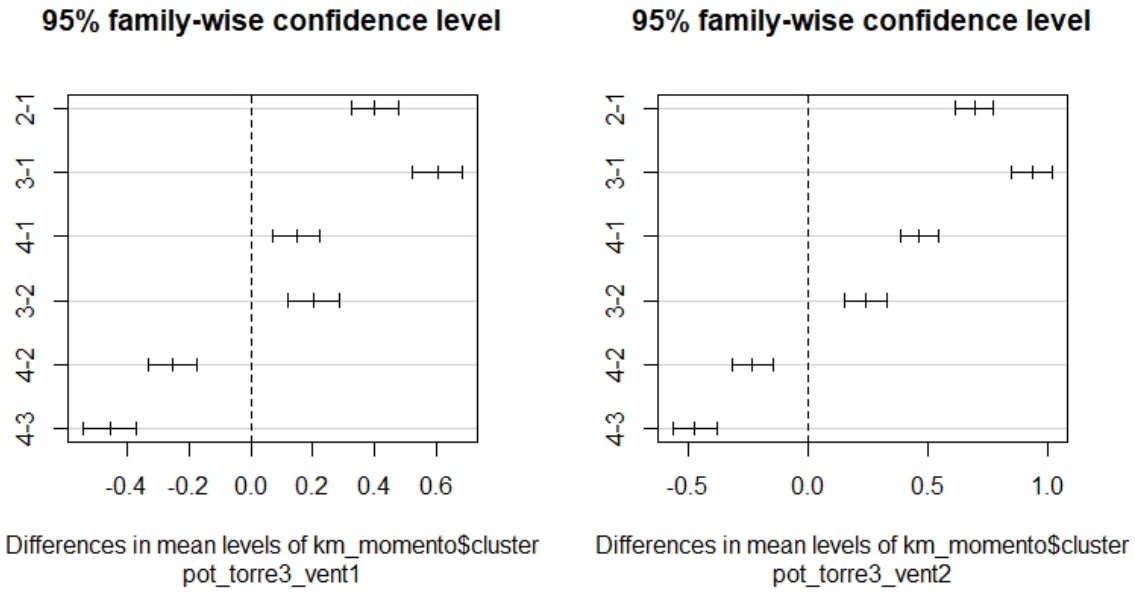


Figura B.26: Gráficos referente al test de Mann−Whitney−Wilcoxon para los grupos clúster.

Anexo C

Código R

Análisis exploratorio

```
# Número de observaciones y variables:
n<-nrow(data); n
p<-ncol(data); p

# Limpieza de datos:
data_complete<-data[complete.cases(data),]
n<-nrow(data_complete); n

# Creación de una nueva variable "mes":
library(lubridate)
library(ggplot2)
mes<-month(data_complete$date)
mes<-factor(mes,labels=c('Enero','Febrero','Marzo','Abril',
'Mayo','Junio','Julio'))
data_complete[,ncol(data_complete)+1]<-mes
colnames(data_complete)[ncol(data_complete)]<-'mes'

# Creación de una nueva variable "id.mes":
id.mes<-factor(month(data_complete$date))
data_complete[,ncol(data_complete)+1]<-id.mes
colnames(data_complete)[ncol(data_complete)]<-'id.mes'
dia<-factor(day(data_complete$date))
data_complete[,ncol(data_complete)+1]<-dia
colnames(data_complete)[ncol(data_complete)]<-'dia'

# Creación de una nueva variable "hora":
hora<-hour(hms(data_complete$time))
data_complete[,ncol(data_complete)+1]<-hora
colnames(data_complete)[ncol(data_complete)]<-'hora'

# Creación de una nueva variable "id":
id<-c(rep(c(1,2,3,4,5),995),c(1,2,3))
id<-factor(id)
data_complete[,ncol(data_complete)+1]<-id
```

```
colnames(data_complete)[ncol(data_complete)]<-'id'

# Creación de las nuevas variables "propor_ch1" y
"propor_ch2" según la ocupación:
propor_ch1<-(data_complete$consum_ch1/
data_complete$consum_chillers)*100
propor_ch2<-(data_complete$consum_ch2/
data_complete$consum_chillers)*100
data_complete[,ncol(data_complete)+1]<-propor_ch1
colnames(data_complete)[ncol(data_complete)]<-'propor_ch1'
data_complete[,ncol(data_complete)+1]<-propor_ch2
colnames(data_complete)[ncol(data_complete)]<-'propor_ch2'

# Creación de data_simple:
data_simple<-data_complete[,c(3,4,5,6,7,8,9,10,11,12,19,20)]
data_simple<-data_simple[complete.cases(data_simple),]

# ANALISIS POR MESES Y GENERAL:
# Enero:
enero<-subset(data_complete,data_complete$mes=='Enero')
data.frame('Ocupacion (%)'=mean(enero$ocupacion_total),
'Consumo Total (kW)'=sum(enero$main_pot_activa),
'Consumo Chillers (kW)'=sum(enero$consum_chillers),
'Temp. media total (ºC)'=mean(c(enero$temp_ch1_in,
enero$temp_ch1_out,
enero$temp_ch2_in,enero$temp_ch2_out)),
'Temp. clima on (ºC)'=mean(enero$temp_clima_on_prom),
'Temp. exterior (ºC)'=mean(enero$temp_ext))
pie(c(round(c(mean(enero$propor_ch1,na.rm=T),
mean(enero$propor_ch2,na.rm=T)),2)), col=c('orange','yellow'),
labels=paste(c('CH1','CH2'),
round(c(mean(enero$propor_ch1,na.rm=T),
mean(enero$propor_ch2,na.rm=T)),2),c('%','%')),
main=paste('Ocupación (%)','Enero'))

# Febrero:
febrero<-subset(data_complete,data_complete$mes=='Febrero')
nrow(febrero)
data.frame('Ocupacion (%)'=mean(febrero$ocupacion_total),
'Consumo Total (kW)'=sum(febrero$main_pot_activa),
'Consumo Chillers (kW)'=sum(febrero$consum_chillers)
,'Temp. media total (ºC)'=mean(c(febrero$temp_ch1_in,
febrero$temp_ch1_out,febrero$temp_ch2_in,
febrero$temp_ch2_out)),
'Temp. clima on (ºC)'=mean(febrero$temp_clima_on_prom),
'Temp. exterior (ºC)'=mean(febrero$temp_ext))
pie(c(round(c(mean(febrero$propor_ch1,na.rm=T),
mean(febrero$propor_ch2,na.rm=T)),2)), col=c('orange','yellow'),
labels=paste(c('CH1','CH2'),
round(c(mean(febrero$propor_ch1,na.rm=T),
mean(febrero$propor_ch2,na.rm=T)),2),c('%','%')),
```

```
main=paste('Ocupación␣(%)','Febrero'))

# Marzo:
marzo<-subset(data_complete,data_complete$mes=='Marzo')
nrow(marzo)
data.frame('Ocupacion␣(%)'=mean(marzo$ocupacion_total),
'Consumo␣Total␣(kW)'=sum(marzo$main_pot_activa),
'Consumo␣Chillers␣(kW)'=sum(marzo$consum_chillers)
,'Temp.␣media␣total␣(ºC)'=mean(c(marzo$temp_ch1_in,
marzo$temp_ch1_out,     marzo$temp_ch2_in,marzo$temp_ch2_out)),
'Temp.␣clima␣on␣(ºC)'=mean(marzo$temp_clima_on_prom),
'Temp.␣exterior␣(ºC)'=mean(marzo$temp_ext))
pie(c(round(c(mean(marzo$propor_ch1,na.rm=T),
mean(marzo$propor_ch2,na.rm=T)),2)),
col=c('orange','yellow'),labels=paste(c('CH1','CH2'),
round(c(mean(marzo$propor_ch1,na.rm=T),
mean(marzo$propor_ch2,na.rm=T)),2),c('%','%')),
main=paste('Ocupación␣(%)','Marzo'))

# Abril:
abril<-subset(data_complete,data_complete$mes=='Abril')
nrow(abril)
data.frame('Ocupacion␣(%)'=mean(abril$ocupacion_total),
'Consumo␣Total␣(kW)'=sum(abril$main_pot_activa),
'Consumo␣Chillers␣(kW)'=sum(abril$consum_chillers)
,'Temp.␣media␣total␣(ºC)'=mean(c(abril$temp_ch1_in,
abril$temp_ch1_out,     abril$temp_ch2_in,abril$temp_ch2_out)),
'Temp.␣clima␣on␣(ºC)'=mean(abril$temp_clima_on_prom),
'Temp.␣exterior␣(ºC)'=mean(abril$temp_ext))
pie(c(round(c(mean(abril$propor_ch1,na.rm=T),
mean(abril$propor_ch2,na.rm=T)),2)), col=c('orange','yellow'),
labels=paste(c('CH1','CH2'),
round(c(mean(abril$propor_ch1,na.rm=T),
mean(abril$propor_ch2,na.rm=T)),2),c('%','%')),
main=paste('Ocupación␣(%)','Abril'))

# Mayo:
mayo<-subset(data_complete,data_complete$mes=='Mayo')
nrow(mayo)
data.frame('Ocupacion␣(%)'=mean(mayo$ocupacion_total),
'Consumo␣Total␣(kW)'=sum(mayo$main_pot_activa),
'Consumo␣Chillers␣(kW)'=sum(mayo$consum_chillers)
,'Temp.␣media␣total␣(ºC)'=mean(c(mayo$temp_ch1_in,
mayo$temp_ch1_out,mayo$temp_ch2_in,mayo$temp_ch2_out)),
'Temp.␣clima␣on␣(ºC)'=mean(mayo$temp_clima_on_prom),
'Temp.␣exterior␣(ºC)'=mean(mayo$temp_ext))
pie(c(round(c(mean(mayo$propor_ch1,na.rm=T),
mean(mayo$propor_ch2,na.rm=T)),2)), col=c('orange','yellow'),
labels=paste(c('CH1','CH2'),
round(c(mean(mayo$propor_ch1,na.rm=T),
mean(mayo$propor_ch2,na.rm=T)),2),c('%','%')),
```

```
main=paste('Ocupación␣(%)','Mayo'))

# Junio:
junio<-subset(data_complete,data_complete$mes=='Junio')
nrow(junio)
data.frame('Ocupacion␣(%)'=mean(junio$ocupacion_total),
'Consumo␣Total␣(kW)'=sum(junio$main_pot_activa),
'Consumo␣Chillers␣(kW)'=sum(junio$consum_chillers)
,'Temp.␣media␣total␣(ºC)'=mean(c(junio$temp_ch1_in,
junio$temp_ch1_out,     junio$temp_ch2_in,junio$temp_ch2_out)),
'Temp.␣clima␣on␣(ºC)'=mean(junio$temp_clima_on_prom),
'Temp.␣exterior␣(ºC)'=mean(junio$temp_ext))
pie(c(round(c(mean(junio$propor_ch1,na.rm=T),
mean(junio$propor_ch2,na.rm=T)),2)), col=c('orange','yellow'),
labels=paste(c('CH1','CH2'),
round(c(mean(junio$propor_ch1,na.rm=T),
mean(junio$propor_ch2,na.rm=T)),2),c('%','%')),
main=paste('Ocupación␣(%)','Junio'))

# Julio:
julio<-subset(data_complete,data_complete$mes=='Julio')
nrow(julio)
data.frame('Ocupacion␣(%)'=mean(julio$ocupacion_total),
'Consumo␣Total␣(kW)'=sum(julio$main_pot_activa),
'Consumo␣Chillers␣(kW)'=sum(julio$consum_chillers)
,'Temp.␣media␣total␣(ºC)'=mean(c(julio$temp_ch1_in,
julio$temp_ch1_out,     julio$temp_ch2_in,julio$temp_ch2_out)),
'Temp.␣clima␣on␣(ºC)'=mean(julio$temp_clima_on_prom),
'Temp.␣exterior␣(ºC)'=mean(julio$temp_ext))
pie(c(round(c(mean(julio$propor_ch1,na.rm=T),
mean(julio$propor_ch2,na.rm=T)),2)), col=c('orange','yellow'),
labels=paste(c('CH1','CH2'),
round(c(mean(julio$propor_ch1,na.rm=T),
mean(julio$propor_ch2,na.rm=T)),2),c('%','%')),
main=paste('Ocupación␣(%)','Julio'))

# Cálculo de medias y medianas por meses:
medias_enero<-colMeans(enero[,c(3:20,26,27)],na.rm = T)
medias_febrero<-colMeans(febrero[,c(3:20,26,27)])
medias_marzo<-colMeans(marzo[,c(3:20,26,27)])
medias_abril<-colMeans(abril[,c(3:20,26,27)])
medias_mayo<-colMeans(mayo[,c(3:20,26,27)])
medias_junio<-colMeans(junio[,c(3:20,26,27)])
medias_julio<-colMeans(julio[,c(3:20,26,27)])
medias<-cbind(medias_enero,medias_febrero,
medias_marzo,medias_abril,medias_mayo,medias_junio,
medias_julio)
medianas_enero<-apply(enero[,c(3:20,26,27)],2,median)
medianas_febrero<-apply(febrero[,c(3:20,26,27)],2,median)
medianas_marzo<-apply(marzo[,c(3:20,26,27)],2,median)
medianas_abril<-apply(abril[,c(3:20,26,27)],2,median)
```

```
medianas_mayo<-apply(mayo[,c(3:20,26,27)],2,median)
medianas_junio<-apply(junio[,c(3:20,26,27)],2,median)
medianas_julio<-apply(julio[,c(3:20,26,27)],2,median)
medianas<-cbind(medianas_enero,medianas_febrero,
medianas_marzo,medianas_abril,medianas_mayo,
medianas_junio,medianas_julio)

# Gráficos de barras por meses:

# Temperaturas 2019:
barplot(medianas[c(5,6,7,8,9,10),],ylim=c(0,37),
beside=T,col=rainbow(6),ylab='ºC',main='Temperaturas 2019',
names.arg =c('Enero','Febrero','Marzo','Abril',
'Mayo','Junio','Julio'))
legend('topleft',rownames(medianas)[c(5,6,7,8,9,10)],
lty=1,lwd=3,cex=1,col=rainbow(6),box.lty = 0)

# Parámetros generales 2019:
barplot(medias[c(1,2),],ylim=c(0,1500),
beside=T,col=rainbow(2),ylab='kW',
main='Parámetros generales 2019',
names.arg =c('Enero','Febrero','Marzo','Abril',
'Mayo','Junio','Julio'))
legend('topleft',rownames(medias)[c(1,2)],
lty=1,lwd=3,cex=1,col=rainbow(2),box.lty =0)
barplot(medias[c(3,4),],ylim=c(0,100),
beside=T,ylab='Porcentaje (%)',col=rainbow(2),
main='Parámetros generales 2019',names.arg =c('Enero',
'Febrero','Marzo','Abril','Mayo','Junio','Julio'))
legend('topleft',rownames(medias)[c(3,4)],
lty=1,lwd=3,cex=1,col=rainbow(2),box.lty =0)
barplot(medias[c(17,18),],ylim=c(0,2),
beside=T,col=rainbow(2),main='Parámetros generales 2019',
names.arg =c('Enero','Febrero','Marzo','Abril',
'Mayo','Junio','Julio'))
legend('topleft',rownames(medias)[c(17,18)],
lty=1,lwd=3,cex=1,col=rainbow(2),box.lty =0)
# Proporción chillers 2019:
barplot(medias[c(19,20),],ylim=c(0,120),
beside=T,col=rainbow(2),main='Proporción chillers 2019',
names.arg =c('Enero','Febrero','Marzo','Abril',
'Mayo','Junio','Julio'))
legend('topleft',rownames(medias)[c(19,20)],
lty=1,lwd=3,cex=1,col=rainbow(2),box.lty =0)

# DIAGRAMA DE ISHIKAWA :
library(SixSigma)
b.effect <- 'Correcto Funcionamiento'
b.groups <- c('Temperatura', 'Consumo','Ocupación','Potencia')
b.causes <- vector(mode = 'list',length = length(b.groups))
b.causes[1] <- list(c('temp_clima_on_prom','temp_ext',
```

```
'temp_ch1_out','temp_ch1_in','temp_ch2_out','temp_ch2_in'))
b.causes[2] <- list(c('consum_chillers',
'consum_ch1','consum_ch2'))
b.causes[3] <- list(c('ocupacion_total'))
b.causes[4] <- list(c('main_pot_activa','pot_torre3_vent1',
'pot_torre3_vent2'))
ss.ceDiag(b.effect, b.groups, b.causes,
sub = 'Datos Hotel Sudamericano',main='ISHIKAWA Diagram')

# DATOS FUNCIONALES:
library(fda.usc)
na_cero<-function(x){ifelse(is.na(x),0,x)}
# Creación de los fdata:
fd1<-matrix(nrow = 212,ncol = 24)
for (i in 1:212) {
fd1[i,]<-data_simple$consum_chillers[(24*(i-1)+1):(24*i)]
}
fd1<-fd1[-c(208:212),]
fdata1<-fdata(fd1)
plot(fdata1,main='Consum_chillers diario',
xlab='Hora',ylab='Consumo')
fd2<-matrix(nrow = 212,ncol = 24)
for (i in 1:212) {
fd2[i,]<-data_simple$main_pot_activa[(24*(i-1)+1):(24*i)]
}
fd2<-fd2[-c(208:212),]
fdata2<-fdata(fd2)
plot(fdata2,main='Main_pot_activa diaria',
xlab='Hora',ylab='Potencia')

# Suavización:
dat.bsp1 = optim.basis(fdata1,numbasis=9) # Suavización bspline
dats1 = dat.bsp1$fdata.est
dat.bsp2 = optim.basis(fdata2,numbasis=9) # Suavización bspline
dats2 = dat.bsp2$fdata.est

par(mfrow=c(1,3))
plot(fdata1,main='Consumo diario de los chiller',
sub='Datos originales',xlab='Hora',
ylab='Consumo_chillers',col=rainbow(100))
plot(dats1,main='Consumo diario de los chiller',
sub='Datos suavizados',xlab='Hora',
ylab='Consumo_chillers',col=rainbow(100))
library(binovisualfields)
legend_image <- as.raster(matrix(rainbow(100)))
plot(c(0,2),c(0,1),type = 'n', axes = F,
xlab = '', ylab = '')
text(x=1.2,y =seq(0.01,1.1,1/6),
labels=c('-Julio','-Junio','-Mayo','-Abril','-Marzo',
'-Febrero','-Enero'))
rasterImage(legend_image, 0, 0, 1,1)
```

```
par(mfrow=c(1,3))
plot(fdata2,main='Potencia activa diaria',
sub='Datos originales',xlab='Hora',
ylab='Main_pot_activa',col=rainbow(100))
plot(dats2,main='Potencia activa diaria',
sub='Datos suavizados',xlab='Hora',
ylab='Main_pot_activa',col=rainbow(100))
legend_image <- as.raster(matrix(rainbow(100)))
plot(c(0,2),c(0,1),type = 'n', axes = F,xlab = '', ylab = '')
text(x=1.2,y =seq(0.01,1.1,1/6),
labels=c('-Julio','-Junio','-Mayo','-Abril','-Marzo',
'-Febrero','-Enero'))
rasterImage(legend_image, 0, 0, 1,1)
par(mfrow=c(1,1))

# ANÁLISIS POR MESES:
# A) CONSUMOS
# ¿Qué pasa con los consumos en cada mes?
color<-c('black','red','green','orange','pink',
'lightblue','brown','yellow','purple','blue')
# ENERO
enero_aux11<-enero[,c(3,6,15,16,23)]
plot(enero_aux11$consum_chillers~enero_aux11$dia,
type='n',xlab='Día',ylab='Consumo (kW)',lty=0,
main='Enero consumo medio (diario)',pch=NA,ylim=c(0,500))
for (h in 1:(ncol(enero_aux11)-1)){
enero_mean<-aggregate(enero_aux11[,h]~dia, enero_aux11, mean)
lines(enero_mean[,2]~enero_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(3,6,15,16)]),
col=color,lty=1,lwd=2,cex=0.8)

# FEBRERO
febrero_aux1<-febrero[,c(3,6,15,16,23)]
plot(febrero_aux1$consum_chillers~febrero_aux1$dia,
type='n',xlab='Día',ylab='Consumo (kW)',lty=0,
main='Febrero consumo medio (diario)',pch=NA,ylim=c(0,500))
for (h in 1:(ncol(febrero_aux1)-1)){
febrero_mean<-aggregate(febrero_aux1[,h]~dia,febrero_aux1, mean)
lines(febrero_mean[,2]~febrero_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(3,6,15,16)]),
col=color,lty=1,lwd=2,cex=0.8)

# MARZO
marzo_aux1<-marzo[,c(3,6,15,16,23)]
plot(marzo_aux1$consum_chillers~marzo_aux1$dia,type='n',
xlab='Día',ylab='Consumo (kW)',lty=0,
main='Marzo consumo medio (diario)',pch=NA,ylim=c(0,500))
for (h in 1:(ncol(marzo_aux1)-1)){
```

```
marzo_mean<-aggregate(marzo_aux1[,h]~dia, marzo_aux1, mean)
lines(marzo_mean[,2]~marzo_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(3,6,15,16)]),
col=color,lty=1,lwd=2,cex=0.8)

# ABRIL
abril_aux1<-abril[,c(3,6,15,16,23)]
plot(abril_aux1$consum_chillers~abril_aux1$dia,type='n',
xlab='Día',ylab='Consumo (kW)',lty=0,
main='Abril consumo medio (diario)',pch=NA,ylim=c(0,500))
for (h in 1:(ncol(abril_aux1)-1)){
abril_mean<-aggregate(abril_aux1[,h]~dia, abril_aux1, mean)
lines(abril_mean[,2]~abril_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(3,6,15,16)]),
col=color,lty=1,lwd=2,cex=0.8)

# MAYO
mayo_aux1<-mayo[,c(3,6,15,16,23)]
plot(mayo_aux1$consum_chillers~mayo_aux1$dia,type='n',
xlab='Día',ylab='Consumo (kW)',lty=0,
main='Mayo consumo medio (diario)',pch=NA,ylim=c(0,500))
for (h in 1:(ncol(mayo_aux1)-1)){
mayo_mean<-aggregate(mayo_aux1[,h]~dia, mayo_aux1, mean)
lines(mayo_mean[,2]~mayo_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(3,6,15,16)]),
col=color,lty=1,lwd=2,cex=0.8)

# JUNIO
junio_aux1<-junio[,c(3,6,15,16,23)]
plot(junio_aux1$consum_chillers~junio_aux1$dia,type='n',
xlab='Día',ylab='Consumo (kW)',lty=0,
main='Junio consumo medio (diario)',pch=NA,ylim=c(0,500))
for (h in 1:(ncol(junio_aux1)-1)){
junio_mean<-aggregate(junio_aux1[,h]~dia, junio_aux1, mean)
lines(junio_mean[,2]~junio_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(3,6,15,16)]),
col=color,lty=1,lwd=2,cex=0.8)

# JULIO
julio_aux1<-julio[,c(3,6,15,16,23)]
plot(julio_aux1$consum_chillers~julio_aux1$dia,type='n',
xlab='Día',ylab='Consumo (kW)',lty=0,
main='Julio consumo medio (diario)',pch=NA,ylim=c(0,500))
for (h in 1:(ncol(julio_aux1)-1)){
julio_mean<-aggregate(julio_aux1[,h]~dia, julio_aux1, mean)
lines(julio_mean[,2]~julio_mean$dia,col=color[h])
}
```

```
legend('topright',colnames(data_complete[,c(3,6,15,16)]),
col=color,lty=1,lwd=2,cex=0.8)


# B) TEMPERATURAS
color<-c('black','red','green','orange','pink',
'lightblue','brown','yellow','purple','blue')
# ENERO
enero_aux2<-enero[,c(7,8,9,10,11,12,23)]
plot(enero_aux2$temp_ext~enero_aux2$dia,type='n',
lty=0,xlab='Día',ylab='Temperatura (ºC)',
main='Enero temperaturas (diaria)',pch=NA,ylim=c(0,35))
for (h in 1:(ncol(enero_aux2)-1)){
enero_mean<-aggregate(enero_aux2[,h]~dia, enero_aux2, mean)
lines(enero_mean[,2]~enero_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(7,8,9,10,11,12)]),
col=color,lty=1,lwd=2,cex=0.8)

# FEBRERO
febrero_aux2<-febrero[,c(7,8,9,10,11,12,23)]
plot(febrero_aux2$temp_ext~febrero_aux2$dia,type='n',
lty=0,xlab='Día',ylab='Temperatura (ºC)',
main='Febrero temperaturas (diaria)',pch=NA,ylim=c(0,35))
for (h in 1:(ncol(febrero_aux2)-1)){
febrero_mean<-aggregate(febrero_aux2[,h]~dia, febrero_aux2, mean)
lines(febrero_mean[,2]~febrero_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(7,8,9,10,11,12)]),
col=color,lty=1,lwd=2,cex=0.8)

# MARZO
marzo_aux2<-marzo[,c(7,8,9,10,11,12,23)]
plot(marzo_aux2$temp_ext~marzo_aux2$dia,type='n',lty=0,xlab='Día',
ylab='Temperatura (ºC)',main='Marzo temperaturas (diaria)',
pch=NA,ylim=c(0,35))
for (h in 1:(ncol(marzo_aux2)-1)){
marzo_mean<-aggregate(marzo_aux2[,h]~dia, marzo_aux2, mean)
lines(marzo_mean[,2]~marzo_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(7,8,9,10,11,12)]),
col=color,lty=1,lwd=2,cex=0.8)
# Se quieren visualizar temperaturas y consumos cuando
#solo funciona el chiller 1 o el chiller 2:
fun_ch2<-marzo[ which(marzo$consum_ch1==0 &
!marzo$consum_ch2==0),c(1,2,4,7)] # Solo funciona el ch2
fun_ch1<-marzo[ which(marzo$consum_ch2==0 &
!marzo$consum_ch1==0),c(1,2,4,7)] # Solo funciona el ch1
no_fun<-marzo[ which(marzo$consum_ch2==0 &
marzo$consum_ch1==0),c(1,2,4,7)] # Ningun ch funciona
mar3<-marzo[marzo$dia==3,]
```

```
par(mfrow=c(2,2))
plot(mar3$main_pot_activa~mar3$time,type='n',
xlab='Hora',ylab='Consumo (kW)',lty=0,
main='Consumo con chiller 1 (03/03/2019)')
lines(mar3$main_pot_activa~mar3$time)
plot(mar3$temp_clima_on_prom~mar3$time,xlab='Hora',
ylab='Temperatura (ºC)',type='n',lty=0,
main='Consumo con chiller 1 (03/03/2019)')
lines(mar3$temp_clima_on_prom~mar3$time)
mar8<-marzo[marzo$dia==8,]
plot(mar8$main_pot_activa~mar8$time,type='n',xlab='Hora',
ylab='Consumo (kW)',lty=0,
main='Consumo con chiller 2 (08/03/2019)')
lines(mar8$main_pot_activa~mar8$time)
plot(mar8$temp_clima_on_prom~mar8$time,xlab='Hora',
ylab='Temperatura (ºC)',type='n',lty=0,
main='Consumo con chiller 2 (08/03/2019)')
lines(mar8$temp_clima_on_prom~mar8$time)
par(mfrow=c(1,1))

# ABRIL
abril_aux2<-abril[,c(7,8,9,10,11,12,23)]
plot(abril_aux2$temp_ext~abril_aux2$dia,type='n',lty=0,xlab='Día',
ylab='Temperatura (ºC)',main='Abril temperaturas (diaria)',
pch=NA,ylim=c(0,65))
for (h in 1:(ncol(abril_aux2)-1)){
abril_mean<-aggregate(abril_aux2[,h]~dia, abril_aux2, mean)
lines(abril_mean[,2]~abril_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(7,8,9,10,11,12)]),
col=color,lty=1,lwd=2,cex=0.7)

# MAYO
mayo_aux2<-mayo[,c(7,8,9,10,11,12,23)]
plot(mayo_aux2$temp_ext~mayo_aux2$dia,type='n',lty=0,xlab='Día',
ylab='Temperatura (ºC)',main='Mayo temperaturas (diaria)',pch=NA,
ylim=c(0,35))
for (h in 1:(ncol(mayo_aux2)-1)){
mayo_mean<-aggregate(mayo_aux2[,h]~dia, mayo_aux2, mean)
lines(mayo_mean[,2]~mayo_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(7,8,9,10,11,12)]),
col=color,lty=1,lwd=2,cex=0.7)

# JUNIO
junio_aux2<-junio[,c(7,8,9,10,11,12,23)]
plot(junio_aux2$temp_ext~junio_aux2$dia,type='n',lty=0,
xlab='Día',ylab='Temperatura (ºC)',
main='Junio temperaturas (diaria)',pch=NA,ylim=c(0,40))
for (h in 1:(ncol(junio_aux2)-1)){
```

```
junio_mean<-aggregate(junio_aux2[,h]~dia, junio_aux2, mean)
lines(junio_mean[,2]~junio_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(7,8,9,10,11,12)]),
col=color,lty=1,lwd=2,cex=0.7)

# JULIO
julio_aux2<-julio[,c(7,8,9,10,11,12,23)]
plot(julio_aux2$temp_ext~julio_aux2$dia,type='n',lty=0,
xlab='Día',ylab='Temperatura (ºC)',
main='Julio temperaturas (diaria)',pch=NA,ylim=c(0,40))
for (h in 1:(ncol(julio_aux2)-1)){
julio_mean<-aggregate(julio_aux2[,h]~dia, julio_aux2, mean)
lines(julio_mean[,2]~julio_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(7,8,9,10,11,12)]),
col=color,lty=1,lwd=2,cex=0.7)

# C)  POTENCIAS
color<-c('black','red','green','orange','pink','lightblue',
'brown','yellow','purple','blue')
# ENERO
enero_aux3<-enero[,c(17,18,19,20,23)]
plot(enero_aux3$pot_torre3_vent1~enero_aux3$dia,type='n',
xlab='Día',ylab='Potencia',lty=0,
main='Enero potencia torre 3 (diario)',
pch=NA,ylim=c(0,1.3))
for (h in 1:(ncol(enero_aux3)-1)){
enero_mean<-aggregate(enero_aux3[,h]~dia, enero_aux3, mean)
lines(enero_mean[,2]~enero_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(17,18,19,20)]),
col=color,lty=1,lwd=2,cex=0.8)
# Se ve que por ejemplo, el día 8 de enero ningun ventilador
# está funcionando de 12:00 a 14:00, para ver que pasa:
enero[ which(enero$torre3_vent1==0 & enero$torre3_vent2==0),1:2]
ene8<-enero[enero$dia==8,]
par(mfrow=c(2,1))
plot(ene8$consum_chillers~ene8$time,type='n',xlab='Hora',
ylab='Consumo (kW)',lty=0,
main='Consumo de chillers 8/01/2019')
lines(ene8$consum_chillers~ene8$time)
plot(ene8$temp_clima_on_prom~ene8$time,xlab='Hora',
ylab='Temperatura (ºC)',type='n',lty=0,
main='Temperatura interior 8/01/2019')
lines(ene8$temp_clima_on_prom~ene8$time)
par(mfrow=c(1,1))

# FEBRERO
febrero_aux3<-febrero[,c(17,18,19,20,23)]
plot(febrero_aux3$pot_torre3_vent1~febrero_aux3$dia,
```

```
type='n',xlab='Día',ylab='Potencia',lty=0,
main='Febrero potencia torre 3 (diario)',
pch=NA,ylim=c(0,3.5))
for (h in 1:(ncol(febrero_aux3)-1)){
febrero_mean<-aggregate(febrero_aux3[,h]~dia, febrero_aux3, mean)
lines(febrero_mean[,2]~febrero_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(17,18,19,20)]),
col=color,lty=1,lwd=2,cex=0.8)

# MARZO
marzo_aux3<-marzo[,c(17,18,19,20,23)]
plot(marzo_aux3$pot_torre3_vent1~marzo_aux3$dia,
type='n',xlab='Día',ylab='Potencia',lty=0,
main='Marzo potencia torre 3 (diario)',pch=NA,ylim=c(0,3.5))
for (h in 1:(ncol(marzo_aux3)-1)){
marzo_mean<-aggregate(marzo_aux3[,h]~dia, marzo_aux3, mean)
lines(marzo_mean[,2]~marzo_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(17,18,19,20)]),
col=color,lty=1,lwd=2,cex=0.8)

# ABRIL
abril_aux3<-abril[,c(17,18,19,20,23)]
plot(abril_aux3$pot_torre3_vent1~abril_aux3$dia,
type='n',xlab='Día',ylab='Potencia',lty=0,
main='Abril potencia torre 3 (diario)',pch=NA,ylim=c(0,3.5))
for (h in 1:(ncol(abril_aux3)-1)){
abril_mean<-aggregate(abril_aux3[,h]~dia, abril_aux3, mean)
lines(abril_mean[,2]~abril_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(17,18,19,20)]),
col=color,lty=1,lwd=2,cex=0.8)

# MAYO
mayo_aux3<-mayo[,c(17,18,19,20,23)]
plot(mayo_aux3$pot_torre3_vent1~mayo_aux3$dia,type='n',
xlab='Día',ylab='Potencia',lty=0,
main='Mayo potencia torre 3 (diario)',pch=NA,ylim=c(0,3.5))
for (h in 1:(ncol(mayo_aux3)-1)){
mayo_mean<-aggregate(mayo_aux3[,h]~dia, mayo_aux3, mean)
lines(mayo_mean[,2]~mayo_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(17,18,19,20)]),
col=color,lty=1,lwd=2,cex=0.8)
# Se ve que el día 13 de 10:00-13:00 ocurre un apagado general:
may13<-mayo[mayo$dia==13,]

par(mfrow=c(2,1))
plot(may13$consum_chillers~may13$time,type='n',xlab='Hora',
ylab='Consumo (kW)',lty=0,
```

```
main='Consumo de chillers 13/05/2019')
lines(may13$consum_chillers~may13$time)
plot(may13$temp_clima_on_prom~may13$time,xlab='Hora',
ylab='Temperatura (ºC)',type='n',lty=0,
main='Temperatura interior 13/05/2019')
lines(may13$temp_clima_on_prom~may13$time)
par(mfrow=c(1,1))

# JUNIO
junio_aux3<-junio[,c(17,18,19,20,23)]
plot(junio_aux3$pot_torre3_vent1~junio_aux3$dia,type='n',
xlab='Día',ylab='Potencia',lty=0,
main='Junio potencia torre 3 (diario)',
pch=NA,ylim=c(0,2.5))
for (h in 1:(ncol(junio_aux3)-1)){
junio_mean<-aggregate(junio_aux3[,h]~dia, junio_aux3, mean)
lines(junio_mean[,2]~junio_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(17,18,19,20)]),
col=color,lty=1,lwd=2,cex=0.8)
# Se ve que el día 5 de 10:00-13:00 ocurre un apagado general:
jun5<-data[3721:3744,]

par(mfrow=c(2,1))
plot(jun5$consum_chillers~jun5$time,type='n',xlab='Hora',
ylab='Consumo (kW)',lty=0,main='Consumo de chillers 05/06/2019')
lines(jun5$consum_chillers~jun5$time)
plot(jun5$temp_clima_on_prom~jun5$time,xlab='Hora',
ylab='Temperatura (ºC)',type='n',lty=0,
main='Temperatura interior 05/06/2019')
lines(jun5$temp_clima_on_prom~jun5$time)
par(mfrow=c(1,1))

# JULIO
julio_aux3<-julio[,c(17,18,19,20,23)]
plot(julio_aux3$pot_torre3_vent1~julio_aux3$dia,type='n',
xlab='Día',ylab='Potencia',lty=0,
main='Julio potencia torre 3 (diario)',pch=NA,ylim=c(0,3.5))
for (h in 1:(ncol(julio_aux3)-1)){
julio_mean<-aggregate(julio_aux3[,h]~dia, julio_aux3, mean)
lines(julio_mean[,2]~julio_mean$dia,col=color[h])
}
legend('topright',colnames(data_complete[,c(17,18,19,20)]),
col=color,lty=1,lwd=2,cex=0.8)


# GRÁFICOS CORRELACIONES:
library(PerformanceAnalytics)
library(gridExtra)
chart.Correlation(data_simple,histogram=TRUE,method="pearson")
library(corrplot)
```

```
M <- cor(data_simple)
round(M,2)
corrplot(M, method="circle",type="upper")

# ESTIMACIÓN NO PARAMÉTRICA DE LA DENSIDAD:
multiplot <- function(..., plotlist=NULL, file,
cols=1, layout=NULL) {
library(grid)
plots <- c(list(...), plotlist)
numPlots = length(plots)
if (is.null(layout)) {
layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
ncol = cols, nrow = ceiling(numPlots/cols))}
if (numPlots==1) {
print(plots[[1]])} else {
grid.newpage()
pushViewport(viewport(layout = grid.layout(nrow(layout),
ncol(layout))))
for (i in 1:numPlots) {
matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))
print(plots[[i]],
vp=viewport(layout.pos.row = matchidx$row,
layout.pos.col = matchidx$col))}}}
p1.1 <- ggplot(data_simple, aes(consum_chillers)) +
geom_histogram(aes(y=..density..), colour="black", fill="white") +
geom_density(alpha=.2, fill="#FF6666")
p2.1 <- ggplot(data_simple, aes(main_pot_activa)) +
geom_histogram(aes(y=..density..), colour="black", fill="white") +
geom_density(alpha=.2, fill="#FF6666")
p3.1 <- ggplot(data_simple, aes(ocupacion_total)) +
geom_histogram(aes(y=..density..), colour="black", fill="white") +
geom_density(alpha=.2, fill="#FF6666")
multiplot(p1.1,p2.1,p3.1, cols = 3)

# OTROS GRÁFICOS DE INTERÉS:
# Gráfico de contorno (x = consum_chillers,
y = temp_clima_on_prom):
commonTheme = list(labs(color="Density",fill="Density"),
theme_bw(),     theme(legend.position=c(0,1),
legend.direction ="horizontal",
legend.justification=c(0,1)))
p8 <-ggplot(data=data_simple,aes(x = consum_chillers,
y = temp_clima_on_prom)) +
stat_density2d(aes(fill=..level..,alpha=..level..),
geom='polygon',colour='black') +
scale_fill_continuous(low="green",high="red") +
geom_smooth(method=lm,linetype=2,colour="red",se=F) +
guides(alpha="none") +  geom_point() + commonTheme
p8
```

```
# CONTRASTES DE NORMALIDAD:
# Shapiro test:
library(mvShapiroTest)
mvShapiro.Test(as.matrix(data_simple))
# Mardia test:
library(MVN)
mvn(data_simple, subset = NULL, mvnTest = c("mardia", "hz",
"royston", "dh", "energy"), covariance = TRUE, tol = 1e-25,
alpha = 0.5, scale = FALSE, desc = TRUE, transform = "none",
R = 1000,univariateTest = c("SW", "CVM", "Lillie", "SF", "AD"),
univariatePlot = "none", multivariatePlot = "none",
multivariateOutlierMethod = "none",
bc = FALSE, bcType = "rounded",
showOutliers = FALSE,   showNewData = FALSE)
# Normal qqplot univariado:
qqplot.data <- function (vec)
{y <- quantile(vec[!is.na(vec)], c(0.25, 0.75))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int <- y[1L] - slope * x[1L]
d <- data.frame(resids = vec)
ggplot(d,aes(sample = resids))+stat_qq()+
geom_abline(slope=slope,intercept=int,
col="blue")+
ggtitle("Normal␣Q-Q␣Plot")}
qqplot.data(data_simple[,1])
for (i in 1:ncol(d_cor)){
qqnorm(d_cor[,i], pch = 1, frame = FALSE,
main=colnames(d_cor)[i])
qqline(d_cor[,i], col = "steelblue", lwd = 2)}

# INDEPENDENCIA DE LOS DATOS:
# Estudio de autocorrelaciones:
par(mfrow=c(3,4)) #Correladas
for (i in 1:ncol(data_simple)){
acf(data_simple[,i],lag=10,las=1,
main=colnames(data_simple)[i])}
par(mfrow=c(1,1))
pvalor<-matrix(rep(0,ncol(data_simple)),
ncol=ncol(data_simple))
for (i in 1:ncol(data_simple)){
pvalor[,i]<-Box.test(data_simple[,i],type='Ljung',
lag=1,fitdf=0)$p.value}
colnames(pvalor)<-colnames(data_simple)
pvalor

# ANÁLISIS DE COMPONENTES PRINCIPALES:
test.pca <- princomp(data_simple,cor=T)
test.pca
summary(test.pca)
# Obtener las varianzas de las componentes,
```

```
# que son los autovalores (gráfico de sedimentación):
test.pca$sdev^2 #Las varianzas de las z_i
# Loadings de cada una de las observaciones:
test.pca$loadings
barplot(loadings(test.pca)[,1:7], beside = TRUE,
main='Coeficientes de la combinación lineal (Autovectores)',
ylim=c(-0.6,1), col=topo.colors(12))
legend('topright',colnames(data_simple),lwd=2,lty=1,
col=topo.colors(ncol(data_simple)),cex=0.8)
# Puntuaciones de los individuos en las componentes:
test.pca$scores
# Graficar las observaciones frente a las 2 primeras componentes:
plot(test.pca$scores[,1],test.pca$scores[,2],
xlab='Comp. 1',ylab='Comp. 2',
main='Observaciones frente a las dos primeras CP')
text(test.pca$scores[,1],test.pca$scores[,2],
labels=data_sim$id)
# Representar el biplot:
biplot(test.pca,main='Observaciones frente a las dos primeras CP',
col=c('blue','black'),lwd=c(1,2),xlim=c(-0.05,0.085))
abline(v=0,lty=2); abline(h=0,lty=2)
# Con el paquete "factoextra":
library(factoextra)
fviz_pca_var(test.pca, col.var = "cos2",
geom.var = "arrow", labelsize = 3, repel = FALSE)
```

Gráficos de control paramétricos

```
# Gráficos de control paramétricos de franjas horarias
# de alta y baja actividad:
# Búcle para calcular la media de cada 24 horas de cada
# día para obtener 1 observación por día:
colnames(data)
data<-data[,-c(13,14,15,16,17,18,21)]
dias<-levels(data$date)
ncol(data)
reducido<-matrix(nrow = 202*12,ncol = 12)
for (i in 1:2424) {
reducido[i,]<-apply(data[(2*(i-1)+1):(2*i),3:14],2,median)
}
fechas<-matrix(numeric(2424))
num<-seq(1,nrow(reducido),12)
for(i in 1:nrow(reducido)){
num1<-num[i]
num2<-num1+1
num3<-num1+2
num4<-num1+3
num5<-num1+4
num6<-num1+5
num7<-num1+6
```

```
num8<-num1+7
num9<-num1+8
num10<-num1+9
num11<-num1+10
num12<-num1+11
fechas[num1,]<-paste(dias[i],'a')
fechas[num2,]<-paste(dias[i],'b')
fechas[num3,]<-paste(dias[i],'c')
fechas[num4,]<-paste(dias[i],'d')
fechas[num5,]<-paste(dias[i],'e')
fechas[num6,]<-paste(dias[i],'f')
fechas[num7,]<-paste(dias[i],'g')
fechas[num8,]<-paste(dias[i],'h')
fechas[num9,]<-paste(dias[i],'i')
fechas[num10,]<-paste(dias[i],'j')
fechas[num11,]<-paste(dias[i],'k')
fechas[num12,]<-paste(dias[i],'l')}
nrow(reducido)==nrow(fechas)
tail(fechas)
tail(reducido)
reducido<-data.frame(fechas,round(reducido,12))
head(reducido)
colnames(reducido)<-colnames(data[,-c(2,16)])
reducido<-reducido[sort(c(num+1,num+9)),]

# CREACIÓN DE LA FUNCIÓN:
t2<-function(clus,q1,q2,type){
library(MSQC)
selec<-clus$temp_ch2_out>quantile(clus$temp_ch2_out,q1) &
clus$temp_ch2_out<quantile(clus$temp_ch2_out,q2)
fase1<-clus[selec,]
fase2<-clus[!selec,]
nrow(fase1)
nrow(fase2)
Xmv<-mult.chart(fase1,type=type,alpha =0.01)$Xmv
# windows()
par(mfrow=c(1,2))
S<-mult.chart(fase1,type=type,alpha =0.01)$covariance
t2_hot<-mult.chart(fase2,type=type,Xmv=Xmv,S=S,
colm=nrow(fase1),alpha=0.01)
par(mfrow=c(1,1))
options(max.print=1000000)
long<-nrow(fase2)-(table(t2_hot$t2>t2_hot$ucl)[1])
outofcontrol<-list()
for (i in 1:(long)){
outofcontrol[[i]]<-reducido[which(reducido$temp_ch2_in==
fase2[which(t2_hot$t2>t2_hot$ucl)[i],]$temp_ch2_in),]$date}
porcentaje<-table(t2_hot$t2>t2_hot$ucl)[2]/nrow(clus)*100
return(list('outofcontrol'=outofcontrol,'porcentaje'=porcentaje))
}
```

```
# Posible aplicación: chi, t2, mewma, mcusum, mcusum2
t2(enero,0.2,0.75,type='t2')
t2(febrero,0.2,0.75,type='t2')
t2(marzo,0.2,0.75,type='t2')
t2(abril,0.2,0.75,type='t2')
t2(mayo,0.2,0.75,type='t2')
t2(junio,0.2,0.75,type='t2')
t2(julio,0.2,0.75,type='t2')

library(mvShapiroTest)
library(MVN)
mvShapiro.Test(as.matrix(enero))
mvShapiro.Test(as.matrix(febrero))
mvShapiro.Test(as.matrix(marzo))
mvShapiro.Test(as.matrix(abril))
mvShapiro.Test(as.matrix(mayo))
mvShapiro.Test(as.matrix(junio))
mvShapiro.Test(as.matrix(julio))

pvalor<-matrix(rep(0,7*ncol(enero)),ncol=ncol(enero))
for (i in 1:ncol(enero)){
pvalor[1,i]<-Box.test(enero[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[2,i]<-Box.test(febrero[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[3,i]<-Box.test(marzo[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[4,i]<-Box.test(abril[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[5,i]<-Box.test(mayo[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[6,i]<-Box.test(junio[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[7,i]<-Box.test(julio[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor<-round(pvalor,4)}
rownames(pvalor)<-c('Enero','Febrero','Marzo',
'Abril','Mayo','Junio','Julio')
colnames(pvalor)<-colnames(enero)
pvalor>0.01 #TRUE means independent
table(pvalor>0.01)[2]/
(table(pvalor>0.01)[1]+table(pvalor>0.01)[2])*100

# Agrupando los meses en los grupos cluster:
uno<-enero
dos<-rbind(febrero,marzo)
tres<-abril
cuatro<-rbind(mayo,junio,julio)

pvalor<-matrix(rep(0,4*ncol(enero)),ncol=ncol(enero))
for (i in 1:ncol(enero)){
```

```
pvalor[1,i]<-Box.test(uno[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[2,i]<-Box.test(dos[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[3,i]<-Box.test(tres[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor[4,i]<-Box.test(cuatro[,i],type='Ljung',lag=1,
fitdf=0)$p.value
pvalor<-round(pvalor,4)}
colnames(pvalor)<-colnames(enero)
rownames(pvalor)<-c('Grupo 1','Grupo 2','Grupo 3','Grupo 4')
pvalor>0.01 #TRUE means independent
table(pvalor>0.01)[2]/
(table(pvalor>0.01)[1]+table(pvalor>0.01)[2])*100

mvShapiro.Test(as.matrix(uno))
mvShapiro.Test(as.matrix(dos))
mvShapiro.Test(as.matrix(tres))
mvShapiro.Test(as.matrix(cuatro))

t2(uno,0.2,0.75,type='t2')
t2(dos,0.2,0.75,type='t2')
t2(tres,0.2,0.75,type='t2')
t2(cuatro,0.2,0.75,type='t2')
```

Gráficos de control no paramétricos

```
# GRÁFICOS MULTIVARIANTES NO-PARAMÉTRICOS:
# Funciones necesarias:
# 1. Rangos de profundidad
rangoprof<-function(n){
r<-vector(length=length(n))
for(i in 1:length(n)){
r[i]<(2/length(n))*(min(sum(n[i]<n),sum(n[i]>n)))+
sum(n[i]==n)/length(n)}
r}
# 2. Función gráficos r:
graf.control.rQ<-function(r,n,alpha){
k<-n
if(k==1){
central<-0.5
LCL<-alpha
tiempo<-1:length(r)
plot(tiempo,r,type='l',col="darkblue",xaxt='n',
xlim=c(0,(length(r)+18))
,ylim=c(min(LCL,min(r))-0.005,max(r)+0.05))
axis(1,seq(from=0, to=(length(r)),by=10),cex.axis=0.8,las=2)
abline(h=LCL,lty=2)
text((max(tiempo)+10),LCL,paste('LCL=',LCL),pos=3,font=2,cex=0.8)
abline(h=0.5,lty=2)
```

```
    text((max(tiempo)+6),central,paste('central=',0.5),
    pos=3,font=2,cex=0.8)
    for(i in 1:length(tiempo)){
    if(r[i]<LCL){
    points(tiempo[i],r[i],pch=4,col="red")}
    else {
    points(tiempo[i],r[i],pch=20)}}
    mtext('Gráfico de control r',side=3,font=2)
    return(list('LCL'=LCL,'r'=r,'tiempo'=tiempo))}
    if(k>=2){
    x<-matrix(r,nrow=k)
    central<-0.5
    tiempo<-1:ncol(x)
    medias<-c(apply(x,2,mean))
    r<-medias
    zalfa<-qnorm(1-alpha/2,0,1)
    if(k<5){
    LCL<-((factorial(nrow(x))*alpha)^(1/nrow(x)))/nrow(x)}
    if(k>=5){
    LCL<-0.5-zalfa*(1/(sqrt(12*nrow(x))))}
    plot(tiempo,r,type='l',xaxt='n',col="darkblue",xlim=c(0,
    (length(r)+18))
    ,ylim=c(min(LCL,min(r))-0.005,max(r)+0.05))
    axis(1,1:(length(r)),cex.axis=0.8,las=2)
    abline(h=LCL,lty=2)
    text((max(tiempo)+6),LCL,paste('LCL=',LCL),pos=3,font=2,cex=0.8)
    abline(h=0.5,lty=2)
    text((max(tiempo)+6),central,paste('central=',0.5),pos=3,font=2,
    cex=0.8)
    for(i in 1:length(tiempo)){
    if(r[i]<LCL){
    points(tiempo[i],r[i],pch=4,col="red")}
    else {
    points(tiempo[i],r[i],pch=20)}}
    mtext('Gráfico de control Q',side=3,font=2)}
    return(list('LCL'=LCL,'r'=r,'tiempo'=tiempo))}

    # 3. Función gráficos Q:
    graf.control.rQ2<-function(r,k,alpha,LCL){
    if(k==1){
    central<-0.5
    tiempo<-1:length(r)
    plot(tiempo,r,type='l',col="darkblue",xaxt='n',xlim=c(0,
    (length(r)+18))
    ,ylim=c(min(LCL,min(r))-0.005,max(r)+0.05))
    axis(1,seq(from=0, to=(length(r)),by=10),cex.axis=0.8,las=2)
    abline(h=LCL,lty=2)
    text((max(tiempo)+10),LCL,paste('LCL=',LCL),pos=3,font=2,cex=0.8)
    abline(h=0.5,lty=2)
    text((max(tiempo)+6),central,paste('central=',0.5),pos=3,font=2,
    cex=0.8)
```

```
    for(i in 1:length(tiempo)){
    if(r[i]<LCL){
    points(tiempo[i],r[i],pch=4,col="red")}
    else {
    points(tiempo[i],r[i],pch=20)}}
    mtext('Gráfico de control r',side=3,font=2)
    return(list('LCL'=LCL,'r'=r,'tiempo'=tiempo)}
    if(k>=2){
    x<-matrix(r,nrow=k)
    central<-0.5
    tiempo<-1:ncol(x)
    medias<-c(apply(x,2,mean))
    r<-medias
    zalfa<-qnorm(1-alpha,0,1)
    plot(tiempo,r,type='l',xaxt='n',col="darkblue",
    xlim=c(0,(length(r)+18))
    ,ylim=c(min(LCL,min(r))-0.005,max(r)+0.05))
    axis(1,1:(length(r)),cex.axis=0.8,las=2)
    abline(h=LCL,lty=2)
    text((max(tiempo)+6),LCL,paste('LCL=',LCL),pos=3,font=2,cex=0.8)
    abline(h=0.5,lty=2)
    text((max(tiempo)+6),central,paste('central=',0.5),pos=3,font=2,
    cex=0.8)
    for(i in 1:length(tiempo)){
    if(r[i]<LCL){
    points(tiempo[i],r[i],pch=4,col="red")}
    else {
    points(tiempo[i],r[i],pch=20)}}
    mtext('Gráfico de control Q',side=3,font=2)
    return(list('LCL'=LCL,'r'=r,'tiempo'=tiempo)}}

    # 4. Función gráficas S:
    grafcontrolS<-function(r,alpha){
    zalfa<-qnorm(1-alpha,0,1)
    if(length(r)<=30){
    LCL<--(zalfa*sqrt(length(r)/12))
    tiempo<-1:length(r)
    CL<-0
    b<-c(r-1/2)
    sumaacumulada<-cumsum(b)
    plot(tiempo,sumaacumulada,type='l',col="darkblue",xaxt='n',
    xlim=(c(0,length(sumaacumulada)+20)),
    ylim=c(min(LCL,min(sumaacumulada))-
    0.5,max(sumaacumulada)+0.5))
    axis(1,seq(from=0, to=(length(r)+6),by=10),cex.axis=0.8,las=2)
    abline(h=LCL,lty=2)
    text((max(tiempo)+10),LCL,paste('LCL=',LCL),pos=3,font=2,cex=0.8)
    abline(h=CL,lty=2)
    text((max(tiempo)+10),CL,paste('CL=',CL),pos=3,font=2,cex=0.7)
    for(i in 1:length(tiempo)){
    if(sumaacumulada[i]<LCL){
```

```
    points(tiempo[i],sumaacumulada[i],pch=4,col="red")}
    else {
    points(tiempo[i],sumaacumulada[i],pch=20)}}
    mtext('Gráfico de control S',side=3,font=2)}
    if(length(r)>30){
    CL<-0
    tiempo<-1:length(r)
    LCL<--zalfa
    b<-c(r-1/2)
    sumaacumulada<-cumsum(b)
    Sn<-sumaacumulada/sqrt(length(r)/12)
    plot(tiempo,Sn,type='l',col="darkblue",xaxt='n',
    xlim=(c(0,length(Sn)+20)),ylim=c(min(LCL,min(Sn))-0.5,
    max(Sn)+0.5))
    axis(1,seq(from=0, to=(length(r)+6),by=10),cex.axis=0.8,las=2)
    abline(h=LCL,lty=2)
    text((max(tiempo)+10),LCL,paste('LCL=',LCL),pos=3,font=2,cex=0.8)
    abline(h=CL,lty=2)
    text((max(tiempo)+10),CL,paste('CL=',CL),pos=3,font=2,cex=0.7)
    for(i in 1:length(tiempo)){
    if(Sn[i]<LCL){
    points(tiempo[i],Sn[i],pch=4,col="red")}
    else {
    points(tiempo[i],Sn[i],pch=20)}}
    mtext('Gráfico de control S*',side=3,font=2)}
    return(list('LCL'=LCL,'suma'=sumaacumulada,'Sn'=Sn,
    'tiempo'=tiempo))}

    # Gráfica de contro r:
    X<-data_simple
    samplemean<-apply(X,2,mean)
    covX<-cov(X)
    solve(covX)
    library(depth)
    m_distance <- mahalanobis(X,samplemean,covX)
    m_depth <- 1/(1 + m_distance)
    r<-rangoprof(m_depth)
    graf.r<-graf.control.rQ(r,n=1,alpha=0.01) #n=1 para gráfico r
    table(graf.r$r<graf.r$LCL)
    # Detección:
    outofcontrol<-list()
    for (i in 1:(table(graf.r$r<graf.r$LCL)[2])){
    outofcontrol[[i]]<-data_complete[which(data_complete$temp_ch2_in==
    data_simple[which(graf.r$r<graf.r$LCL)[i],]$temp_ch2_in &
    data_complete$temp_ch1_in==data_simple[which
    (graf.r$r<graf.r$LCL)[i],]$temp_ch1_in),1:2]}
    outofcontrol

    X<-data_new
    samplemean<-apply(X,2,mean)
    covX<-cov(X)
```

```
solve(covX)
m_distance <- mahalanobis(X,samplemean,covX)
m_depth <- 1/(1 + m_distance)
r<-rangoprof(m_depth)
graf.r<-graf.control.rQ(r,alpha=0.01,n=1) #n=1 para gráfico r
table(graf.r$r<graf.r$LCL)
# Detección:
data_complete[which(data_complete$temp_ch2_in==data_simple
[which.min(graf.r$r),]$temp_ch2_in),]$date
outofcontrol<-list()
for (i in 1:(table(graf.r$r<graf.r$LCL)[2])){
outofcontrol[[i]]<-data_complete[which(data_complete$temp_ch2_in==
data_simple[which(graf.r$r<graf.r$LCL)[i],]$temp_ch2_in &
data_complete$temp_ch1_in==data_simple[
which(graf.r$r<graf.r$LCL)[i],]$temp_ch1_in),1:2]}
outofcontrol

# Gráfica de control Q:
X<-data_simple
samplemean<-apply(X,2,mean)
covX<-cov(X)
solve(covX)
library(depth)
m_distance <- mahalanobis(X,samplemean,covX)
m_depth <- 1/(1 + m_distance)
r<-rangoprof(m_depth)
graf.q<-graf.control.rQ(r,n=2,alpha=0.01)
table(graf.q$r<graf.q$LCL)
141/(2348+141)*100
# Detección:
data_complete[which(data_complete$temp_ch2_in==data_simple
[which.min(graf.q$r),]$temp_ch2_in),]$date
outofcontrol<-list()
for (i in 1:(table(graf.q$r<graf.q$LCL)[2])){
outofcontrol[[i]]<-data_complete[which(
data_complete$temp_ch2_in==data_simple[which
(graf.q$r<graf.q$LCL)[i],]$temp_ch2_in &
data_complete$temp_ch1_in==data_simple[which(graf.q$r<
graf.q$LCL)[i],]$temp_ch1_in),1:2]}
outofcontrol

X<-data_new
samplemean<-apply(X,2,mean)
covX<-cov(X)
solve(covX)
m_distance <- mahalanobis(X,samplemean,covX)
m_depth <- 1/(1 + m_distance)
r<-rangoprof(m_depth)
graf.q<-graf.control.rQ(r,n=2,alpha=0.01)
table(graf.q$r<graf.q$LCL)
```

```
# Gráfica de contro S^*:
X<-data_simple
samplemean<-apply(X,2,mean)
covX<-cov(X)
solve(covX)
library(depth)
m_distance <- mahalanobis(X,samplemean,covX)
m_depth <- 1/(1 + m_distance)
r<-rangoprof(m_depth)
graf.s<-grafcontrolS(r,alpha=0.01)
table(graf.s$Sn<graf.s$LCL)
# Detección:
outofcontrol<-list()
for (i in 1:(table(graf.s$Sn<graf.s$LCL)[2])){
outofcontrol[[i]]<-data_complete[which(
data_complete$temp_ch2_in==data_simple[
which(graf.s$Sn<graf.s$LCL)[i],]$temp_ch2_in &
data_complete$temp_ch1_in==data_simple[which(
graf.s$Sn<graf.s$LCL)[i],]$temp_ch1_in),1:2]}
outofcontrol
options(max.print=22000)
X<-data_new
samplemean<-apply(X,2,mean)
covX<-cov(X)
solve(covX)
m_distance <- mahalanobis(X,samplemean,covX)
m_depth <- 1/(1 + m_distance)
r<-rangoprof(m_depth)
graf.s<-grafcontrolS(r,alpha=0.01)
table(graf.s$Sn<graf.s$LCL)
```

Gráficos de control para datos autocorrelados

```
# Definición de la función:
T2.robusto=function(x,method=c("MCD","MVE","Trimmed"),alpha,beta){
library(MASS)
n=nrow(x)
p=ncol(x)
set.seed(1234)
#  MÉTODO MCD
if(method=="MCD"){
sigma.inv=solve(cov.mcd(x)$cov)
med=cov.mcd(x)$center
med=matrix(rep(med,n),nrow=n,ncol=p,byrow=TRUE)
dif=x-med
dif=as.matrix(dif)
dift=t(dif)

T=numeric()
```

```
for(i in 1:n){
T[i]=(dif[i,])%*%sigma.inv%*%(dift[,i])}

plot(T,type="o",main="Método␣MCD",xlab="",
ylab=expression(paste(T^2)),lwd=2)
L=(n-1)^2/n*(qbeta(0.0027, p/2, (n-p-1)/2, lower.tail = FALSE))
abline(h=L,col=2,lwd=2)
return(T)}

#  MÉTODO MVE
if(method=="MVE"){
sigma.inv=solve(cov.mve(x)$cov)
med=cov.mve(x)$center
med=matrix(rep(med,n),nrow=n,ncol=p,byrow=TRUE)
dif=x-med
dif=as.matrix(dif)
dift=t(dif)

T=numeric()

for(i in 1:n){
T[i]=dif[i,]%*%sigma.inv%*%dift[,i]}
plot(T,type="o",main="Método␣MVE",xlab="",
ylab=expression(paste(T^2)),lwd=2)
L=(n-1)^2/n*(qbeta(0.0027, p/2, (n-p-1)/2, lower.tail = FALSE))
abline(h=L,col=2,lwd=2)
return(T)}

#  MÉTODO MEDIAS TRUNCADAS
if(method=="Trimmed"){
a_media=function(x,alpha){
n=length(x)
x=sort(x)
a=floor(n*alpha/100)
mean(x[(a+1):(n-a)])}

a_trunc=numeric()
for (i in 1:p){
a_trunc[i]=a_media(x=x[,i],alpha=alpha[i])}


S=matrix(0,nrow=p,ncol=p)
cbeta=matrix(0,nrow=p,ncol=p)

for (u in 1:p){
for (v in 1:p){
if(beta[u,v]==0) {cbeta[u,v]<-1}
if(beta[u,v]!=0){
y=rchisq(1000,1)
y=sort(y)
b=floor(beta[u,v]*1000/100)
```

```
    cuv=1/mean(y[(b+1):(1000-b)])
    cbeta[u,v]=1000*cuv/(1000-1)}
    S[u,v]=cbeta[u,v]*a_media((x[,u]-a_trunc[u])*(x[,v]-a_trunc[v]),
    beta[u,v])}}

    med=matrix(rep(a_trunc,n),nrow=n,ncol=p,byrow=TRUE)
    dif=x-med
    dif=as.matrix(dif)
    dift=t(dif)
    sigma.inv=solve(S)

    T=numeric()

    for(i in 1:n){
    T[i]=dif[i,]%*%sigma.inv%*%dift[,i]}
    L=(n-1)^2/n*(qbeta(0.0027,p/2,(n-p-1)/2,lower.tail=FALSE))
    plot(T,type="o",main="Método␣Medias␣Truncadas",ylim=c(0,max(T)),
    xlab="",ylab=expression(paste(T^2)),lwd=2)
    L=(n-1)^2/n*(qbeta(0.0027,p/2,(n-p-1)/2,lower.tail=FALSE))
    abline(h=L,col=2,lwd=2)
    return(T)}}


    # Estimación de alpha y beta:
    alpha=c(8,15,0,20);alpha
    beta=matrix(c(15,0,0,0.15,0,0,0,0,0,0,0.1,0.2,0.15,0,0.2,50),
    nrow=4,ncol=4,byrow=TRUE)

    data[1000:1200,c(1,9,10,11,12)]
    apply(misdatos,2,sd)
    colMeans(misdatos)

    # Aplicación de la función:
    par(mfrow=c(1,3))
    n<-nrow(misdatos)
    p<-ncol(misdatos)
    L=(n-1)^2/n*(qbeta(0.0027,p/2,(n-p-1)/2,lower.tail=FALSE))
    Tmcd=T2.robusto(misdatos,method="MCD")
    abline(h=L,col=2,lwd=2)
    Tmve=T2.robusto(misdatos,method="MVE")
    abline(h=L,col=2,lwd=2)
    Ttrunc=T2.robusto(misdatos,method="Trimmed",
    alpha=alpha,beta=beta)
    abline(h=L,col=2,lwd=2)
    par(mfrow=c(1,1))

    # Identificación de los "fuera de control":
    table(data[which(Tmcd>L),1])
    table((Tmcd>L))
    table(data[which(Tmve>L),1])
    table((Tmve>L))
```

```
table(data[which(Ttrunc>L),1])
table((Ttrunc>L))

# Gráficos completos:
par(mfrow=c(1,3))
n<-nrow(misdatos)
p<-ncol(misdatos)
L=(n-1)^2/n*(qbeta(0.0027,p/2,(n-p-1)/2,lower.tail=FALSE))
plot(Tmcd,type='n',ylim=c(0,max(Tmcd)),lwd=2,main='MCD')
lines(1:n,Tmcd,type='o',lwd=2)
abline(h=L,col=2,lwd=2)
plot(Tmve,type='n',ylim=c(0,max(Tmve)),lwd=2,main='MVE')
lines(1:n,Tmve,type='o',lwd=2)
abline(h=L,col=2,lwd=2)
plot(Ttrunc,type='n',ylim=c(0,max(Ttrunc)),lwd=2,
main='Medias␣truncadas')
lines(1:n,Ttrunc,type='o',lwd=2)
abline(h=L,col=2,lwd=2)
par(mfrow=c(1,1))

# Gráficos ampliados a escala:
par(mfrow=c(1,3))
L=(n-1)^2/n*(qbeta(0.0027,p/2,(n-p-1)/2,lower.tail=FALSE))
plot(Tmcd,type='n',ylim=c(0,250),lwd=2,main='MCD')
lines(1:n,Tmcd,type='o',lwd=2)
abline(h=L,col=2,lwd=2)
plot(Tmve,type='n',ylim=c(0,250),lwd=2,main='MVE')
lines(1:n,Tmve,type='o',lwd=2)
abline(h=L,col=2,lwd=2)
plot(Ttrunc,type='n',ylim=c(0,250),lwd=2,
main='Medias␣truncadas')
lines(1:n,Ttrunc,type='o',lwd=2)
abline(h=L,col=2,lwd=2)
par(mfrow=c(1,1))
```

Bibliografía

[1] Abdi H. y Williams L. *Principal Component Analysis.* John Wiley & Sons, 2010

[2] Albert P.B. y Xavier T.M. *Métodos estadísticos. Control y mejora de la calidad.* Universitat Politecnica de Cataluña, 2004

[3] Alt F.B. *Multivariate quality control. In Encyclopedia of Statistical Sciences.* Edited by Kotz and Johnson, Vol. 6, New York. John Wiley & Sons, 1985

[4] Altaf B.A. *Application of Dynamic Partial Least Squares to Complex Processes.* School of Chemical Engineering and Advanced Materials. Newcastle University, 2013

[5] Apley D.W. y Tsung F. The autoregressive T-squared chart for monitoring univariate autocorrelated processes. *Journal of Quality Technologies*, Vol. 34, 80-96, 2002

[6] Baillo A. y Granó A. *100 Problemas resueltos de Estadística Multivariante implementados en Matlab.* Delta Publicaciones, 2007

[7] Box G.E.P. y Cox D.R. An analysis of transformations. *Statistical Society*, Vol. 26(2), 211-252, 1964

[8] Box G.E.P. y Jenkins G. *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco (1976)

[9] Brook D. y Evans D. An Approach to the Probability Distribution of Cusum Run Length. *Biometrika*, Vol. 59(3), 539-549, 1972

[10] Chou Y.M. y Polansky A.M. Transforming non-normal data to normality in statistical process control. *Journal of Quality Technologies*, Vol. 30(2), 133-141, 1998

[11] Crosier R. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics*, Vol. 30(3), 291-303, 1988

[12] Cuesta-Albertos JA., Fraiman R. y Ransford T. Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Boletim da Sociedade Brasileira de Matematica*, Vol. 37, 1-25, 2007

[13] Cuevas A., Febrero M. y Fraiman R. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, Vol. 22(3), 481-496, 2007

[14] Doganaksoy N., Faltin F.W. y Tucker W.T. Identification of out-of-control multivariate characteristic in a multivariable manufacturing environment. *Common Statististical Methodes*, Vol. 20, 2775-2790, 1991

[15] Dong Y. y Qin S.J. Dynamic Inner Partial Least Squares for Dynamic Data Modeling. *International Federation of Automatic Control*, 117-122, 2015

[16] Duffy J. y Spatz S. *Static Switching Today, Electrical manufacturing.* General Electric, Application Manual, Transistorized Static Control, 1959

[17] Edgar S.F. y Michele S. MPCI: An R Package for Computing Multivariate Process Capability Indices. *http://www.jstatsoft.org Journal of Statistical Software, 47(7), 1-15, 2012*

[18] Efron B. The Jackknife, the Bootstrap and other Resampling Plans. *CMBF-NSF Regional Conference Series in Applied Mathematics*, New York, Vol. 83, 353-360, 1982

[19] Ferrer A. Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC PCA): Some Rejections and a Case Study in Autobody Assembly Process. *Quality Engineering*, 311-325, 2007

[20] Fraiman R. y Muniz G. *Trimmed means for functional data.* Test 10, 419-440, 2001

[21] García-Díaz J. y Aparisi F. *Optimización de los gráficos de control estadístico de procesos EWMA y MEWMA mediante algoritmos genéticos.* Congreso Nacional de Estadística e Investigación Operativa, 2003

[22] Healy J.D. A note on multivariate CUSUM procedures. *Technometrics*, Vol. 29, 409-412, 1987

[23] Henze N. y Zirkler B. A class of invariant consistent tests for multivariate normality. *Theory Methods*, Vol. 19(10), 3595-3617, 1990

[24] Holmes D.S. y Mergen A.E. Improving the performance of T-square control chart. *Quality Engineering*, Vol. 5(4), 619-625, 1993

[25] Hotelling H. *Multivariate Quality Control.* McGraw-Hill, New York, 1947

[26] Hubele N.F., Shahriari H. y Cheng C.S. A Bivariate Process Capability Vector in Statistical Process Control in Manufacturing. *J.B and Montgomery*, New York, 299-310, 1991

[27] Jackson J. *A Users Guide to Principal Components.* John Wiley & Sons, New York, 1991

[28] Jolliffe I. *Principal Component Analysis.* Springer, New York, 2002

[29] Johnson N.L. Systems of frequency curves generated by methods of translation. *Biometrika*, Vol. 36, 149-176, 1949

[30] Kalagonda A.A. y Kulkarni S.R. Multivariate quality control chart for autocorrelated processes. *Applied Statistics*, Vol. 31(3), 317-327, 2004

[31] Kaspar M. y Ray W. Dynamic pls modelling for process control. *Chemical Engineering Science*, 3447-3461, 1993

[32] Kresta J., MacGregor J. y Marlin T. Multivariate statistical monitoring of process operating performance. *Chemical Engineering Science*, 35-47, 1995

[33] Krzysztof Ciupke. Multivariate Process Capability Vector Based on One-Sided Model. *Quality and Reliability Engineering*, Vol. 31(2), 313-327, 2015

[34] Ku W., Storer R.H. y Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 179-196, 1995

[35] Lakshminarayanan S., Shah S. y Nandakumar K. Modeling and control of multivariable processes: dynamic pls approach. *AIChE Journal*, 2307-2322, 1997

[36] Lowry C.A. y Montgomery D.C. A review of multivariate control charts. *IIE Trans*, Vol. 27(6), 800-810, 1995

[37] Lowry C.A., Woodall W.H., Champ C.W. y Rigdon S.E. A multivariate exponentially weighted moving average control chart. *Technometrics*, Vol. 34(1), 46-53, 1992

[38] MacGregor J. y Kourti T. Statistical process control of multivariate processes. *Pergamon*, Vol. 3, 403-414, 1995

[39] Mardia K.V. Measures of multivariate skewness and kurtosis. *Biometrika*, Vol. 57, 519-530, 1970

[40] Mason T.J., Paniwnyk L. y Lourimer J.P. Theóusesíof ultrasound in food technology. *Ultrasonics Sonochem*, Vol. 3, 253-260, 1996

[41] Miguel Flores qcr: Quality Control Review. R package version 1.0. *https://CRAN.R-project.org/package=qcr*

[42] Murphy B.J. *Selecting Out of Control variables with the t2 multivariate quality control procedure.* Journal of the Royal Statistical Society Serie D (The Statistician), Vol. 36, 571-583, 1987

[43] Nickerson D.M. Construction of a conservative confidence region from projections of an exact confidence region in multiple linear regression. *The American Statistician*, Vol. 48(2), 120-124, 1994

[44] Otoole A., Abdi H., Deffenbacher K. y Valentin D. A low dimensional representation of faces in the higher dimensions of the space. *Journal of Optical Society of America*, 405-411, 1993

[45] Page E.S. Cumulative sum charts. *Technometrics*, Vol. 3(1), 1-9, 1993

[46] Pan J.N. y Lee C.Y. New capability indices for evaluating the performance of multivariate manufacturing processes. *Quality and Reliability Engineering International*, Vol. 26(1), 3-15, 2010

[47] Peres-Neto P., Jackson D. y Somers K. How many principal components stopping rules for determinin the number of non-trivial axes revisited. *Computational Statistics Data Analysis*, 974-997, 2005

[48] Pignatiello J. y Runger G.C. Comparisons of Multivariate CUSUM Charts. *Journal of quality technology*, Vol. 22(3), 173-186, 1990

[49] Qin S. y McAvoy T. Nonlinear r modeling via a neural net pls approach *Computers & chemical engineering*, Vol. 20(2) , 147-159, 1996

[50] Quenouille M. Notes on bias and estimation. *Biometrika*, 353-360, 1956

[51] Rato T. y Reis M. *Defining the structure of dpca model and its impact on process monitoring and prediction activities.* Chemometrics and Intelligent Laboratory Systems, 2013

[52] Regina y Liu. Control charts for multivariate processes. *Journal of the American Statistical Association*, Theory and Methods, Vol. 90(432), 1995

[53] Roberts S.W. Control chart tests based on geometric moving averages. *Technometrics*, Vol. 42(1), 97-102, 1959

[54] Royston J.P. Approximating the Shapiro-Wilk W-Test for non-normality. *Statistics and Computing*, Vol. 2(3), 117-119, 1992

[55] Scrucca L. qcc: an R package for quality control charting and statistical process control. *https://cran.r-project.org/doc/Rnews/ News 4/1, 11-17, 2004*

[56] Shapiro S. y Wilk M. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, Vol. 52(34), 591-611, 1965

[57] Shahriari H., Hubele N.F. y Lawrence F.P. A multivariate process capability vector. *Proceedings of the 4th Industrial Engineering Research Conference*, Vol. 1, 304-309, 2009

[58] Slifker J.F. y Shapiro S.S. The Johnson system: selection and parameter estimation. *Technometrics*, Vol. 22(2), 239-246, 1980

[59] Sullivan J.H. y Woodall W.H. A comparison of multivariate control charts for individual observations. *Journal of Quality Technologies*, Vol. 28, 398-408, 1996

[60] Taam W., Subbaiah P. y Liddy W.J. A note on multivariate capability indices. *Journal of Applied Statistics*, Vol. 20, 339-351, 1993

[61] Thode H.C. Testing for normality. *Statistics*, Vol. 164, 2002

[62] Thomas Roth qualityTools: Statistics in Quality Science. R package version 1.55. *http://www.r-qualitytools.org*

[63] Vanhatalo E. y Kulahci M. *Impact of correlation on principal components and their used in statistical process control.* Quality and Reliability Engineering International, Vol. 32(4), 2015

[64] Vanhatalo E., Kulahci M. y Bergquist B. *On the structure of dinamic principal component analysis used in statistical process monitoring.* Chemometrics and Intelligent Laboratory Systems, 1-11, 2017

[65] Wang C.H. Constructing multivariate process capability indices for short-run production. *International Journal of Advanced Manufacturing Technology*, Vol. 26(11-12), 1306-1311, 2005

[66] Wang F.K. y Chen J.C. Capability index using principal components analysis. *Quality Engineering*, Vol. 11, 21-27, 1998

[67] Wierda S.J. *Multivariate statistical process control.* Statistica Neerlandica

[68] Wold S. Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics*, Vol. 20(4), 397-405, 1978

[69] Wold H. Soft modelling, The basic design and some extensions. *Systems Under In direct Observation*, North-Holland, Amsterdam, Vols. I and II, 211-228, 1982

[70] Wold S., Ruhe A., Wold H. y Dunn W. The collinearity problem in linear regression, The partial least squares approach to generalized inverses. *SIAM Journal on Scientific Computing*, Vol. 5, 735-743, 1984

[71] Wold S. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, Vol. 2(1-3), 37-52, 1987

[72] Woodall W.H. y Ncube M.M. Multivariate CUSUM quality-control procedures. *Technometrics*, Vol. 3(3), 285-292, 1985

[73] Xekalaki E. y Perakis M. The use of principal component analysis in the assessment of process capability indices. *Proceedings of the Joint Statistical Meetings of the American Statistical Association*, The Institute of Mathematical Statistics, The Canadian Statistical Society, New York, 2002