Universitat de Girona

GRÀFIC DE CONTROL T2 DE HOTELLING PER A DADES COMPOSICIONALS

Marina Vives Mestres

Dipòsit legal: Gi. 2039-2014
http://hdl.handle.net/10803/284756

Universitat de Girona

TESI DOCTORAL

Gràfic de control T^2 de Hotelling per a dades composicionals

Marina Vives Mestres

2014

Universitat de Girona

TESI DOCTORAL

Gràfic de control T^2 de Hotelling per a dades composicionals

Marina Vives Mestres

2014

Programa de Doctorat en Tecnologia

Directors

Dr. Josep A. Martín Fernández

i

Dr. Josep Daunis i Estadella

Memòria presentada per optar al títol de doctora per la Universitat de Girona

Universitat de Girona

El Dr. Josep Antoni Martín Fernández i el Dr. Josep Daunis i Estadella, professors del departament d'Informàtica, Matemàtica Aplicada i Estadística de la Universitat de Girona,

DECLAREM:

Que el treball titulat *Gràfic de control T^2 de Hotelling per a dades composicionals*, que presenta Marina Vives Mestres per a l'obtenció del títol de doctora, ha estat realitzat sota la nostra direcció.

I, perquè així consti i tingui els efectes oportuns, signem aquest document.

Signatures,

Josep A. Martín Fernández

Josep Daunis i Estadella

Girona, 15 de juliol de 2014

A l'Aina,
la llavor de la meva vida,
i a l'Albert,
per acompanyar-me pel camí.

Agraïments

Primer de tot agrair a l'Albert, la meva parella i company de viatge, pels esforços que ha fet per intentar entendre el contingut de la tesi, per alegrar-se dels nous avenços i encoratjar-me en els moments baixos. No deu haver estat fàcil suportar l'envestida de la tempesta, sobretot cap al final de la tesi, quan els nervis estaven a flor de pell. Gràcies per ser al meu costat.

Gràcies a la meva mare i a en Salvador per tenir cura de la meva princesa mentre treballava, pel suport i l'interès en el significat dels gargots que poblaven el meu escriptori. També gràcies a la Mª Àngels, en Josep, en David i la Quimeta, per tots els dimarts que m'han dedicat, els innombrables "tuppers" i els sopars de pernil. Sé que és difícil per tots vosaltres entendre aquesta feina, els horaris asimètrics, tantes hores davant l'ordinador... cosa que fa que encara tingui més mèrit el vostre suport.

Gràcies a les crisàlides, que ara ja són papallones: l'Estel (i la seva família), la Isabel (i la seva família), la Marta, la Carlota, l'Anna, la Oriana, la Ingrid, la Iris, la Natàlia i la Carla. Heu estat un suport inestimable, vosaltres i les vostres petites llavors que han fet que es creuessin els nostres camins. M'heu ensenyat a escoltar, a compartir i a expressar des del cor. La vivència amb vosaltres no la oblidaré mai.

Gràcies als companys de la UdG: en Santi, la Glòria, la Vera, en Carles, en Marc, l'Iván, la Natàlia i molt especialment als meus tutors Martin i Pepus. Gràcies Carles pel teu interès i suport a les noves generacions, ets realment un exemple a seguir. Glòria, gràcies per ser el referent més proper dels passos a seguir i per mostrar-te sempre tant disposada. Marc i Iván, gràcies per la "camaderia" de despatx, realment és un gust treballar al vostre costat. Gràcies a la Berta per haver-me servit d'exemple a la fase final. També vull donar les gràcies als companys de departament que s'han interessat per la tesi.

Vull dedicar un paràgraf d'agraïment a en Martin i a en Pepus, els tutors de tesi. Des del dia que em vas fer l'entrevista, Martin, vaig veure que eres un apassionat del món CoDa. Malgrat no entendre massa les implicacions del terme "composicional" o "restringit" o "símplex" (fins al moment per mi era un mètode d'optimització), vaig veure clar que de la teva convicció sobre el potencial CoDa en podia aprendre molt. És un luxe i un plaer treballar amb tu, per la teva capacitat de síntesi, de comprensió i d'expressió. Espero poder

continuar aprenent al teu costat. Pepus, tu m'has aportat la vessant més humana i m'has donat suport en els moments més durs i de replantejament de futur. Gràcies per mostrar la cara més divertida de l'estadística, per les teves explicacions planeres, les tertúlies, les bromes i el bon humor. Feu la parella de tutors perfecte.

Aquesta tesi ha estat finançada mitjançant una beca de recerca (BR) de la Universitat de Girona. També a través dels projectes de recerca: Research Group in COmpositional DAta [RGCODA] de l' Agència de Gestió d'Ajuts Universitaris i de Recerca (Ref: 2009SGR424), Análisi Estadístico de Datos Composicionales y Otros Datos con Espacio Muestral Restringido del Ministerio de Ciéncia e Innovación (Ref: MTM2009-13272) i Métodos Estadísticos en Espacios Restringidos del Ministerio de Ciéncia e Innovación (Ref: MTM2012-33236).

Publicacions

La present tesi es presenta com a compendi dels següents articles:

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014), **Individual T^2 Control Chart for Compositional Data**. Journal of Quality Technology, 46(2), pp. 127-139.
  Factor d'impacte: primer quartil (Q1) del *Journal Citation Report* (JCR) de l'*Institute of Scientific Information*.

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014), **Out-of-Control Signals in Three-Part Compositional T^2 Control Chart**. Quality and Reliability Engineering International, 30 (3), pp. 337-346.
  Factor d'impacte: tercer quartil (Q3) del *Journal Citation Report* (JCR) de l'*Institute of Scientific Information*.

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. **Signal Interpretation in Hotelling's T^2 Control Chart for Compositional Data**. Submitted to IIE Transactions.
  Factor d'impacte: segon quartil (Q2) del *Journal Citation Report* (JCR) de l'*Institute of Scientific Information*.

A banda d'aquests articles en revistes, del treball de la tesi se'n desprenen les següents aportacions a congressos:

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2013), **Interpretation of out-of-control signals in a compositional T^2 control chart**. Abstracts of the Workshop on Compositional Data Analysis (CoDaWork 2013), Vorau, Austria, 4-7 June 2013.

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2012), **Compositional T^2 control chart: Interpretation of out of control signals**. Abstracts of the 12th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS-12), Ljubljana, Slovenia, 9-13 September 2012.

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2012), **A theoretical approach to T^2 control chart for compositional data**. Book of abstracts of XIII Chemometrics in Analytical Chemistry. Budapest, Hungary, 25-29 June 2012.

- Vives-Mestres, M., Daunis-i-Estadella, J., and Martín-Fernández, J. A. (2012), **Gráfico T^2 de Hotelling para datos composicionales**. XXXIII Congreso Nacional de Estadística e Investigación Operativa y de las VII Jornadas de Estadística Pública. Madrid, 17-20 April 2012.

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2011), **P Control Charts: a new approach**. Proceedings of the 11th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS-11), CD-ROM. Coimbra, Portugal, 5-7 September 2011.

- Vives-Mestres, M., Daunis-i-Estadella, J., and Martín-Fernández, J. A. (2011), **Could be the CODA methodology useful in control chart techniques?**. Abstracts of the Workshop on Compositional Data Analysis (CoDaWork 2011), Sant Feliu de Guíxols, Girona, 9-13 May 2011.

Llista d'abreviatures

alr	Additive Log-Ratio
ARL	Average Run Length
CC	Control Chart
CCL	Center Control Line
clr	Centered Log-Ratio
CoDa	Compositional Data
CUSUM	Cumulative Sum
EWMA	Exponentially Weighted Moving Average
ilr	Isometric Log-Ratio
LCL	Lower Control Line
MYT	Descomposició de Mason-Young-Tracy
NN	Nearest Neighbour
RL	Run Length
SBP	Sequential Binary Partition
SPC	Statistical Process Control
UCL	Upper Control Line

Índex de figures

Índex de taules

Índex

Resum

La present tesi estudia el gràfic de control multivariant T^2 de Hotelling per al control de dades composicionals (CoDa). Una composició és un vector d'elements estrictament positius que sumen una constant. Les CoDa tenen la peculiaritat de viure en un espai restringit.

En un primer estadi, hem analitzat les propostes trobades a la literatura per controlar, mitjançant l'estadístic T^2, processos en els que la característica de qualitat és una composició. Hem determinat que aquestes propostes incompleixen el principi de coherència subcomposicional que ha de complir tota anàlisi de CoDa. Aquest principi estableix que la inferència sobre subcomposicions (una part de la composició) ha de ser consistent, independentment de si la inferència es basa en la subcomposició o la composició completa. D'altra banda, també hem vist com les regions de control quedaven dibuixades fora de l'espai restringit.

En base a les observacions prèvies, hem proposat un nou gràfic de control que hem anomenat T^2 composicional i anotat T^2_C. Aquest es basa en una representació de les CoDa en coordenades a l'espai real, on llavors es calcula l'estadístic T^2. Aquestes coordenades es calculen mitjançant transformacions logràtio dels components.

Hem utilitzat indicadors de funcionament per comparar el mètode de control clàssic amb el mètode composicional i hem vist que el T^2_C tenia una taxa menor de falses alarmes i que aquesta es mantenia constant fos quin fos el centre de la distribució de les dades.

En el control de processos, és important identificar els senyals fora de control però també ho és identificar les causes del senyal per tal de poder dur a terme accions correctores. És per això que hem proposat un mètode gràfic per facilitar la interpretació dels senyals fora de control en el cas de les composicions de tres parts.

Finalment, presentem un algoritme que permet identificar les causes del senyal per un nombre qualsevol de components. Tant l'algoritme com el mètode gràfic es basen en trobar quin és el logràtio de components que, de forma univariant, contribueix més al valor global de la T^2_C.

Al llarg de la tesi hem aplicat els mètodes i conceptes proposats a exemples simulats i a exemples reals industriats extrets de la literatura.

Resumen

La presente tesis estudia el gráfico de control multivariante T^2 de Hotelling para el control de datos composicionales (CoDa). Una composición es un vector de elementos estrictamente positivos de suma constante. Los CoDa tienen la peculiaridad de vivir en un espacio restringido.

En un primer estadio, hemos analizado las propuestas encontradas en la literatura para controlar, mediante el estadístico T^2, procesos en los que la característica de calidad es una composición. Hemos determinado que estas propuestas incumplen el principio de coherencia subcomposicional que debe cumplir todo análisis CoDa. Este principio establece que la inferencia sobre subcomposiciones (parte de una composición) debe ser consistente, independientemente de si la inferencia se basa en la subcomposición o la composición completa. Por otro lado, también hemos visto como las regiones de control caen fuera del espacio restringido.

En base éstas observaciones, hemos propuesto un nuevo gráfico de control que hemos llamado T^2 composicional (T^2_C). Éste se basa en una representación de los CoDa en coordenadas en el espacio real, done luego se calcula el estadístico T^2. Las coordenadas se calculan mediante transformaciones logràtio de los componentes.

Hemos utilizado indicadores de funcionamiento para comparar el método de control clásico con el método composicional y hemos visto que el T^2_C tenía una menor tasa de falsas alarmas y que ésta se mantenía constante fuera cuál fuera el centro de la distribución de los datos.

En el control de procesos, es importante identificar las señales fuera de control, pero también lo es identificar las causas de la señal para poder llevar a cabo acciones correctoras. Es por ello que hemos propuesto un método gráfico para facilitar la interpretación de las señales fuera de control en el caso de las composiciones de tres partes.

Finalmente, presentamos un algoritmo que permite identificar las causas del señal para cualquier número de componentes. Tanto el algoritmo como el método gráfico se basan en encontrar el logràtio de componentes que, de forma univariante, contribuyen mas al valor global de la T^2_C.

A lo largo de la tesis hemos aplicado los métodos y conceptos propuestos a ejemplos simulados y a ejemplos reales industriales encontrados en la literatura.

Abstract

This thesis studies the multivariate Hotelling T^2 control chart for monitoring compositional data (CODA). A composition is a vector of positive elements that add to a constant sum. CoDa have the peculiarity of living in a restricted space.

Firstly, we analyze the proposals found in the literature to control processes in which the quality characteristic is a composition by the use of the T^2 statistic. We have determined that these proposals violate the principle of subcomposicional coherence. This principle states that the inference on a subcomposition (a part of a composition) should be consistent regardless of whether the inference is based on the entire composition or a subcomposition. Furthermore, we have also seen that the control regions fall out of the sample space.

Based on these observations, we have proposed a new control chart called compositional T^2 control chart and denoted by T_C^2. It is based on a representation of CoDa onto coordinates in the real space, where the T^2 statistic can be calculated. The coordinates are calculated by the use of a logratio transformation of the components.

We used performance indicators to compare the classical method with the compositional method, and we have seen that the T_C^2 has a lower false alarm rate, which remained constant regardless of the center of the distribution of the data.

In process control, it is not only important to identify the out of control signals, but also to identify the causes of the anomaly in order to carry out corrective actions. This is why we have proposed a graphical method to interpret the out of control signals in the case of three part compositions.

Finally, we present an algorithm to interpret the causes of the signal in the general case. Both the algorithm and the graphical method are based on finding the logratio of components that greatly contributes to the overall value of the T_C^2.

Throughout the thesis we have applied the proposed methods and concepts to simulated examples and real industrial examples from the literature.

Capítol 1

Introducció

1.1 Motivació

El treball d'investigació a desenvolupar durant la tesi s'emmarca dins d'una de les línies de recerca principals del projecte METRICS "Métodos estadísticos en espacios restringidos" (Ref: MTM2012-33236; Ministerio de Economía y Competitividad) que té com a objectiu l'"Adaptació de mètodes multivariants a dades restringides per la seva aplicació, entre d'altres, a les ciències de la vida, economia i a les ciències ambientals i de la terra". El tema de tesis doctoral forma part de la sublínia específica dedicada a les tècniques estadístiques multivariants per al control de processos en els quals la informació a controlar té caràcter composicional.

El profund coneixement en dades composicionals que ha desenvolupat el grup de recerca UdG GRCT0035 Estadística i Anàlisi de Dades (EAD), mai abans havia estat aplicat en el camp del control estadístic de processos. L'estat de l'art revisat durant aquesta tesi revela que, fins al moment, s'havien aplicat les tècniques estadístiques estàndards per a controlar dades composicionals, sense tenir en compte les peculiaritats d'aquestes, tot donant lloc a resultats poc consistents.

La unió de les dues branques de recerca, d'una banda el control estadístic de processos i de l'altra l'anàlisi de dades composicionals, obre un gran ventall de possibilitats de recerca de les quals la present tesi en representa només l'inici. Entre aquestes possibilitats hi trobem l'estudi dels gràfics de control T^2 per al control de mitjanes de composicions o bé l'estudi del disseny d'experiments per a barreges, entre d'altres.

1.2 Situació dins la recerca

Els gràfics de control (CC de l'anglès *Control Chart*) són una eina del control estadístic de processos (SPC de l'anglès *Statistical Process Control*) que permeten assegurar que un procés es manté en estat de control estadístic,

és a dir, en absència de causes assignables i on els canvis en les mesures de centre i variabilitat són estadísticament previsibles (causes comunes).

La majoria de processos requereixen el control de múltiples variables. L'ús de múltiples CC univariants no és recomanat en aquest cas, perquè s'ignora l'estructura de correlació entre les variables i no es controla la probabilitat global d'obtenir un fals fora de control (Montgomery, 2009). Basant-se en aquestes necessitats es va desenvolupar el control estadístic de processos multivariants. Es poden trobar revisions dels avenços més significatius a Lowry and Montgomery (1995), MacGregor and Kourti (1995), Reynolds and Cho (2006), Bersimis et al. (2007) i en el cas específic del control d'atributs Topalidou and Psarakis (2009).

El CC multivariant més elemental és el de la T^2 de Hotelling (Hotelling, 1947). Aquest gràfic utilitza com a estadístic la distància de Mahalanobis entre l'observació (o la mitjana d'un grup d'observacions) i la mitjana del procés tot tenint en compte la correlació entre les variables. Més informació sobre l'ús de la T^2 es pot trobar a Tracy et al. (1992), Fuchs and Kenett (1998) i Montgomery (2009).

La interpretació dels senyals fora de control dels CC multivariants resulta important per tal de poder dur a terme accions correctives per millorar el procés. Aquesta tasca resulta especialment complicada en l'anàlisi multivariant pel fet que es condensa en un únic estadístic la informació referent a un conjunt de variables. De forma general, una observació fora de control en un CC multivariant pot indicar un canvi en el centre o variabilitat d'una o més variables, o bé un canvi en la relació entre múltiples variables, o fins i tot una combinació d'ambdós efectes.

Pel cas concret del CC T^2 de Hotelling, s'han desenvolupat diferents mètodes per interpretar els senyals fora de control. Un resum dels mètodes més destacats així com una comparació entre ells es presenta a Das and Prakash (2007). Un d'ells és el proposat per Mason et al. (1995), conegut com el mètode de descomposició MYT (dels seus creadors Mason-Young-Tracy) que es basa en la descomposició de l'estadístic T^2 en parts ortogonals directament interpretables. El mètode MYT està descrit amb detall i exemples al llibre Mason and Young (2002).

Malgrat que el control de processos s'ha aplicat majoritàriament a línies de producció, en els darrers temps s'ha avançat molt en aplicacions a l'enginyeria, ciències ambientals, biologia, genètica, epidemiologia, medicina, finances, aplicació de la llei i atletisme (Montgomery, 2009). És d'especial interès l'aplicació dels gràfics de control a l'atenció sanitària i la vigilància de la salut pública (Woodall, 2006), per exemple, per al control de les ràtios d'infeccions o bé dels temps d'espera en un hospital. En resum, es pot aplicar a qualsevol procés amb una sortida mesurable.

La present tesi es centra en el control de processos en els que la característica de qualitat és una composició. Una composició és un vector d'elements estrictament positius que sumen una constant. Són exemples

de variables composicionals aquelles que es mesuren en percentatges, parts per u, ppm, molaritats o bé qualsevol altra unitat de concentració. Les dades composicionals tenen la peculiaritat de viure en un espai restringit, anomenat símplex, degut a la suma constant dels seus components.

Les dades composicionals (CoDa de l'anglès *Compositional Data*) es troben àmpliament a la indústria química, petroquímica, farmacèutica, alimentària ... També les trobem en moltes i diverses aplicacions com ara en l'anàlisi de l'ús del temps (sociologia), la composicions de minerals a les roques (geologia), l'abundància d'espècies (biologia), la distribució dels recursos d'una empresa entre departaments (economia), percentatges de població (demografia)... Totes tenen en comú que les dades descriuen quantitativament les parts d'un total.

La diversitat tant dels camps d'aplicació dels gràfics de control com de l'existència de dades composicionals, fa que la unió de les dues àrees conformin una línia de recerca prometedora. La present tesi inicia aquesta línia de recerca amb l'adaptació del gràfic T^2 de Hotelling per al control de composicions.

El primer inconvenient que es troba l'investigador en intentar elaborar un gràfic T^2 per controlar un conjunt de CoDa amb la tècnica tradicional, és que no pot calcular l'estadístic T^2 perquè la matriu de covariàncies és singular, i per tant no invertible mitjançant els mètodes estàndards. Aquest fet ve donat per la suma constant dels components. Hem trobat a la literatura tres escenaris d'aplicació del gràfic T^2 per al control de CoDa;

- En els casos en els que hi hagi colinearitat entre les variables, Mason and Young (2002) proposen eliminar una de les variables implicada en la colinearitat i calcular l'estadístic T^2 amb les variables restants. Una altra proposta, equivalent, consisteix en reconstruir la matriu de covariàncies tot eliminant els vectors propis associats als valors propis nuls o gairebé nuls, i calcular l'estadístic T^2 amb els més grans.

- Es pot donar el cas que existeixi colinearitat entre les variables però que aquesta no sigui detectada. Aquest fet pot ser causat per errors de mesura en les dades que facin que la matriu de covariàncies sigui gairebé singular. Aquest fet ja ha estat estudiat prèviament (Mason and Young, 2002) i s'ha alertat que l'estadístic T^2 queda greument afectat, de forma que els punts fora de control ja no són creïbles i el procés de control perd sentit.

- En un darrer cas, considerem aquells processos en els que es mesura només una part de la composició, i que per tant, la suma dels components no és constant. Aquest és el cas més habitual a la literatura revisada, alguns exemples es poden trobar a Mason et al. (1997, 2001); Ortiz-Estarelles et al. (2001); Gonzalez-de la Parra and Rodriguez-Loaiza (2003). En aquests casos l'estadístic T^2 es pot calcular sense

problemes, però es poden donar resultats poc coherents amb les dades originals (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011).

Cap de les estratègies descrites més amunt té en compte la particularitat de les dades composicionals. Tal i com ja va anunciar Aitchison en el seu treball Aitchison (1986), les dades composicionals representen parts d'un total i per tant només contenen informació relativa, és a dir, l'única informació rellevant en una composició està continguda en les ràtios de les seves components, i no en els seus valors individuals.

Només s'han trobat a la literatura dos articles que mencionen les peculiaritats de les CoDa a l'hora d'implementar un esquema de control. El primer intent el trobem a Boyles (1997). Boyles va desenvolupar un gràfic khi quadrat per controlar dades multinomials i Dirichlet, que és la distribució més coneguda sobre el símplex. La distribució Dirichlet és molt rígida ja que exigeix una completa independència entre les components a partir de les quals es construeix. És de difícil aplicació a la pràctica, ja que no permet modelar cap estructura de dependència CoDa.

Boyles (1997) utilitza gràfics descriptius simples, com ara diagrames bivariants, per comparar el gràfic χ^2 amb el T^2 basat en una transformació logràtio. La transformació emprada per Boyles utilitza com a divisor del logràtio el darrer component de la composició. Aquesta transformació es coneix amb el nom de alr de l'anglès *additive log-ratio transformation* i el seu principal inconvenient és que no és una transformació isomètrica. L'autor conclou que el gràfic T^2 basat en logràtio és mes sensible que el χ^2 però es decanta pel χ^2 perquè "*the computational complexity of the optimal approach [...] makes it impractical in many shopfloor situations*". Des del nostre punt de vista, i després dels darrers avenços en el camp de la producció automàtica (Stoumbos et al., 2000), considerem que l'ús de l'enfoc òptim passa per sobre de la complexitat computacional.

Yang et al. (2004) fa una altra proposta per controlar dades composicionals. En el seu treball utilitzen dades dels percentatges de partícules que passen determinats tamisos per tal de controlar que la gradació d'uns agregats per a la indústria de l'asfalt es mantingui dins els límits establerts. Proposen dos mètodes per definir regions d'acceptació. El primer consisteix en elaborar múltiples gràfics de control univariants, cosa que ja hem comentat anteriorment que no és recomanable. El segon mètode, es basa en un enfoc additiu (no logràtio) i per tant no adequat per a CoDa.

La proposta d'aquesta tesi es basa en la metodologia composicional, que consisteix en analitzar els quocients o ràtios entre components, o més ben dit, els logràtios entre components (anomenats coordenades) ja que el logaritme facilita el maneig de les dades.

1.3 Presentació dels articles

Un cop feta la introducció general i repassada la literatura disponible, queda clar que hi ha una línia oberta en la recerca feta fins al moment pel que fa al control de processos on la característica de qualitat és una composició. La present tesi obre aquest camí de recerca fent les següents aportacions innovadores.

- Demostra que les propostes fetes fins al moment per a controlar composicions mitjançant la T^2 de Hotelling, i per tant per esquivar la colinearitat entre les variables, no són coherents amb la natura de les CoDa.
- Proposa un nou gràfic de control T^2 per a dades composicionals (T^2_C) que té en compte els principis bàsics de les CoDa.
- Demostra que els mètodes clàssics per interpretar el punts fora de control del gràfic T^2 no són adequats per a CoDa i fa una aproximació geomètrica a la interpretació dels senyals fora de control del T^2_C.
- Proposa un algorisme que permet interpretar els senyals fora de control del gràfic T^2_C.

El primer article que conforma aquesta tesi es titula **Individual T^2 Control Chart for Compositional Data** i ha estat publicat al Journal of Quality Technology (Vives-Mestres et al., 2014a). Una transcripció de l'article es troba a la pàgina 42. En aquest article es fa evident la línia de recerca per explorar pel que fa al control de processos de CoDa. Consta d'una llarga i planera secció introductòria a la metodologia composicional, ja que no és habitual en aquest camp de recerca. L'article presenta el gràfic T^2_C per al control de CoDa (observacions individuals) basat en la transformació logràtio de les dades i inclou una comparació gràfica i numèrica entre el gràfic T^2 clàssic i el CoDa T^2_C. Finalitza amb una aplicació pràctica tot utilitzant unes dades conegudes de Holmes and Mergen (1993) que també han estat utilitzades per Sullivan and Woodall (1996) i Montgomery (2009).

El segon article es titula **Out-of-Control Signals in Three-Part Compositional T^2 Control Chart** i ha estat publicat al número especial de l'European Network for Business and Indistrial Statistics 2011 (ENBIS) al Quality and Reliabiity Engineering International (Vives-Mestres et al., 2014b). En aquest article es demostra gràficament que el mètode clàssic per interpretar els senyals fora de control no es pot aplicar a les coordenades amb les que es construeix el T^2_C, i mostra una aproximació geomètrica que facilita la interpretació. L'article finalitza amb una aplicació pràctica utilitzant les dades de l'article precedent.

El darrer article titulat **Signal Interpretation in Hotelling's T^2 Control Chart for Compositional Data** ha estat enviat a la revista

IIE Transactions. En aquest article es proposa un nou algorisme que permet identificar la ràtio de components que és principalment responsable del senyal fora de control del gràfic T^2_C. De fet, es proposen dos algorismes: un és més adequat per processos en els que el nombre de components és inferior a 11 mentre que l'altre és més ràpid per a vectors composicionals de més de 11 parts. L'aplicació del mètode proposat es demostra utilitzant les dades de Gonzalez-de la Parra and Rodriguez-Loaiza (2003) que reporten el perfil d'impureses d'un medicament del qual es vol controlar que es mantenen dins els límits establerts.

1.4 Estructura de la tesi

La present tesi s'estructura de la següent manera. Un cop situada la tesi dins la recerca actual i presentats els articles, es descriuen els objectius generals i específics al Capítol 2. El Capítol 3 presenta breument i sintètica els aspectes metodològics bàsics del control estadístic de processos i de les dades composicionals. En tots dos casos es fa un repàs de la literatura més destacada a la que el lector pot recórrer en cas de desitjar informació més detallada. El Capítol 4 conforma el nucli central de la tesi, i conté una còpia dels dos articles publicats en el mateix format que a la revista i una transcripció de l'article enviat. La tesi continua (Capítol 5) amb una síntesi dels principals resultats i una discussió d'aquests. I finalment es presenten les conclusions així com les futures línies d'investigació.

Capítol 2

Objectius

La present tesi té com a objectiu general l'adaptació de les tècniques de control estadístic de processos, concretament dels gràfics de control, a les dades composicionals. Altrament dit, aplicar les tècniques composicionals per a controlar processos en els que la característica de qualitat (la variable a controlar) és una composició.

S'han treballat tres tipus de gràfics de control: el gràfic de control p, el gràfic multivariant T^2 de Hotelling per al control d'observacions individuals i el gràfic univariant de sumes acumulatives (CUSUM de l'anglès *Cumulative Sums*).

De l'estudi del gràfic p se'n desprenen dues aportacions a congressos i el CUSUM està encara en fase de treball. L'estudi del gràfic de control T^2 de Hotelling conforma el nucli central d'aquesta tesi amb dos articles acceptats i un tercer d'enviat, a banda d'altres aportacions a congressos. És per aquest motiu que la tesi se centra en el treball de la T^2 mentre que el referent als altres dos gràfics de control es descriuen sintèticament al Capítol 5.

2.1 Objectius del nucli central de la tesi

Pel que fa al gràfic de control multivariant T^2 de Hotelling, podem definir els objectius específics de la forma següent.

O1. Comparar, gràficament, les regions de control que s'obtenen del gràfic de Hotelling per al control de CoDa tot utilitzant la metodologia estàndard (T^2) i la metodologia composicional (T^2_C). Fer la comparació en termes de la capacitat per seguir la distribució de les dades composicionals i la restricció de la regió de control al símplex.

O2. Comparar, mitjançant indicadors de funcionament, el rendiment de la metodologia estàndard i el de la metodologia composicional a l'hora de controlar CoDa. Concretament, utilitzar l'indicador RL (de l'anglès *run length*) i els estadístics de la seva distribució (mitjana i quartils).

El RL es defineix com el nombre de mostres que es prenen fins que el gràfic detecta una observació fora de control.

O3. Proposar una solució geomètrica a l'espai $\mathbb{R}^2$, on es representen les coordenades de les composicions de tres parts, que permeti identificar els components causants de les observacions fora de control del gràfic T^2_C.

O4. Aportar una interpretació gràfica al diagrama ternari dels elements de la descomposició de l'estadístic T^2_C mitjançant el mètode de descomposició ortogonal MYT. Comparar aquesta interpretació amb l'enfoc clàssic.

O5. Desenvolupar un algoritme automàtic que, donada una observació fora de control en el gràfic T^2_C, retorni el quocient de components que més contribueixen a la causa de la anomalia.

O6. Mostrar els conceptes i metodologia proposats amb aplicacions pràctiques de casos reals.

La relació entre els objectius i els tres articles es resumeix a la Taula 2.1.

	Objectiu					
Article	O1	O2	O3	O4	O5	O6
Individual T^2 Control Chart for Compositional Data	✓	✓				✓
Out-of-Control Signals in Three-Part Compositional T^2 Control Chart			✓	✓		✓
Signal Interpretation in Hotelling's T^2 Control Chart for Compositional Data					✓	✓

Taula 2.1: Relació entre objectius del nucli central de la tesi i els articles publicats i enviat.

2.2 Altres objectius

Pel que fa al gràfic per al control de proporcions p:

- Proposar un enfoc adequat per controlar proporcions basat en la metodologia composicional i comparar aquest amb els mètodes clàssics.
- Estudiar si l'enfoc composicional aporta ua solució òptima en el control de processos en el quals les proporcions a controlar limitin amb la

frontera, és a dir, siguin properes a 0 o 1. En aquests casos la versió estàndard presenta problemes ja que el límit inferior sovint dona valors negatius, cosa que no és possible en el control de proporcions.

Pel que fa al gràfic per al control de sumes acumulatives CUSUM per al control de proporcions:

- Proposar un nou gràfic de sumes acumulatives CUSUM per controlar proporcions basat en la metodologia composicional i comparar aquest amb els mètodes clàssics.

Capítol 3

Metodologia

En aquest capítol es fa una revisió dels aspectes essencials a tenir en compte referents a les dues temàtiques que tracta la tesi; són d'una banda el control estadístic de processos (SPC de l'anglès *Statistical Process Control*), i més concretament del gràfic de control T^2 de Hotelling, i d'altra banda les dades composicionals (CoDa de l'anglès *Compositional Data*).

3.1 Control Estadístic de Processos (SPC)

La idea clau darrere del SPC és diagnosticar i reduir les causes de la variabilitat per produir el major nombre possible d'unitats no defectuoses. Cal distingir entre els dos tipus de causes que generen variació: les causes comunes i les causes especials o assignables.

El tipus de causa que produeixi la variabilitat marcarà l'estratègia a seguir per eliminar-la. Les causes assignables s'eliminen mitjançant plans correctius i preventius mentre que si l'objectiu és reduir les causes comunes de variabilitat (i.e. la variabilitat total), és necessari introduir canvis importants en el procés, com ara noves màquines o serveis.

La principal eina que permet distingir entre les dues causes de variació són els gràfics de control (CC de l'anglès *Control Chart*) desenvolupats per Dr. Andrew Shewhart en el seu treball Shewhart (1931), també coneguts com a gràfics de Shewhart. Shewhart va crear una eia simple que podia ser utilitzada per personal no estadístic i que permetia aplicar conceptes estadístics per representar la variabilitat pròpia d'un procés de fabricació.

D'altres eines utilitzades en SPC són el disseny d'experiments, l'anàlisi de capacitat, els diagrames causa-efecte o els diagrames d'Ishikawa. Aquestes eines bàsiques, juntament amb els mètodes de SPC, s'enfronten a nous reptes d'adaptació a un entorn de fabricació canviant que inclou noves tendències cap cicles de producció més curts, major quantitat de dades, requisits de més qualitat, i a una major capacitat computacional (Woodall and Montgomery, 1999).

3.1.1 Gràfics de control

Els gràfics de control permeten assegurar que el procés es manté en un estat de control estadístic. En aquest estat, el procés funciona correctament a un nivell estable.

En un gràfic de control es representen els valors d'un estadístic, com ara la mitjana, la desviació, o la proporció, calculat a partir d'una mostra de mida n presa a intervals regulars, en funció del temps o número de mostra. Cada mostra es considera un subgrup racional: els elements de la mostra estan subjectes a variacions aleatòries (les causes assignables apareixen entre subgrups). Quan no és possible prendre mostres majors que 1 el CC es basa en les observacions individuals ($n = 1$).

La Figura 3.1 mostra un esquema d'un gràfic de control. La línia central (CCL de l'anglès *Center Control Line*) representa una mesura de centre mentre que les línies superior i inferior (UCL i LCL, de l'anglès *Upper Control Line* i *Lower Control Line* respectivament) senyalen una amplada dins la qual és probable que es mogui l'estadístic de control si el procés està en estat de control estadístic.

Quan es disposa d'una nova mostra, es calcula l'estadístic i s'afegeix un nou punt al CC. Si el nou punt se situa entre les UCL i LCL, indica que el procés està sota control, mentre que si està fora d'aquests límits és probable que alguna causa hagi afectat el procés (causa assignable). Patrons no aleatoris dins els límits de control també poden indicar la presència de causes assignables, per exemple, 9 punts seguits per sobre o per sota la CCL, 6 punts seguits creixent o decreixent contínuament, ...

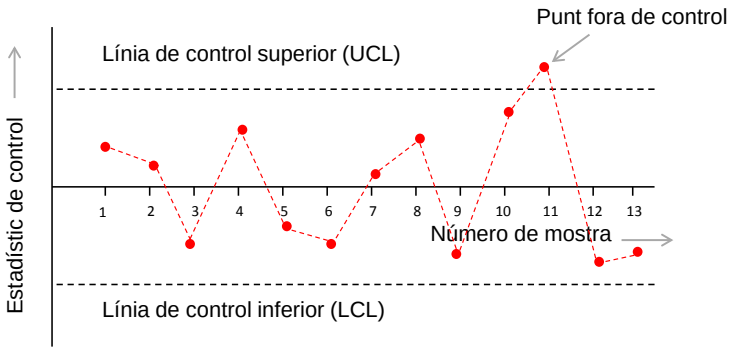


Figura 3.1: Esquema d'un gràfic de control.

En el gràfics Shewhart els límits UCL i LCL se situen a una distància tres cops la desviació estàndard de l'estadístic (3σ) des de la CCL. El valor 3 va ser triat per Shewhart (Shewhart, 1931) perquè, sota el supòsit de normalitat, "*seems to be an acceptable economic value*", és a dir, ofereix un balanç raonable entre la detecció de falses alarmes i la detecció ràpida

de canvis en el procés. L'ús de $k\sigma$ límits és d'especial interès en camps d'aplicació com la diagnosi clínica on la detecció ràpida és més important que les falses alarmes.

Els CC més habituals assumeixen que la distribució de la variable de control és una d'entre Normal, Poisson o Binomial i que les observacions són independents. Els gràfics Shewhart són molt senzills però s'allunyen de l'òptim ja que per exemple són ineficients per a detectar canvis petits en la mitjana del procés.

3.1.2 Fases del procés de control

Hem de distingir entre dues fases en el procés de posada en marxa d'un gràfic de control.

La primera fase anomenada Fase I o fase retrospectiva, consisteix en portar el procés a un estat de control estadístic i després definir què s'entén per estar sota control estadístic, és a dir, definir els paràmetres de la distribució subjacent a partir d'una mostra de mida m que està sota control. Si els paràmetres ja són coneguts (i.e. per la disponibilitat d'informació passada), es poden fer servir amb cautela en comptes d'estimar-los.

A la Fase II o fase prospectiva el CC s'utilitza per controlar el procés i verificar, a l'arribada de cada observació, si l'estat de control ha canviat. En aquesta fase s'utilitzen els límits de control i/o els estimadors obtinguts a la Fase I.

Woodall WH. (2000) destaca la importància de la Fase I i assegura que "*researchers tend to neglect Phase I applications and the vitally important practical considerations of quality characteristic selection, measurement and sampling issues, and rational subgrouping*". A més a mes "*authors sometimes do not clearly specify whether their work applies to Phase I or Phase II*" (Woodall and Montgomery, 1999). És evident que l'estimació de paràmetres de la Fase I té un impacte directe en el funcionament del CC. De fet, els límits de control de la Fase II són variables aleatòries, i per determinar de forma precisa els límits es necessiten més dades de les que tradicionalment es recomanen. L'efecte de l'estimació de paràmetres en el rendiment del CC s'ha estudiat per a pocs tipus de CC (Jensen et al., 2006).

3.1.3 Indicadors de funcionament

El funcionament d'un gràfic de control va estretament lligat al seu disseny. El disseny d'un CC implica triar l'estadístic de control, els límits de control, la mida de l'increment/disminució a detectar (δ), la mida de la mostra (n) així com la freqüència de mostreig entre d'altres. De vegades, alguns d'aquests paràmetres es trien per aconseguir un nivell específic de rendiment.

Existeixen diferents indicadors per mesurar el rendiment dels CC i sovint s'utilitzen per comparar diferents esquemes. L'indicador més popular és

el valor esperat del RL (de l'anglès *run length*). El RL d'un CC és el nombre de mostres que es prenen fins que el gràfic detecta una observació fora de control. La variable aleatòria discreta que defineix el RL es denota típicament per N i la seva distribució s'anomena *run length distribution*.

Assumint que les mostres aleatòries són independents i que la probabilitat de donar un senyal és la mateixa per a totes les mostres (independentment de la distribució assumida sobre les dades), quan els paràmetres són coneguts la distribució RL és geomètrica

$$P(N = j) = \beta(k, \delta, n)^{j-1}(1 - \beta(k, \delta, n)), j = 1, 2, ...$$

on $\beta(k, \delta, n)$ és l'error tipus II o la probabilitat de que el gràfic no doni senyal estant el procés fora de control, és a dir, $\beta(k, \delta, n) = P(\text{No senyal} \mid \text{Canvi en el procés})$. Quan els paràmetres són estimats, la distribució RL no és geomètrica.

L'esperança de la distribució N, anomenada ARL (de l'anglès *Average Run Length*) es calcula com

$$ARL = E[N] = [1 - \beta(k, \delta, n)]^{-1} \tag{3.1}$$

Si no hi ha canvis en el procés, la distribució de N s'anomena distribució RL sota control i s'expressa com N_0 i la seva mitjana ARL_0 que representa el nombre esperat de subgrups obtinguts fins que el gràfic detecta erròniament una mostra fora de control. Paral·lelament, si hi ha un canvi en el procés, la distribució de N s'anomena RL fora de control i s'expressa com N_δ i la seva mitjana ARL_δ, és el nombre esperat de subgrups obtinguts fins que el CC detecta que el procés està fora de control.

Un CC eficient té un ARL_0 gran i un ARL_δ petit. Quan es compara el rendiment de dos o més CC, l'ARL_0 se sol fixar a un nivell acceptablement alt, i aquell que tingui un menor ARL_δ per un determinat desplaçament δ és el guanyador. El valor sovint acceptat de l'ARL sota control és $ARL_0 = 370.4$ que prové de l'assumpció de normalitat de les dades amb mitjana i variància conegudes i $k = 3$. En aquestes condicions la distribució N és geomètrica amb probabilitat d'èxit 0.0027.

En els darrers anys l'ús de l'ARL com a mesura del rendiment dels CC ha estat subjecte a crítiques, perquè d'una banda la desviació estàndard de la RL és molt gran i de l'altra la distribució geomètrica és molt esbiaixada i per tant la mitjana de la RL no és una mesura correcta de centre (valor típic). En contrapartida, a la literatura es proposa no només fixar-se en els valors de l'ARL sinó també en els quartils de la distribució RL (Kenett and Pollak, 2011).

3.1.4 Tipus de gràfics de control

Segons sigui la natura de les dades, distingirem entre els CC destinats al control de dades contínues i els CC per a dades discretes o atributs. Pel que

fa a les dades contínues els més utilitzats són els gràfics de Shewhart per a la mitjana (x barra o $\bar{x}$) i pel control d'observacions individuals. Aquests gràfics sovint van acompanyats per CC per al control de la variabilitat, essent els més bàsics els gràfics R i s. Pel que fa a les dades discretes el més comú és el gràfic p per al control de proporcions o bé el gràfic np per a unitats defectuoses. També hi trobem els gràfics c i u, que permeten controlar el nombre total de disconformitats per unitat i la proporció de disconformitats en una mostra respectivament.

S'han proposat a la literatura millores als gràfics Shewhart per detectar canvis petits en la mitjana del procés i per tractar els casos en els que la distribució de les dades s'allunya lleugerament de la normalitat. És el cas dels gràfics CUSUM (de l'anglès *Cumulative Sum*) i EWMA (de l'anglès *Exponentially Weighted Moving Average*). Ambdós fan ús d'observacions presents i passades.

Altres esquemes de control avançats inclouen adaptacions per tenir en compte la correlació entre variables, per cicles curts de producció, per nombre de mostres canviants, ... Per més informació sobre aquests i els gràfics de control de Shewhart es pot consultar Kenett and Zacks (1998); Montgomery (2009); Kenett et al. (2014).

La majoria de gràfics de control assumeixen que les dades segueixen algun tipus de distribució paramètrica. Quan la distribució no és coneguda i s'allunya de la normalitat s'utilitzen CC no paramètrics (Chakraborti et al., 2001).

Els gràfics presentats fins al moment serveixen per controlar una variable en una dimensió. No obstant, les dades multivariants són molt més informatives que una col·lecció de dades univariants. Per això s'han desenvolupat els gràfics de control multivariants, que permeten controlar simultàniament un conjunt de variables. A diferència dels CC univariants, els CC multivariants requereixen una mesura per detectar desviacions respecte als valors objectiu de cada variable i al mateix temps una mesura per avaluar l'estructura de covariància de les dades.

L'ús de múltiples gràfics univariants per controlar un conjunt de variables està desaconsellat perquè l'error de tipus I i la probabilitat que un punt assenyali correctament l'estat de control del conjunt no són iguals als valors dels gràfics individuals. A més a més, en els gràfics univariants no es té en compte la relació entre les variables.

En un esquema multivariant s'assumeix que $p \times 1$ vectors aleatoris $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ...$ s'observen i controlen al llarg del temps. Cada vector conté p característiques de qualitat. En la majoria de casos, les observacions es consideren independents i seguint una distribució normal multivariant amb vector mitjana $\boldsymbol{\mu}$ i matriu de covariàncies $\boldsymbol{\Sigma}$.

El primer control de qualitat multivariant va ser proposat per Hotelling (1947), que va desenvolupar el més conegut dels CC multivariants: el gràfic T^2 de Hotelling per al control de la mitjana d'un procés. És l'extensió

multivariant del gràfic de Shewhart $\bar{x}$.

La majoria d'esquemes de control assumeixen dades independents i idènticament distribuïdes. Per a esquemes de control per processos més complexos com ara processos canviants en el temps es pot consultar el llibre de Xie and Kruger (2012).

Descrivim a continuació amb detall els gràfics de control per a proporcions p i el gràfic de control multivariant T^2 de Hotelling.

Gràfic de control per a proporcions

En el gràfic de control p, es controla la fracció d'ítems que posseeixen una determinada característica del total d'unitats estudiades (sovint es parla d'unitats defectuoses). Cada punt del gràfic p es calcula com $p_i = x_i/n_i$ on x_i és el nombre d'unitat que posseeixen la característica en l'i-èssim subgrup de mida n_i. El gràfic es basa en la distribució binomial ($Bin(n, p)$) de les observacions, assumint que la classificació dels individus en les dues categories és independent l'una de l'altra. Els límits de control es calculen com

$$\begin{aligned} UCL &= \bar{p} + k\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \\ CCL &= \bar{p} \\ LCL &= \bar{p} - k\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \end{aligned} \tag{3.2}$$

on $\bar{p}$ és l'estimador de p, calculat com la mitjana de la proporció d'unitats defectuoses de la mostra inicial de la Fase I, i n és la mida del subgrup. Si p és conegut, el valor de $\bar{p}$ de l'Equació 3.2 es substitueix per p.

En els casos en els que la mida de la mostra n varia, els UCL i LCL es recalculen per cada mostra mentre que la CCL es manté constant. Un altre enfoc consisteix en calcular els límits amb una mitjana de la mida de la mostra $\bar{n}$ assumint que n_i varia poc d'una mostra a l'altra. Un tercer enfoc consisteix en estandarditzar la variable p_i de forma que els punts del gràfic vinguin mesurats en unitats de desviació estàndard.

Un dels principals problemes del gràfic p és que no sempre té límit inferior (LCL). Quan el valor de $\bar{p}$ és petit (e.g. $\bar{p} = 0.1$) i donat n constant (e.g. $n = 10$) el límit inferior de la Equació 3.2 pren valors negatius (e.g. $LCL = -0.18$), cosa que no és raonable tenint en compte la natura de les dades. En aquests casos el gràfic no pot detectar disminucions en el valor de p, és a dir, no pot detectar que el procés millora.

Els límits de l'Equació 3.2 es basen en l'aproximació de la distribució binomial per la normal. Aquesta aproximació és correcta per valors alts de n i valors de p que no s'allunyen excessivament de 0.5 ja que en aquests casos la distribució binomial és menys esbiaixada. Altrament el biaix de la

distribució binomial fa que l'aproximació per la normal no sigui adequada. Una norma general per poder fer aquesta aproximació ve donada per $np(1-p) > 9$ o $np > 5$ per $0 < p \leq 0.5$ i $n(1-p) > 5$ per $0.5 < p < 1$ (Schader and Schmid (1989)).

Quan el valor de p és conegut, $k = 3$ i es satisfan les condicions de normalitat, l'error tipus I del gràfic no s'allunya gaire del valor nominal de 0.0027. No obstant, quan p és desconegut, el valor real de l'error tipus I no només depèn de n sinó també del valor real de p (Jensen et al., 2006). Una bona discussió sobre els indicadors de funcionament dels gràfics p, tant pel cas de paràmetres estimats com coneguts, es pot trobar a Chakraborti and Human (2006).

Woodall (1997) va assenyalar que les distribucions dels estadístics utilitzats en els gràfics de control per atributs són majoritàriament esbiaixades per la dreta, cosa que implica que la probabilitat de detectar un increment en p sigui en general diferent de la de detectar una disminució en p fins i tot si el procés està sota control.

A la literatura es proposen diverses alternatives per corregir el biaix de la distribució binomial. Alguns enfocs aposten per modificar/ajustar els límits de control (Acosta-Mejia, 1999). D'altres opten per transformar les dades, per exemple, amb la funció logit o la inversa de la funció sinus (Newcombe, 2001).

Per detectar millores en el procés, també és possible utilitzar límits de control calculats de forma que la probabilitat d'obtenir una falsa alarma en els dos extrems del gràfic sigui igual a $\alpha/2$. Aquest s'anomena *mètode exacte* i els intervals que se'n deriven es coneixen com interval Clopper-Pearson (Blyth and Still, 1983).

D'altres solucions impliquen canviar l'estadístic de control pel nombre d'unitats observades fins que n'apareix una de no conforme, fet que es modela amb una distribució geomètrica. Aquests esquemes són adequats per processos amb percentatges de defectes molt baixos, coneguts en anglès com *high yield processes*. Aquest gràfic és apropiat quan els ítems es produeixen de forma seqüencial i s'inspecciona el 100% de la producció.

D'altres gràfics es basen en transformacions de la distribució geomètrica per tal que sigui aproximadament normal, per després aplicar els esquemes clàssics. Algunes transformacions són la potència del tipus $y^{1/3.6}$ on y és un comptatge geomètric (Nelson, 1994) o l'ús de l'estadístic Q (Quesenberry, 1991).

Alguns articles on es comparen els diferents mètodes són Newcombe (1998) Pires (1998) McCool and Joyner-Motley (1998) i pel cas concret de p petita a Chang and Gan (2007) i Wang (2009).

Gràfic T^2 de Hotelling

L'extensió multivariant del gràfic $\bar{x}$ és el de T^2 Hotelling, que utilitza com a mesura del desplaçament la distància de Mahalanobis, que és l'arrel quadrada del paràmetre

$$\lambda^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{3.3}$$

Distingirem entre els CC per al control d'observacions individuals i per al control de mitjanes. Dins de cada tipus, separarem entre la fase I (estimació de paràmetres) i la fase II (control del procés).

Gràfic T^2 de Hotelling per a observacions individuals En els casos en els que no és possible disposar de mides de mostra superiors a u ($n = 1$), s'utilitza el gràfic de control de Hotelling per a observacions individuals. A cada instant de temps s'observa el vector $\mathbf{x} = (x_1, x_2, ...x_p)$ que conté p característiques de qualitat.

En el cas més senzill en el que els paràmetres $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ de la distribució són coneguts, es representa la distància de Mahalanobis des de cada punt al valor objectiu $\boldsymbol{\mu}$, és a dir,

$$T^2 = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

En aquest cas, l'estadístic T^2 segueix una distribució χ^2 amb p graus de llibertat.

En la major part d'aplicacions, els paràmetres de la distribució no són coneguts i s'han d'estimar a partir d'una mostra de mida m a la fase I. L'estimació de la mitjana es calcula

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, ...\bar{x}_p)' \quad \text{on} \quad \bar{x}_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij}$$

i la matriu de covariàncies

$$\mathbf{S} = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})'$$

L'estadístic que s'utilitza a la fase I del CC de Hotelling per a observacions individuals amb paràmetres estimats és

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})' S^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \tag{3.4}$$

Aquest no segueix una distribució χ^2 ni F, sinó que segueix una distribució beta degut a la dependència entre les $\mathbf{x}$ i l'estimació de la mitjana i la matriu de covariàncies (Tracy et al., 1992). En concret

$$T^2 \sim \frac{(m-1)^2}{m} B(p/2, (m-p-1)/2 \tag{3.5}$$

Tracy et al. (1992) recomana en aquesta fase l'ús de límits de control superiors i inferiors ja que l'estadístic de la Equació 3.4 és sensible als canvis en la mitjana així com en la matriu de covariàncies; valors petits de l'estadístic poden indicar canvis en la matriu de covariàncies. Malgrat aquesta recomanació, en les aplicacions pràctiques no s'utilitza el límit inferior.

A la fase II les observacions són independents dels paràmetres estimats $\bar{\mathbf{x}}$ i $\boldsymbol{S}$ i per tant al CC es dibuixa l'estadístic

$$T_f^2 = (\mathbf{x}_f - \bar{\mathbf{x}})' \boldsymbol{S}^{-1} (\mathbf{x}_f - \bar{\mathbf{x}})$$

on el subíndex f fa referència a observacions futures, i que segueix una distribució

$$T_f^2 \sim \frac{p(m+1)(m-1)}{m(m-p)} F(p, m-p)$$

En cada cas, el límit de control (UCL), es calcula a partir del percentil escollit de la corresponent distribució: ja sigui la Khi-quadrat, la Beta o la F de Fisher.

Gràfic T^2 de Hotelling per al control de mitjanes Quan la mida de la mostra és superior a u ($n > 1$), s'assumeix que a la fase I es disposa de m mostres, cadascuna de les quals conté n vectors amb p característiques de qualitat. De cada mostra, també anomenat subgrup, s'utilitza el $(p \times 1)$ vector de mitjanes mostral ($\bar{\mathbf{x}}$) i la matriu de covariàncies mostral $(p \times p)$ per estimar $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$.

$$\bar{\bar{\mathbf{x}}} = \frac{1}{m} \sum_{i=1}^{m} \bar{\mathbf{x}} \quad \text{i} \quad \bar{\boldsymbol{S}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{S}$$

A la fase I, Ryan (2011) proposa utilitzar l'estadístic.

$$T^2 = n(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})' \bar{\boldsymbol{S}}^{-1} (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})$$

que segueix una distribució

$$T^2 \sim \frac{p(m-1)(n-1)}{mn-m-p+1} F_{p,mn-m-p+1}$$

En canvi, a la fase II s'utilitza l'estadístic

$$T_f^2 = n(\bar{\mathbf{x}}_f - \bar{\bar{\mathbf{x}}})' \bar{\boldsymbol{S}}^{-1} (\bar{\mathbf{x}}_f - \bar{\bar{\mathbf{x}}})$$
$$T_f^2 \sim \frac{p(m+1)(n-1)}{mn-m-p+1} F_{p,mn-m-p+1}$$

Altra vegada el límit de control (UCL) es calcula, en cada cas, a partir del percentil escollit de la corresponent distribució F.

La Taula 3.1 resumeix l'estadístic a utilitzar en cadascuna de les fases, en funció de la mida de la mostra i el coneixement sobre els paràmetres de la distribució.

Taula 3.1: Resum dels estadístics del gràfic T^2 de Hotelling i les seves distribucions en funció de la mida de la mostra (n), la fase de control i el coneixement o no dels paràmetres de la distribució.

Mostra n	Fase	Paràmetres μ i Σ	Estadístic i distribució
$n = 1$	Fase I o II	Conegut	$T^2 = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ $T^2 \sim \chi^2_p$
$n = 1$	Fase II	μ Conegut Σ Desconegut	$T^2_f = (\mathbf{x}_f - \boldsymbol{\mu})' \boldsymbol{S}^{-1}_m (\mathbf{x}_f - \boldsymbol{\mu})$ $T^2_f \sim \frac{p(m-1)}{m(m-p)} F_{(p,m-p)}$
$n = 1$	Fase I	Desconegut	$T^2 = (\mathbf{x} - \bar{\mathbf{x}})' \boldsymbol{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$ $T^2 \sim \frac{(m-1)^2}{m} B_{(p/2,(m-p-1)/2)}$
$n = 1$	Fase II	Desconegut	$T^2_f = (\mathbf{x}_f - \bar{\mathbf{x}})' \boldsymbol{S}^{-1} (\mathbf{x}_f - \bar{\mathbf{x}})$ $T^2_f \sim \frac{p(m+1)(m-1)}{m(m-p)} F_{(p,m-p)}$
$n > 1$	Fase I o II	Conegut	$T^2 = (\bar{X}_i - \mu_0)' \Sigma^{-1} (\bar{X}_i - \mu_0)$ $T^2 \sim \chi^2_p$
$n > 1$	Fase I	Desconegut	$T^2 = n(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})' \bar{\boldsymbol{S}}^{-1} (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})$ $T^2 \sim \frac{p(m-1)(n-1)}{mn-m-p+1} F_{(p,mn-m-p+1)}$
$n > 1$	Fase II	Desconegut	$T^2_f = n(\bar{\mathbf{x}}_f - \bar{\bar{\mathbf{x}}})' \bar{\boldsymbol{S}}^{-1} (\bar{\mathbf{x}}_f - \bar{\bar{\mathbf{x}}})$ $T^2_f \sim \frac{p(m+1)(n-1)}{mn-m-p+1} F_{(p,mn-m-p+1)}$

Tant per als gràfics per al control d'observacions individuals com per al control de mitjanes, la mida de la mostra m que permet obtenir un bons estimadors dels paràmetres de la distribució ha estat objecte d'estudi i de debat. Lowry and Montgomery (1995) fa recomanacions de m per als dos tipus de gràfics $n = 1$ i $n > 1$. Utilitza el criteri de l'error relatiu comparant els límits exactes de la Fase II i els límits aproximats amb la χ^2. Posteriorment s'ha vist que les seves recomanacions són massa baixes, ja que no tenen en compte la RL, i per tant les mides de mostra que proposen resulten

en un nombre alt de falses alarmes i valors baixos d'ARL.

Mason et al. (2003) proporciona una bona discussió per determinar la mida de la mostra en gràfics per observacions individuals ($n = 1$). Proposen tres mesures per determinar la mida per a la Fase I i Fase II amb límits basats en la distribució Beta i F respectivament.

Champ et al. (2005) estudia els gràfics de mitjanes ($n > 1$) i dóna recomanacions sobre la mida de la mostra m de forma que s'obtinguin valors d'ARL a la Fase II a una distància raonable de $1/\alpha$ utilitzant límits basats en la distribució F.

En un article més recent, Chen and Pan (2011), recomana valors de m per controlar observacions individuals ($n = 1$) quan l'estadístic T^2 es calcula utilitzant la matriu de covariàncies estimada a partir de les diferències successives de les observacions (S_D).

El gràfic de control T^2 de Hotelling permet condensar la informació de múltiples variables en un sol estadístic. Com que la T^2 té en compte la mitjana i la matriu de covariàncies del procés, els punts fora de control poden indicar canvis en la mitjana, en la variància o fins i tot en la relació entre les variables. Aquest inconvenient, juntament amb el fet que la T^2 no té sentit pràctic, fan difícil la interpretació els punts fora de control.

Molts articles tracten el tema de la interpretació dels signes fora de control i presenten diferents alternatives per diagnosticar i identificar les variables que causen la fallada. El més antic dels procediments va ser proposat per Jackson (1985), que va utilitzar els components principals de l'estimació de la matriu de covariàncies per descompondre el valor de la T^2. Aquest mètode es pot utilitzar per identificar el grup de variables que contribueix al senyal, però no permet una interpretació directe en termes de les variables originals.

Un mètode més eficient és el proposat per Mason et al. (1995), conegut com el mètode de descomposició MYT. El mètode consisteix en descompondre l'estadístic T^2 en termes independents, anomenats termes no condicionals i termes condicionals. La descomposició permet a l'usuari identificar quines són les variables que tenen una contribució significant a la causa de la desviació.

Das and Prakash (2007) compara els dos mètodes, entre d'altres. Els autors avaluen el funcionament en termes de la capacitat dels diferents mètodes per detectar les variables que causen la desviació, assumint que la matriu de covariàncies es manté constant, ja que els mètodes comparats no permeten identificar si la causa del problema prové d'un canvi en el vector de mitjanes o en la matriu de covariàncies.

D'altres mètodes per a la interpretació dels punts fora de control són el proposat per Hawkins (1993) que utilitzava la variables de regressió ajustada per millorar la diagnosi. Per més detalls sobre els mètodes existents veure Mason and Young (2006).

3.2 Dades Composicionals

En aquesta secció es resumeix els aspectes bàsics que defineixen les dades composicionals (CoDa de l'anglès *Compositional Data*). El contingut es basa en Pawlowsky-Glahn et al. (2010) i les tesis doctorals de Mateu-Figueras (2003) i Martín-Fernández (2001), a les qual el lector es pot remetre per més informació. També és d'especial interès el llibre Pawlowsky-Glahn and Buccianti (2011) i les referències que es detallen al final de cada capítol.

Les dades composicionals descriuen quantitativament les parts d'un total, i per tant habitualment les unitats són parts per u, percentatges, ppm, ppb o concentracions. La restricció de suma constant que acostuma a caracteritzar les CoDa complica l'anàlisi estadística. Per exemple, el fet de modificar una component qualsevol d'un vector composicional provoca necessàriament canvis en com a mínim una de les altres components. Aquest fet obliga a replantejar-se el concepte d'independència estocàstica quan es treballa amb vectors aleatoris composicionals.

Igualment, el coeficient de correlació entre dues components no es pot interpretar de forma habitual. Pearson (1897) va ser el primer a assenyalar que les components que tenen un mateix denominador introdueixen una falsa o espúria correlació entre elles. Aquestes són algunes de les dificultats que impedeixen aplicar l'anàlisi estadística estàndard quan es treballa amb dades composicionals.

Aitchison (1986) va ser el primer a desenvolupar una metodologia específica basada en la idea de que les CoDa representen parts d'un total i que per tant només contenen informació relativa, és a dir, l'única informació rellevant en una composició està continguda en les ràtios de les seves components, i no en els seus valors absoluts.

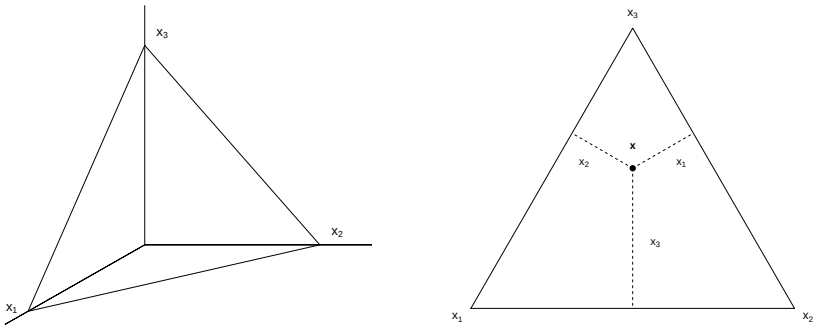
Tota la metodologia composicional està basada en les ràtios, o més ben dit, en les logràtios ja que el logaritme facilita el maneig i simetritza les dades.

3.2.1 Conceptes bàsics

La introducció a la metodologia de les dades composicionals es pot fer a partir de les classes d'equivalència presentades amb molt de detall a Barceló-Vidal et al. (2001). No obstant, per simplificar, hem optat per la introducció de les dades composicionals a partir dels representants lineals de suma constant definits a partir de l'operació clausura, és a dir, considerant que la suma de les parts de la composició és constant. Recordem però que una composició es defineix com un vector de components que representen parts d'un total, i que la suma no té perquè ser sempre constant.

Ens referirem als elements d'una composició com a part o component.

Definició 3.1 Una composició amb D-parts és un vector $(D \times 1)$ els components del qual $x_1, x_2, \ldots, x_D$ són nombres reals estrictament positius $x_1 >$



(a) Símplex a $\mathbb{R}^3$. (b) Diagrama ternari.

Figura 3.2: Representacions equivalents de composicions de tres parts a (a) $\mathbb{R}^3$ i (b) al diagrama ternari.

$0, x_2 > 0, \ldots, x_D > 0$, que sumen un valor constant $x_1 + x_2 + \ldots + x_D = \kappa$ i que contenen informació relativa.

Usualment, $\kappa = 1$ o $\kappa = 100$, és a dir, que les mesures s'han fet, o transformat, a proporcions o percentatges, respectivament. També són composicions les dades mesurades en unitats de concentracions, com ara mg/l o molaritats.

Definició 3.2 L'espai mostral natural de les composicions és el símplex $\mathcal{S}^D$, definit com

$$\mathcal{S}^D = \{(x_1, x_2, \ldots, x_D) | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa\}.$$

Les composicions de 3 parts ($D = 3$) es troben inscrites en un triangle equilàter a $\mathbb{R}^3$, situat al pla perpendicular al vector $(1, 1, 1)$ (Figura 3.2a). No obstant, és més habitual representar les dades al diagrama ternari (Figura 3.2b), que és una representació equivalent. Un diagrama ternari és un triangle equilàter tal que la mostra genèrica $\mathbf{x} = (x_1, x_2, x_3)$ es troba a una distància x_1 del costat oposat al vèrtex X_1, a una distància x_2 del costat oposat al vèrtex X_2 i a una distància x_3 del costat oposat al vèrtex X_3. En el cas de $D = 4$ el símplex es representa en un tetraedre regular d'alçada unitat.

Per canviar les unitats de la composició, e.g. passar-les a percentatge o a proporció, és necessària una operació anomenada clausura.

Definició 3.3 Per qualsevol vector de D components real positius $\mathbf{x} = (x_1, x_2, \ldots, x_D) \in \mathbb{R}_+^D$ la clausura de $\mathbf{x}$ es defineix com

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right)'$$

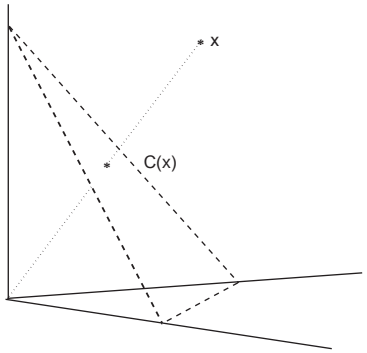


Figura 3.3: Representació gràfica de l'operació clausura.

El resultat de la clausura és el mateix vector reescalat de forma que la suma dels components sigui κ. La interpretació gràfica de l'operació clausura es mostra a la Figura 3.3: la clausura de $\mathbf{x}$ mou el punt al llarg de la recta que va des de l'origen fins a $\mathbf{x}$ fins a la intersecció amb el pla $\sum x_i = \kappa$.

Notem que el canvi en una de les parts d'una composició provoca necessàriament el canvi en com a mínim una de les altres parts. Una composició $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ és un vector de dimensió $D-1$ ja que queda completament especificat si es coneixen $D-1$ parts. La composició $\mathbf{x}$ també queda totalment determinada si es coneixen els $D-1$ quocients x_i/x_D per a $i = 1, 2, \ldots, D-1$ (Aitchison, 1986).

El valor de la suma constant κ no és informatiu, ja que només depèn de l'escala utilitzada per expressar les dades composicionals. Per tant, un fet molt important a tenir en compte és que les dades composicionals només aporten informació sobre les magnituds relatives $x_i/x_j (i, j = 1, 2, \ldots, D; i \neq j)$ de les components que l'integren. Qualsevol tècnica estadística que es vulgui aplicar a aquest tipus de dades s'ha de dirigir als quocients o ràtios entre components en comptes de dirigir-se a les magnituds absolutes d'aquests.

3.2.2 El símplex com a espai vectorial

Posteriorment als treballs d'Aitchison, es va demostrar que la metodologia CoDa basada en transformacions logràtio sobre el símplex es podia explicar equivalentment a partir de l'estructuració del símplex com a espai vectorial euclidià (Barceló-Vidal et al., 2001; Egozcue and Pawlowsky-Glahn, 2006).

Sobre el símplex es defineixen dues operacions bàsiques: la pertorbació, definida per a dues composicions $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$, i la potència, definida entre una composició $\mathbf{x} \in \mathcal{S}^D$ i un escalar $\alpha \in \mathbb{R}$.

Definició 3.4 Siguin $\mathbf{x}, \mathbf{x}^*$ dues composicions amb D parts. Llavors l'operació

$$\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}\left(x_1 x_1^*, x_2 x_2^*, \ldots, x_D x_D^*\right)'$$

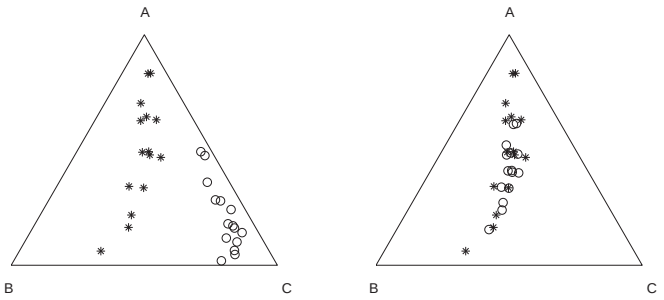


Figura 3.4: A l'esquerra, pertorbació de les composicions inicials $\circ$ per $p = (0.1, 0.1, 0.8)$ que resulten en $*$. A la dreta, potència de les composicions inicials $*$ per $\alpha = 0.2$ resultant en $\circ$.

s'anomena pertorbació.

Definició 3.5 Sigui $\mathbf{x}$ una composició amb D parts i sigui α un escalar de $\mathbb{R}$. Llavors l'operació

$$\alpha \otimes \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)'$$

s'anomena potència.

Les operacions pertorbació i potència, que s'indiquen amb els símbols $\oplus$ i $\otimes$ respectivament, indueixen en el símplex una estructura d'espai vectorial sobre el cos $\mathbb{R}$. La pertorbació actua com a operació interna a $\mathcal{S}^D$ i és l'anàleg a la suma a l'espai real. La potència actua com a operació externa respecte dels elements del cos $\mathbb{R}$ i és l'anàleg a la multiplicació per un escalar a l'espai real. La Figura 3.4 mostra gràficament un exemple del resultat d'aplicar aquestes operacions a un conjunt de composicions a $\mathcal{S}^3$.

Si afegim la definició de producte escalar, norma i distància a les definicions anteriors, és possible proveir al símplex d'una estructura d'espai vectorial euclidià. El producte escalar permet projectar les composicions en determinades direccions i permet verificar l'ortogonalitat entre vectors composicionals. La norma permet mesurar la llargada d'una composició. Totes aquestes operacions permeten operar al símplex de la mateixa manera que es fa a l'espai real (Pawlowsky-Glahn and Egozcue, 2001). Molts problemes composicionals es poden treballar i resoldre mitjançant aquesta estructura algebraico-geomètrica (Mateu-Figueras, 2003; Barceló-Vidal, 2011).

Per a més informació sobre el producte escalar, la norma i la distància definides al símplex, així com les propietats de les operacions de pertorbació i potència, podeu adreçar-vos a Pawlowsky-Glahn et al. (2010).

3.2.3 Subcomposicions

Quan només ens interessen algunes parts (no totes) de la composició $\mathbf{x} \in \mathcal{S}^D$, es diu que treballem amb una subcomposició, definida com

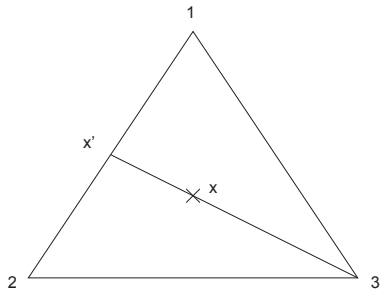


Figura 3.5: Subcomposició $\mathbf{x}' \in \mathcal{S}^2$ representada com a projecció lineal de $\mathbf{x} \in \mathcal{S}^3$.

Definició 3.6 Donada una composició $\mathbf{x}$, una subcomposició $\mathbf{x}_s$ amb s parts, s'obté aplicant l'operació clausura al subvector $(x_{i1}, x_{i2}, \ldots, x_{is})$ de $\mathbf{x}$. Els subíndex $i_1, \ldots, i_s$ indiquen quines parts de la composició se seleccionen, no necessàriament les s primeres.

Destaquem aquí que les ràtios entre els components de la subcomposició són iguals als mateixos ràtios de la composició completa. Aquesta propietat implica que els resultats obtinguts sobre les ràtios entre les parts de la subcomposició són iguals als obtinguts entre les mateixes parts de la composició completa.

Gràficament, una subcomposició és una composició en un símplex de dimensió inferior. Per exemple, a la Figura 3.5 es mostra com la subcomposició $\mathbf{x}' \in \mathcal{S}^2$ formada amb les dues primeres parts de $\mathbf{x} \in \mathcal{S}^3$ és el resultat de la projecció de $\mathbf{x}$ sobre el costat 12 des del vèrtex 3.

És habitual l'ús de subcomposicions quan es treballa amb composicions que contenen moltes parts ja que es fa difícil, si no impossible, mesurar tots els components. Aquest fet es dóna sovint en les mesures de concentracions.

3.2.4 Principis de l'anàlisi de dades composicionals

Qualsevol mètode estadístic aplicat a una composició ha de complir tres principis per ser coherents amb l'estructura de les dades: invariància per escala, invariància per permutació i coherència subcomposicional (Aitchison, 1986).

El principi d'invariància per escala postula que els resultats d'una anàlisi han de ser els mateixos siguin quines siguin les unitats de la composició. L'anàlisi de ràtios compleix aquest principi ja que la ratio $x_1/x_2 = (\lambda x_1)/(\lambda x_2)$ perquè les unitats es cancel·len. No obstant, la ràtio depèn de l'ordre de les parts, és a dir $x_1/x_2 \neq x_2/x_1$. Una transformació adequada utilitza logratios, de la forma $\log x_1/x_2$. D'aquesta manera la inversió dels components produeix un canvi de signe, cosa que dóna una simetria respecte a l'ordre de les parts.

El principi d'invariància per permutació postula que les conclusions d'una anàlisi composicional no han de dependre de l'ordre de les parts.

Finalment, el principi de coherència subcomposicional, estableix que la inferència sobre subcomposicions ha de ser consistent, independentment de si la inferència es basa en la subcomposició o la composició completa. A l'espai real aquest principi es tradueix en que la inferència sobre un subconjunt de variables ha de ser la mateixa independentment de si basem la inferència en un subconjunt de variables o el conjunt complet.

3.2.5 Representació en coordenades

En aquest apartat presentem tres transformacions que ens permeten treballar en coordenades: logràtio additiva (alr de l'anglès *additive log-ratio*), logràtio centrada (clr de l'anglès *centred log-ratio*), logràtio isomètrica (ilr de l'anglès *isometric log-ratio*), totes elles basades en ràtios de components. Denotarem les coordenades fruit de les tres transformacions per $\boldsymbol{w}$, $\boldsymbol{z}$ i $\boldsymbol{y}$ respectivament.

Farem servir la notació log per referir-nos al logaritme natural (en base e) seguint la tendència marcada pels programes estadístics que fan servir per defecte la base natural.

Transformació alr

Aitchison (1986) defineix la transformació logràtio additiva com

Definició 3.7 Donada una composició amb D parts, la transformació logràtio additiva de $\mathbf{x} \in \mathcal{S}^D$ a $\boldsymbol{w} \in \mathcal{R}^{D-1}$ es defineix com

$$\boldsymbol{w} = \text{alr}(\mathbf{x}) = \left(\log \frac{x_1}{x_D}, \log \frac{x_2}{x_D}, \ldots, \log \frac{x_{D-1}}{x_D} \right) \tag{3.6}$$

La transformació alr és bijectiva i la seva inversa és l' alr^{-1} que es defineix com

$$x_i = \frac{\exp w_i}{\sum_{j=1}^{D-1} \exp w_j + 1} \qquad (i = 1, 2, \ldots, D-1),$$
$$x_D = 1 - \left(\sum_{i=1}^{D-1} x_i \right) = \frac{1}{\sum_{j=1}^{D-1} \exp w_j + 1}.$$

La transformació alr és un isomorfisme entre $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$ però no és una isometria, és a dir, no conserva ni les distàncies, ni el producte escalar ni la norma. Això fa que, per exemple, sigui erroni realitzar una classificació en grups d'un conjunt de composicions aplicant la distància ordinària sobre les dades alr-transformades.

Un dels inconvenients de la transformació alr és la seva falta de simetria, ja que la component que figura en el denominador de cada logràtio adquireix

un protagonisme especial respecte de la resta de components. Certament podríem escollir qualsevol altra component com a comú denominador.

La distància euclidiana entre coordenades alr depèn del component del denominador, cosa que fa que aquesta transformació no sigui invariant per permutació.

Transformació clr

Aitchison (1986) defineix la transformació logràtio centrada com:

Definició 3.8 Donada una composició amb D parts, la transformació clr de $\mathbf{x} \in \mathcal{S}^D$ a $\boldsymbol{z} \in \mathbb{R}^{D-1}$ es defineix com

$$\boldsymbol{z} = \text{clr}(\mathbf{x}) = \left(\log \frac{x_1}{g(\mathbf{x})}, \log \frac{x_2}{g(\mathbf{x})}, \ldots, \log \frac{x_D}{g(\mathbf{x})} \right) \tag{3.7}$$

on $g(\mathbf{x}) = (x_1 \cdot x_2 \cdots x_D)^{1/D}$ és la mitjana geomètrica de les D components de $\mathbf{x}$.

En aquest cas, la transformació és simètrica entre les parts. Les dades transformades se situen a l'hiperplà V de $\mathbb{R}^D$ que passa per l'origen i és ortogonal al vector d'unitats $(1, 1, \ldots, 1)$, és a dir, $V = \text{clr}(\mathcal{S}^D) = \{\boldsymbol{z} \in \mathbb{R}^D; \sum_{i=1}^D z_i = 0\}$. Això comporta una nova dificultat, ja que la suma de les components del vector transformat és igual a 0.

La transformació clr és bijectiva entre el símplex i l'hiperplà V; la seva inversa és la clr^{-1} que es defineix com

$$\mathbf{x} = \text{clr}^{-1}(\boldsymbol{z}) = \mathcal{C}(e^{z_1}, e^{z_2}, \ldots, e^{z_D})$$

La distància d'Aitchison entre dues composicions $\mathbf{x}$ i $\mathbf{x}^*$ es defineix com la distància euclidiana entre els respectius vectors clr transformats. Existeix una relació entre les transformacions alr i clr que es pot trobar a Aitchison (1986).

L'inconvenient de la falta de simetria a la definició de la transformació alr no apareix a la definició de la transformació clr. No obstant, l'estructura de covariàncies associada a aquesta transformació no s'allibera de l'inconvenient de la singularitat d'aquesta matriu.

Transformació ilr

Egozcue et al. (2003) defineixen una isometria entre els espais $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$. La motivació principal d'aquesta nova transformació és superar els inconvenients de les dues transformacions anteriors que aporten certes dificultats a l'hora d'interpretar resultats

La transformació isomètrica sorgeix de manera natural si observem la transformació clr. La condició $\sum z_k = 0$ que satisfan les components dels

vectors del subespai $V = \text{clr}(\mathcal{S}^D)$ ens indica que el vector $(1, 1, \ldots, 1)$ és ortogonal a aquest hiperplà. Si escollim una base de l'espai $\mathbb{R}^D$ formada per $D-1$ vectors ortonormals del subespai V i per un vector unitari i normal a V, és a dir, $1/\sqrt{D}(1, 1, \ldots, 1)$, i expressem els vectors clr transformats en aquesta nova base, obtindrem que la seva última component és igual a 0. A continuació, podem eliminar aquesta última component aplicant una projecció sobre l'hiperplà $\text{clr}(\mathcal{S}^D)$.

Aquest procediment, transformació clr seguida d'un canvi de base ortonormal i de la projecció ortogonal sobre el subespai V, dóna lloc a una isometria entre els espais $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$. Egozcue et al. (2003) defineixen aquesta transformació i la denoten per ilr .

Definició 3.9 Donada una base ortonormal del símplex $\mathcal{S}^D$, $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1})$, i la base ortogonal $(D-1 \times D)$ a $\mathbb{R}^{D-1}$ $\mathbf{\Psi} = \begin{pmatrix} \text{clr}(\mathbf{e}_1) \\ \text{clr}(\mathbf{e}_2) \\ \ldots \\ \text{clr}(\mathbf{e}_{D-1}) \end{pmatrix}$, es defineix la transformació ilr d'una composició $\mathbf{x} \in \mathcal{S}^D$ a un vector $\boldsymbol{y} \in \mathbb{R}^{D-1}$ com

$$\boldsymbol{y} = \text{ilr}(\mathbf{x}) = clr(\mathbf{x}) \cdot \mathbf{\Psi}' \tag{3.8}$$

La transformació isomètrica no és única, donat que en la seva definició no queda especificada la base ortonormal de $\mathcal{S}^D$ i per tant tenim la llibertat d'escollir-la. Egozcue et al. (2003) proposen definir-la a partir d'una partició seqüencial binària (SBP de l'anglès *sequential binary partition*).

Una SBP és una jerarquia de les parts duna composició: en un primer pas, la composició es divideix en dos grups; i en el passos seguents, cada grup es divideix al seu torn en dos grups. A cada pas, el nombre de parts $(x_{j_1}, \ldots, x_{j_r})$ en el primer grup, codificades per $+1$, s'enregistra a r i el nombre de parts $(x_{k_1}, \ldots, x_{k_s})$ en el segon grup, codificades per -1, s'enregistra a s. Les coordenades ilr y_i obtingudes al pas i de la SBP i el corresponent element de la base $\boldsymbol{\psi}_i$ es calculen de la següent manera

$$y_i = \sqrt{\frac{r_i \cdot s_i}{r_i + s_i}} \log \frac{(x_{j_1} x_{j_2} \cdots x_{j_r})^{1/r_i}}{(x_{k_1} x_{k_2} \cdots x_{k_s})^{1/s_i}}, \quad \boldsymbol{\psi}_i = (\psi_1, \ldots, \psi_D) \begin{cases} \psi_j = +\sqrt{\frac{s_i}{r_i(r_i+s_i)}} \\ \psi_k = -\sqrt{\frac{r_i}{s_i(r_i+s_i)}} \\ \psi_0 = 0 \end{cases}, \tag{3.9}$$

on ψ_j és el coeficient per cada part $x_{j_1}, \ldots, x_{j_r}$ al numerador de y_i (codificat $+1$ al SBP), ψ_k és el coeficient per cada part $x_{k_1}, \ldots, x_{k_s}$ al denominador de y_i (codificat -1 al SBP) i ψ_0 és el coeficient per les parts que no intervenen.

Com ja hem dit anteriorment, la suma dels elements de $\boldsymbol{\psi}_i$ és zero perquè el vector es troba a l'hiperplà V. A més a mes, com que formen una base ortonormal, $\boldsymbol{\psi}_i \cdot \boldsymbol{\psi}_l$=0, per $i, l = 1, \ldots, D-1$, $i \neq l$, i $\|\boldsymbol{\psi}_i\|$=1.

Les coordenades de la composició en la base $\boldsymbol{\Psi}$ s'anomenen balanços (y_i) i les composicions de la base ($\mathbf{e}_i$) s'anomenen *balancing elements*. Cada element ilr de la base $\boldsymbol{\psi}_i$ és un log-contrast, és a dir, una combinació lineal de logaritmes de les dades composicionals amb coeficients de suma zero.

A la Taula 3.2 es mostra un exemple de SBP pel cas de $D = 3$.

Taula 3.2: Exemple de partició seqüencial binària (SBP): coordenades ilr $\boldsymbol{y} = (y_1, y_2)$ i base Ψ.

Ordre	x_1	x_2	x_3	r	s	Coordenada
1	-1	-1	1	1	2	$y_1 = \sqrt{\frac{1\cdot 2}{1+2}} \log \frac{x_3}{\sqrt{x_1 x_2}}$
2	-1	1	0	1	1	$y_2 = \sqrt{\frac{1\cdot 1}{1+1}} \log \frac{x_2}{x_1}$
$\boldsymbol{\Psi}$	$-\sqrt{\frac{1}{6}}$	$-\sqrt{\frac{1}{6}}$	$+\sqrt{\frac{2}{3}}$			
	$-\sqrt{\frac{1}{2}}$	$+\sqrt{\frac{1}{2}}$	0			

Podem utilitzar les operacions estàndards de l'espai real, treballar amb la distància euclidiana i aplicar el producte escalar ordinari sobre les dades ilr transformades.

Egozcue et al. (2003) donen les relacions entre les tres transformacions: alr, clr i ilr.

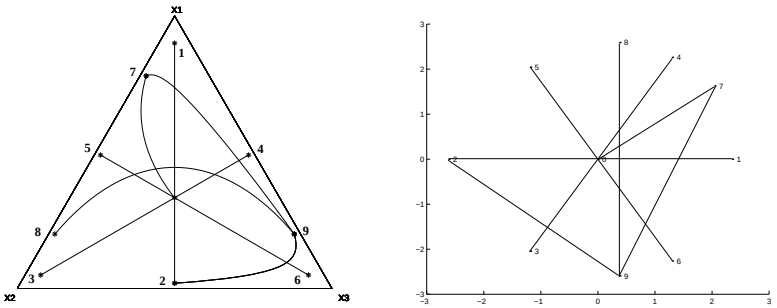
3.2.6 Tractament de zeros

Una dificultat important de la metodologia CoDa és la impossibilitat d'aplicar les transformacions logràtio a les composicions que tenen alguna de les seves components igual a zero.

Quan el valor nul es pot interpretar com un valor inferior a un determinat límit de detecció, té sentit reemplaçar el zero per un valor positiu molt petit utilitzant les tècniques de substitució que es fan servir habitualment en el tractament de les dades mancants, tot mantenint la coherència composicional de les dades.

Quan el valor nul representa un zero absolut, és a dir, absència total d'aquell component, l'estratègia a seguir es fonamenta en el fet de suposar que aquests zeros caracteritzen determinades subpoblacions, les quals han de ser modelitzades mitjançant models condicionals.

El detall dels mètodes que s'apliquen en aquestes dues circumstàncies es pot trobar a Martín-Fernández et al. (2011) i Palarea-albaladejo et al. (2014).



(a) Rectes composicionals al símplex $\mathcal{S}^3$

(b) Equivalència de les rectes composicionals a l'espai real $\mathbb{R}^2$.

Figura 3.6: Per visualitzar les relacions, angles, distàncies, ... cal representar les dades en coordenades.

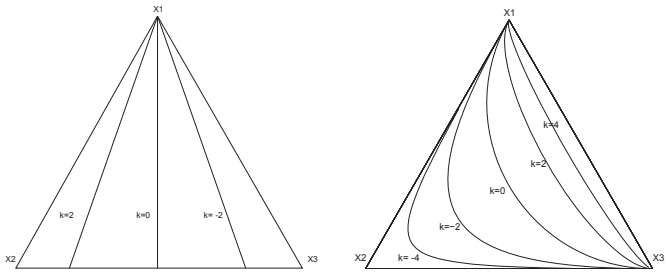


Figura 3.7: Rectes paral·leles al símplex. A l'esquerra, $\log x_2 - \log x_3 = k$ per a $k = -2, 0, 2$. A la dreta, $\log x_1 - 2\log x_2 + \log x_3 = k$ per a $k = -4, -2, 0, 2, 4$.

3.2.7 Geometria al símplex

Es mostren a continuació algunes figures que pretenen mostrar gràficament que la geometria al símplex és diferent a la geometria euclidiana amb la que estem acostumats a treballar a l'espai real.

La Figura 3.6 mostra rectes composicionals al símplex $\mathcal{S}^3$ i les rectes equivalents a l'espai transformat. És evident com les rectes perpendiculars de l'espai real 12 i 89 queden deformats un cop representats a l'espai restringit. El mateix passa amb els angles. Vegeu com l'angle recte entre els segments $\overline{50}$ i $\overline{07}$ de l'espai real (Figura 3.6b) queda deformat en la seva representació al símplex de la Figura 3.6a.

Les Figures 3.7 i 3.8 mostren exemples de famílies de rectes paral·leles i ortogonals en el símplex $\mathcal{S}^3$. A partir d'aquests gràfics resulta evident el fet que les imatges gràfiques que tenim de recta, paral·lelisme i ortogonalitat procedents de l'espai real no són vàlides a l'espai de les composicions, malgrat ser ambdós espais mètrics euclidians.

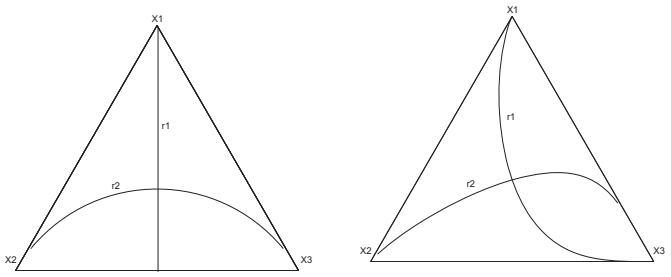


Figura 3.8: Rectes ortogonals a $\mathcal{S}^3$. A l'esquerra, $r_1 : x_2 = x_3$ i $r_2 : 2\log x_1 - \log x_2 - \log x_3 = 0$. A la dreta, $r_1 : \log x_1 - 3\log x_2 + 2\log x_3 = 0$ i $r_2 : 5\log x_1 - \log x_2 - 4\log x_3 = 0$

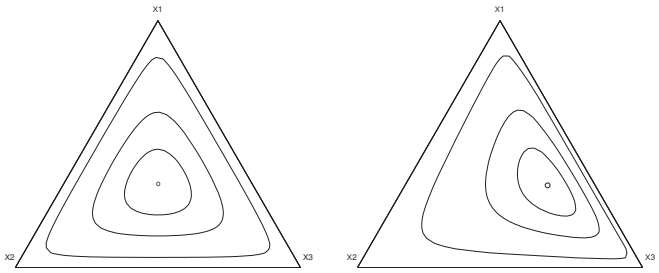


Figura 3.9: Circumferències a $\mathcal{S}^3$ de radis $r = 0.5, 1, 2$. A l'esquerra amb centre ($\circ$) a $(1/3, 1/3, 1/3)$ que és el baricentre del triangle i a la dreta a $(2/6, 1/6, 3/6)$.

Així, per exemple, observant les rectes de la Figura 3.7, resulta clar que el camí més curt entre dos punts del símplex no sempre és el segment rectilini entès en la forma "estàndard". Naturalment, però, si apliquéssim la transformació logràtio centrada (clr) a totes les rectes de representades a les Figures 3.7 i 3.8, obtindríem imatges estàndard de rectes paral·leles i ortogonals contingudes en el pla $z_1 + z_2 + z_3 = 0$ de $\mathbb{R}^3$.

Per acabar, la Figura 3.9, mostra les gràfiques d'unes quantes circumferències representades sobre $\mathcal{S}^3$. Igual com passava amb les rectes, els perfils d'aquestes circumferències composicionals no tenen res a veure amb els perfils estàndard d'aquestes figures. La proximitat a la frontera del símplex provoca distorsions en els perfils, des d'un punt de vista euclidià. Això és pel fet que la distància entre dos punts molt "propers" entre si (en el sentit estàndard del terme) situats gairebé tocant la frontera del triangle és molt més gran que la distància de dos punts amb la mateixa proximitat situats en la zona central del símplex.

Capítol 4

Articles

Universitat de Girona

El Dr. Josep Antoni Martín Fernández, com a coautor dels articles següents:

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014) **Individual T^2 Control Chart for Compositional Data**. *Journal of Quality Technology*, 46(2), pp. 127-139.

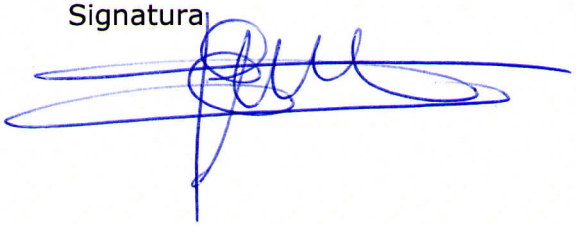
- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014) **Out-of-Control Signals in Three-Part Compositional T2 Control Chart**. *Quality and Reliabiity Engineering International*, 30 (3), pp. 337-346.

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. **Signal Interpretation in Hotelling's T2 Control Chart for Compositional Data**. *Submitted to Technometrics*

Accepto que la Sra. Marina Vives Mestres presenti els articles esmentats com a autor principal i com a part de la seva tesi doctoral, i que aquests articles no puguin, per tant, formar part de cap altra tesi doctoral.

I perquè així consti i tingui els efectes oportuns, signo aquest document.

Signatura



Girona, 15 de juliol de 2014

Universitat de Girona

El Dr. Josep Daunis i Estadella, com a coautor dels articles següents:

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014) **Individual T^2 Control Chart for Compositional Data**. *Journal of Quality Technology*, 46(2), pp. 127-139.

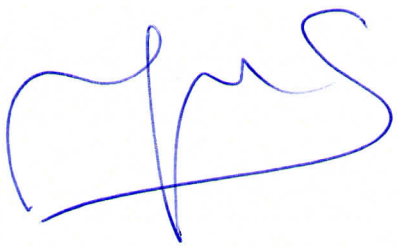
- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014) **Out-of-Control Signals in Three-Part Compositional T2 Control Chart**. *Quality and Reliabiity Engineering International*, 30 (3), pp. 337-346.

- Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. **Signal Interpretation in Hotelling's T2 Control Chart for Compositional Data**. *Submitted to Technometrics*

Accepto que la Sra. Marina Vives Mestres presenti els articles esmentats com a autor principal i com a part de la seva tesi doctoral, i que aquests articles no puguin, per tant, formar part de cap altra tesi doctoral.

I perquè així consti i tingui els efectes oportuns, signo aquest document.

Signatura



Girona, 15 de juliol de 2014

4.1 JQT

Aquest primer article cobreix els objectius O1, O2 i O6 descrits a la secció 2.1. En resum es pretén comparar gràfica i numèricament el gràfic de control T_C^2 composicional amb l'enfoc T^2 clàssic aplicat a composicions. L'article també mostra una aplicació pràctica industrial utilitzant un cas conegut de la literatura SPC.

L'article ha estat publicat a la revista Journal of Quality Technology.
Volum: 46 , Número: 2 , Pàgines: 127-139 , Publicat: Abril 2014
ISSN: 00224065
Factor d'impacte: primer quartil (Q1).

Vives-Mestres, M., Daunis-i-Estadella, J. and Martn-Fernandez, J. A. “Individual T2 Control Chart for Compositional Data”. *Journal of Quality Technology*. Vol. 46, issue 2 (2014) : 127-139

http://asq.org/pub/jqt/past/vol46-issue2/index.html

Abstract

The usual Hotelling T^2 control chart is not appropriate for monitoring processes where the quality characteristic is a mixture. The composition of mixtures are vectors of positive elements that represent parts of a whole, to which standard multivariate techniques are not appropriate due to their restricted sample space. There are many applications where a mixture is monitored against time, such as in the chemical industry, product composition, impurity profile, or gas components analysis. In this paper, a multivariate control chart for individual compositional observations based on the T^2 statistic is proposed and compared with the typical one in terms of average run length. The authors show how results are more consistent with compositional data nature and illustrate implementation in a real-world example.

Keywords

Hotelling's T2 statistic, Control charts, Mixture variables, Multivariate control charts, T2 control chart, Simplex, Statistical process control (SPC), Average run length (ARL)

4.2 QREI

El segon article cobreix els objectius O3, O4 i O6 descrits a la secció 2.1. Aquests objectius pretenen aportar una interpretació gràfica per determinar la causa de les observacions fora de control del gràfic T_C^2 pel cas de composicions de tres parts. En aquest cas també s'utilitza un exemple pràctic per mostrar l'aplicació del mètode proposat.

L'article ha estat publicat a la revista Quality and Reliability Engineering International.
Volum: 30, Número: 3, Pàgines: 337-346 , Publicat: Abril 2014 (online Octubre 2013)
DOI: 10.1002/qre.1583
Factor d'impacte: tercer quartil (Q3).

Vives-Mestres, M., Daunis-i-Estadella, J. and Martn-Fernandez, J. A. “Out-of-Control Signals in Three-Part Compositional T^2 Control Chart”. Quality and Reliability Engineering International. Vol. 30, issue 3 (2014) : 337-346

http://dx.doi.org/10.1002/qre.1583

http://onlinelibrary.wiley.com/doi/10.1002/qre.1583/abstract

Abstract

Interpretation of out-of-control signals in multivariate control charts is a persistent problem. Identifying the variables contributing to the signal is of crucial importance for process control. Both procedures turn out to be more complicated when observations are compositions (variables adding to a constant sum). Compositional T^2 control chart (T^2_C) is suitable for monitoring compositions because it is based on a transformation of the data, which moves it from restricted to real space. This paper proposes a method for interpreting the out-of-control signals of the T^2_C control chart for three-part compositions on the basis of an appropriate selection of the transformation. An example of industrial application illustrates the introduced techniques

Keywords

compositional data; Hotelling's T^2 statistic; multivariate control chart; signal interpretation; log-ratio

4.3 Enviat

El tercer article (enviat) cobreix els objectius O5 i O6 descrits a la secció 2.1. L'article proposa un algorisme que permet determinar la causa de les observacions fora de control del gràfic T_C^2 per a qualsevol nombre de parts de la composició.

L'article ha esta tenviat a la revista IIE Transactions.
Factor d'impacte: segon quartil (Q2).

Vives-Mestres, M., Daunis-i-Estadella, J. and Martn-Fernandez, J. A. “Signal Interpretation in Hotelling's T^2 Control Chart for Compositional Data. Submitted to *IIE Transactions*

http://www.tandfonline.com/loi/uiie20#.VJPasl4AOA

Abstract

Nowadays, control of concentrations of elements is of crucial importance in industry. Concentrations are expressed in terms of proportions or percentages which means that they are compositional data (CoDa). CoDa are defined as vectors of positive elements that represent parts of a whole and usually add to a constant sum. Classical T^2 control chart is not appropriate for CoDa, for which is better to use a compositional T^2 control chart (T^2_C CC). The T^2C CC is based on a transformation of the data into log ratios of components that moves it from restricted to real space. This paper fills the gap on interpretation of the out-of-control signals on the individual T^2_C CC. We show, for the three-component case, that the typical orthogonal decomposition procedure is misleading in order to identify the cause of the out-of-control signal. Two new methods based on finding the ratio of components for which the univariate T^2 statistic is maximum are proposed. We illustrate the T^2_C CC interpretation with a practical example from the chemical and pharmaceutical industry.

Keywords

Composition, Hotelling’s Statistic, Log ratio, Mixture, Multivariate Process Control, Signal Interpretation

Capítol 5

Resultats i discussió

En aquesta secció repassem els principals resultats que es deriven d'aquesta tesi, principalment pel que fa referència a l'adaptació a les dades composicionals del gràfic de control per a la T^2 de Hotelling i el gràfic p. També es detallen els primers passos pel que fa l'adaptació del gràfic CUSUM.

En literatura CoDa la mida de la composició es denota per D mentre que en el SPC multivariant, la mida del vector observat es denota per p. En la següent discussió utilitzarem la nomenclatura de SPC, que correspon amb la que hem fet servir als articles.

5.1 Gràfic T_C^2

Com ja hem comentat a la Secció 3.2, degut a la restricció de suma constant de les dades composicionals, la matriu de covariàncies clàssica d'un conjunt de CoDa és singular, i per tant no invertible mitjançant els mètodes estàndards. Aquest fet fa que no sigui possible calcular l'estadístic clàssic T^2 de l'Equació 3.3 per a CoDa.

Al la Secció 1.2 hem analitzat les solucions proposades a la literatura per resoldre la dificultat de la singularitat de la matriu de covariàncies.

La primera solució, i la més àmpliament acceptada, per evitar la singularitat de la matriu de covariàncies consisteix en eliminar una de les variables de la composició ja que, de fet, coneixent les altres components sempre es pot deduir el valor del component eliminat. Aquest mètode proporciona una regió de control amb forma hiperel·líptica al voltant de la mitjana aritmètica de les variables.

És fàcil intuir que aquestes formes (hiper)el·líptiques no tenen perquè ajustar-se a l'espai restringit de les dades composicionals. Això es pot constatar quan es representa l'el·lipse resultant d'aplicar aquest procediment a un conjunt de composicions de tres parts ($p = 3$). Per exemplificar-ho hem simulat dos conjunts de dades amb característiques diferents: són, d'una banda, un conjunt en forma d'arc situat a la part alta del diagrama ternari

i, de l'altra, un conjunt proper a un dels vèrtexs, que es representen a la Figura 5.1 juntament amb les regions de control clàssiques.

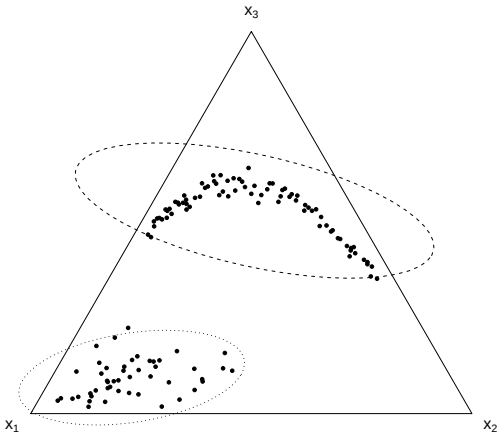


Figura 5.1: El·lipses de control aplicant el mètode T^2 clàssic al conjunt de dades simulades en forma d'arc (línia discontinua) i al conjunt proper al vèrtex x_1 (línia puntejada).

Recordem que la regió de control obtinguda, així com el valor de l'estadístic T^2 és el mateix sigui quin sigui la part eliminada.

En la Figura 5.1 s'observa com el conjunt en forma d'arc no queda ben ajustat per la regió de control, i que aquesta surt de l'espai del símplex. El mateix passa amb el conjunt situat prop del vèrtex x_1: la regió de control accepta valors que no són possibles donada la característica restringida de les dades. Observem com, en aquest cas, el gràfic de control no serà capaç de detectar increments de x_1 en el procés.

La solució proposada a la literatura que consisteix en aplicar components principals a les CoDa i retenir només els més grans, tampoc aporta un resultat satisfactori. Tornant a l'exemple de les dades en forma d'arc, seleccionar el primer component principal és equivalent a seleccionar una recta continguda en el pla $x_1 + x_2 + x_3 = \kappa$, i és evident que aquesta recta mai podrà ajustar un conjunt de dades amb aquesta forma d'arc.

Per finalitzar amb el repàs dels inconvenients del mètodes proposats a la literatura clàssica, veiem el cas en el que només es treballa amb l'espai determinat per $D - 1$ parts de la composició, i per tant s'aplica el mètode estàndard a un conjunt de components que no sumen una constant. En aquest cas, no hi ha cap problema a l'hora de calcular l'estadístic T^2. Per mostrar l'inconvenient de l'ús d'aquest mètode hem fet, altre cop, una simulació de composicions de tres parts. El conjunt està representat en el diagrama ternari a la Figura 5.2, i li hem afegit una observació atípica (■).

A la Figura 5.2, podem veure com l'observació atípica, ho és perquè té un

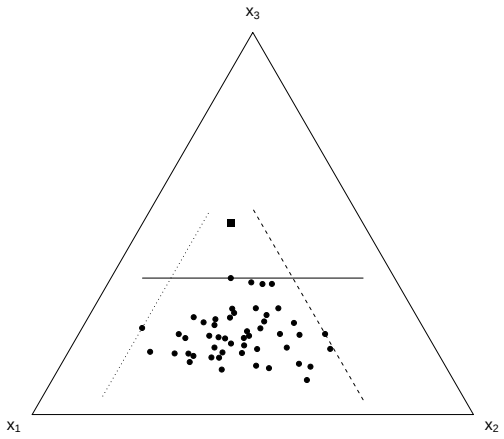


Figura 5.2: Dades CoDa simulades amb l'adició d'una observació atípica (■). Delimitació dels menor valor de x_1 (línia puntejada), el menor x_2 (línia discontinua) i el major x_3 (línia sòlida).

valor de x_3 superior a la resta, mentre que els valors de x_1 i x_2 no són extrems. Si l'observador, en comptes de recollir dades de la composició completa, només recull dades de la marginal bivariant (x_1, x_2), estarà treballant amb la projecció dels punts del pla $x_1 + x_2 + x_3 = \kappa$ sobre el pla $x_3 = 0$ (Figura 5.3a). La regió de control de la T^2 dibuixada en aquest pla (Figura 5.3b), mostra com l'observació atípica, ho és perquè la relació entre els valors x_1 i x_2 no és correcta. Aquesta conclusió no correspon amb el que mostra el diagrama ternari i per tant amb l'anàlisi de la composició completa. Aquest fet es dona perquè el mètode clàssic del càlcul de la T^2 no compleix amb el principi de coherència subcomposicional descrit a la Secció 3.2.4.

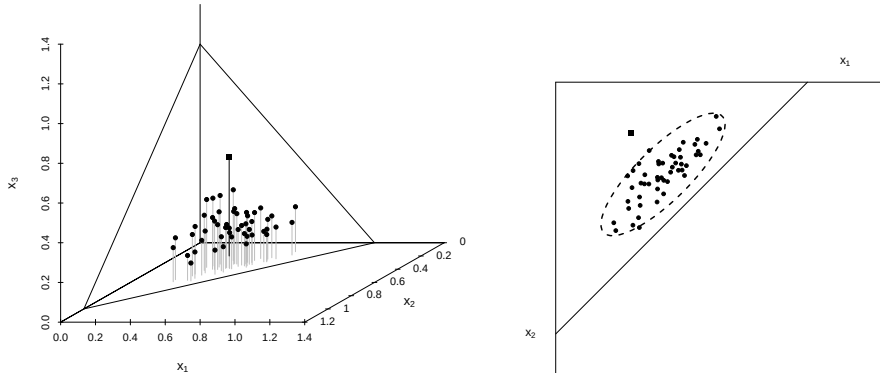
No comentem aquí els principis d'invariància per escala ni invariància per permutació, perquè aquests ja es compleixen amb la T^2 clàssica, és a dir, el valor de la T^2 és el mateix siguin quines siguin les unitats de les dades i o l'ordre dels components.

Un cop vist que els mètodes proposats a la literatura per al control de CoDa amb l'estadístic T^2 no són coherents amb les característiques de les dades, hem proposat un nou esquema de control. El CC proposat s'anomena gràfic de control T^2 per dades composicionals i el denotem amb el subíndex C: T^2_C.

L'estadístic T^2_C es defineix de la següent manera: donat $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, una composició de p parts i $\boldsymbol{y} = (y_1, \ldots, y_{p-1})$, les seves coordenades ilr definides usant l'Equació 3.9, el valor de T^2_C es calcula com

$$T^2_C = (\boldsymbol{y} - \boldsymbol{\mu}_y)' \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_y) \tag{5.1}$$

On $\boldsymbol{z}$ és la coordenada de la composició observada i $\boldsymbol{\mu}_y$ i $\boldsymbol{\Sigma}_y$ són el vector de mitjana i la matriu de covariàncies de les coordenades logràtio, respecti-



(a) Marginal (x_1, x_2) (b) Projecció al pla $x_3 = 0$

Figura 5.3: Treballar amb la marginal (x_1, x_2) equival a projectar el conjunt de dades al pla $x_3 = 0$ (a). La regió de control de la projecció es porta a una conclusió errònia sobre la causa de l'anomalia a l'observació atípica ■ (b).

vament. A la pràctica, cal estimar ambdós paràmetres amb un conjunt de dades sota control a la fase I, de la mateixa manera que es fa en els mètodes estàndard.

Assumim que els vectors $\boldsymbol{y}$ són independents i idènticament distribuïts segons una distribució normal multivariant $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$.

L'estadístic T_C^2 no queda afectat per la base triada per calcular les coordenades ilr. A la pràctica, l'usuari ha de triar una base que li resulti fàcil per interpretar les dades. De fet, el T_C^2 també es pot calcular a partir de les coordenades clr ($\boldsymbol{z}$); només cal substituir les $\boldsymbol{y}$ de l'Equació 5.1 per $\boldsymbol{z}$ després d'eliminar un component qualsevol, i canviar la mitjana i la matriu de covariàncies de les coordenades ilr per les de la clr després d'eliminar el mateix component que abans.

L'estadístic T_C^2 compara cada observació amb la mitjana geomètrica del conjunt, que és una millor mesura de centre que la mitjana aritmètica ja que, de forma habitual, la distribució univariant de les dades composicionals no segueix una distribució normal sinó log-normal.

Aquesta millor determinació del centre es pot veure perfectament en el conjunt de dades en forma d'arc. A la Figura 5.4 es mostren els gràfics de control T^2 fent servir el mètode clàssic i el mètode composicional. Veiem com la mitjana aritmètica, representada per un $\triangle$, es troba separada del conjunt de dades mentre que la mitjana geomètrica, representada per un $\square$, queda centrada entre les dades. De fet, la mitjana aritmètica està tant separada del conjunt, i precisament en una direcció amb poca variabilitat, que apareix com a atípica en el gràfic T_C^2. Un altre aspecte a destacar és que, usant el mètode clàssic, les observacions que estan a la cua de l'arc apareixen com a atípiques mentre que amb el T_C^2 estan dins la regió de control. Per al càlcul dels límits de control de la Figura 5.4 s'ha utilitzat la distribució

Beta amb $\alpha = 0.96$.

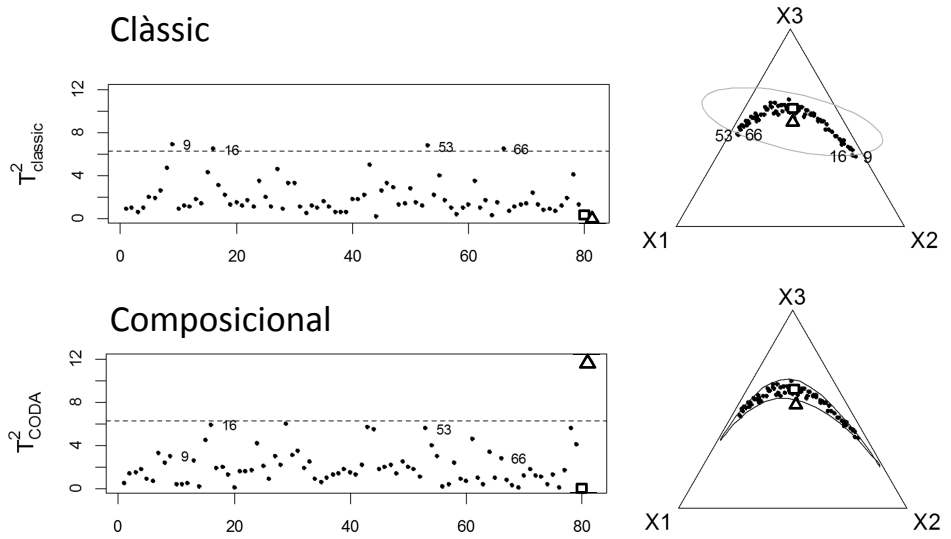


Figura 5.4: Gràfic de control T^2 clàssic (superior) i composicional (inferior). La mitjana geomètrica (□) és una millor mesura de centre de la distribució que la mitjana aritmètica (△).

Un cop definit el gràfic de control composicional T^2_C, hem comparat el funcionament d'aquest amb el T^2 clàssic en termes de ARL i dels quartils de la RL. Per al càlcul de la RL hem fet servir un esquema similar al proposat per Champ et al. (2005), per al qual és necessari simular les dades del procés, que es van comparant amb el límit de control.

El primer inconvenient que hem trobat està en com simular les dades composicionals. Una opció consisteix en simular $p-1$ components normals multivariants i calcular el restant fent la diferència, però aquest mètode no assegura l'existència del p-èssim component. És per això que hem simulat coordenades normals multivariants, concretament de dimensió $p-1=2$, és a dir, dades normals al símplex $\mathcal{N}_{\mathcal{S}^3}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ amb paràmetres coneguts

$$\boldsymbol{\mu}_y = (0,0) \quad \boldsymbol{\Sigma}_y = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$$

On $\boldsymbol{\mu}_y$ i $\boldsymbol{\Sigma}_y$ són els paràmetres de la distribució normal multivariant a l'espai de coordenades $\mathbb{R}^2$.

Observem que la matriu de covariàncies és diagonal, és a dir, que la correlació entre les logràtios és nul·la i que la mitjana està situada al centre del diagrama ternari. Aquest és un dels escenaris utilitzats per comparar el funcionament del clàssic T^2 i el T^2_C, que hem anomenat escenari 0 a la Taula 5.1.

Hem considerat 7 escenaris més, que corresponen a distribucions normals multivariants al símplex amb la mateixa matriu de covariàncies diagonal anterior, però amb un vector de mitjanes que es desplaça cap al vèrtex x_3, és a dir, on el component x_3 és present cada cop amb més quantitat. D'aquesta manera podrem comparar el funcionament dels dos gràfics en distribucions homogènies (escenari 0 de la Taula 5.1) i molt heterogènies (escenari 7 de la Taula 5.1).

Taula 5.1: Valors del vector de mitjanes considerats en els 8 escenaris de la simulació per calcular l'ARL del gràfic clàssic T^2 i el composicional T^2_C.

	$\boldsymbol{\mu}_x$				$\boldsymbol{\mu}_x$		
Escenari	x_1	x_2	x_3	Escenari	x_1	x_2	x_3
0	0.33	0.33	0.33	4	0.17	0.17	0.67
1	0.29	0.29	0.42	5	0.12	0.12	0.75
2	0.25	0.25	0.50	6	0.08	0.08	0.83
3	0.21	0.21	0.58	7	0.04	0.04	0.92

Notem que els paràmetres de la distribució normal es defineixen a l'espai de coordenades ($\boldsymbol{z}$). No obstant, per al càlcul de l'estadístic T^2 clàssic, són necessaris aquests paràmetres a l'espai de composicions ($\mathbf{x}$). Per obtenir la mitjana $\boldsymbol{\mu}_x$ de la composició fem $\boldsymbol{\mu}_x = \text{ilr}^{-1}(\boldsymbol{\mu}_z)$, però no existeix cap formula exacta per obtenir $\boldsymbol{\Sigma}_x$ a partir de $\boldsymbol{\Sigma}_z$.

Estimem $\boldsymbol{\Sigma}_z$ a partir de la ilr^{-1} d'un milió de mostres de la distribució $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ per cadascun dels vectors de mitjana de la Taula 5.1. Considerarem que aquesta és la matriu de covariàncies coneguda de les $\mathbf{x}$.

Els resultats pel que fa a la mitjana de la RL (ARL) es mostren a la Figura 5.5. Cada punt del gràfic correspon a la mitjana de $100,000$ valors de RL, és a dir, el nombre mitjà de punts que apareixen fins que el gràfic en senyala un fora de control.

Recordem, tal i com hem vist a la Secció 3.1.3, que quan les observacions són independents, idènticament distribuïdes i el UCL és constant, la RL segueix una distribució geomètrica amb mitjana ARL$= 1/\alpha$. Per $\alpha = 0.005$ la mitjana és ARL$= 200$. Aquest valor es manté constant pel cas del gràfic T^2_C, però no és així en el cas del gràfic T^2 clàssic quan la distribució es mou cap al vèrtex.

Del gràfic de la Figura 5.5 se'n desprèn que els dos esquemes es comporten de manera similar quan la distribució està centrada al símplex. No obstant, quan la distribució es desplaça cap a un dels extrems, el gràfic CoDa manté el mateix rendiment, mentre que en el cas clàssic baixa dràsticament. Això és degut a la forma com intereseccionen les regions de control dels dos gràfics.

El gràfic de control T^2_C s'ha aplicat a un exemple industrial clàssic de la literatura del control de processos. Les dades apareixen publicades per

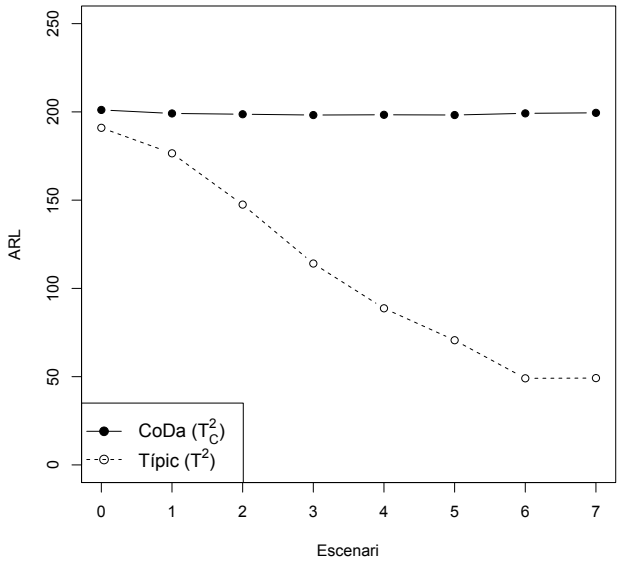


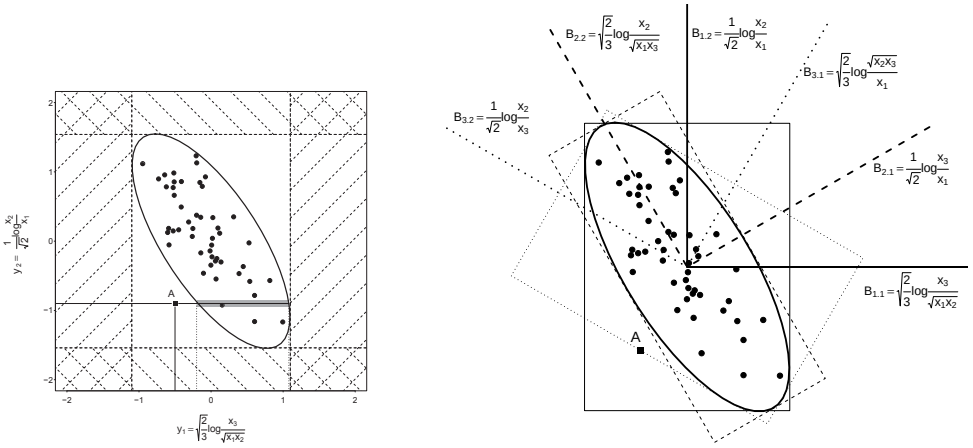
Figura 5.5: Comparació de l'ARL del gràfic T^2_C amb el clàssic T^2 pel als 8 escenaris simulats.

primer cop a Holmes and Mergen (1993) i s'utilitzen també a Sullivan and Woodall (1996) i es reprodueixen a Montgomery (2009). Les dades descriuen el percentatge de partícules que es classifiquen en petites mitjanes i grans, i són clarament composicionals. L'aplicació del CC T^2_C permet comprovar com les observacions indicades fora de control amb l'esquema clàssic, en realitat no ho són, mentre que hi ha una observació que no s'havia detectat prèviament que sí que es troba fora de control. L'anàlisi detallat de les dades i la seva representació gràfica permet veure clarament la diferència entre els dos mètodes i el significat de les observacions fora de control.

Tot el que s'ha discutit fins al moment representa el contingut de l'article: Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014), **Individual T^2 Control Chart for Compositional Data**. Journal of Quality Technology, 46(2), pp. 127-139.

Un cop proposat el nou esquema de control, comparat el seu funcionament amb l'esquema clàssic i aplicat a un exemple industrial, hem anat un pas més enllà i hem treballat la interpretació de les observacions fora de control. Com ja hem comentat a la Secció 3.1.4, és de vital importància la identificació de les causes que provoquen el senyal per tal de poder emprendre mesures correctives.

A partir d'exemples numèrics i gràfics, hem demostrat que la interpretació dels signes fora de control del T^2_C CC mitjançant la descomposició



(a) Interpretació del terme condicional.

(b) Possibles bases ilr a $\mathbb{R}^2$.

Figura 5.6: Dades simulades amb una observació atípica A. Utilitzant la base per defecte $\mathrm{B}_1 = \{\mathrm{B}_{1.1}, \mathrm{B}_{1.2}\}$ s'obté un senyal en el terme condicional $T^2_{y_1.y_2}$ mentre que si s'utilitza la base $\mathrm{B}_2 = \{\mathrm{B}_{2.1}, \mathrm{B}_{2.2}\}$ el terme significatiu és el no condicional $T^2_{y_1}$.

MYT clàssica, és enganyosa. Per resumir-ho direm que, com que els termes condicionats són logràtios de components i comparteixen components, la conclusió sobre quin és el component responsable del fora de control no és evident.

Per mostrar aquesta ambigüitat, simulem unes coordenades normals multivariants a $\mathbb{R}^2$ i afegim un atípic (■) que denotem per A. Les dades es representen a la Figura 5.6a. Utilitzant la base ilr per defecte B_1 que es mostra a la Figura 5.6b obtenim que el terme que contribueix significativament al senyal fora de control és el terme condicional $T^2_{z_1.z_2}$. Com que z_1 i z_2 comparteixen els components x_1 i x_2 no és possible saber on és la causa real del problema.

En canvi de la Figura 5.6b es desprèn que, utilitzant la base $\mathrm{B}_2 = \{\mathrm{B}_{2.1}, \mathrm{B}_{2.2}\}$ obtindrem que el terme que contribueix significativament al fora de control és el corresponent a l'eix 1 d'aquesta base ($\mathrm{B}_{2.1}$), és a dir, a la ràtio $\frac{1}{\sqrt{2}} \log \frac{x_3}{x_1}$. Aquesta informació és molt més precisa que no pas la informació que ens donava el terme condicional inicial.

A partir d'aquestes observacions, proposem un mètode per identificar les causes de les observacions fora de control per $p = 3$ basat en trobar la base en la qual la descomposició de la T^2_C té un pes superior en els termes no condicionals i menor en els condicionals. Aquesta base, o més ben dit, el primer eix d'aquesta base, es troba maximitzant el valor de $T^2_{y_1} = \frac{(y_1 - \mu_{y_1})^2}{\sigma^2_{y_1}}$ de forma que sigui igual al valor global de T^2_C. Com que probablement l'eix

que defineix y_1 no es pot expressar en forma de balanç de components amb exponents enters, proposem aproximar-ho pel balanç més proper.

Aquest mètode gràfic i intuïtiu per interpretar els senyals fora de control en el cas de $p = 3$, així com la justificació de que la descomposició MYT no es pot aplicar directament a les coordenades, es presenta a l'article Vives-Mestres, M., Daunis-i-Estadella, J. i Martín-Fernández, J. A. (2014), **Out-of-Control Signals in Three-Part Compositional T^2 Control Chart**. Quality and Reliability Engineering International, 30 (3), pp. 337-346.

En el mateix article es mostra l'aplicació del mètode a l'exemple industrial del percentatge de partícules grans (L), mitjanes (M) i petites (S) de l'article anterior. Es mostra com la causa de l'única observació fora de control del gràfic T_C^2, es troba en un valor erroni de la ràtio $\sqrt{\frac{2}{3}}\log\frac{S}{\sqrt{LM}}$. De les dades podem veure com, aquesta ràtio, és gairebé 6 vegades més gran que la mateixa ràtio de la mitjana geomètrica, cosa que justifica el seu valor atípic.

El darrer avenç que aporta aquesta tesi consisteix en proposar un algorisme que permet identificar la causa del senyal fora de control per a qualsevol mida de p. Aquest mètode es basa en el principi establert a l'article anterior; els termes condicionals són enganyosos a l'hora d'identificar la causa de l'anomalia. Per tant, proposa buscar el logràtio que fa que el terme no condicional sigui màxim.

Concretament es proposen dos mètodes. El primer consisteix en calcular tots els termes no condicionals per a tots els balanços de components, és a dir calcular

$$T^2 = \frac{(z-\mu_z)^2}{\sigma_z^2}, \tag{5.2}$$

on z és un logràtio de components i μ_z i σ_z^2 són, respectivament, la mitjana i la variància del mateix logràtio. El logràtio que faci màxim el terme T^2 és el responsable del senyal ja que una descomposició de l'estadístic T_C^2 utilitzant aquest component donarà el màxim pes en el terme no condicionat. Aquest mètode és costós computacionalment perquè, quan el nombre de components augmenta, també augmenta de forma exponencial el nombre de possibles logràtios. És per aquest motiu que es proposa un segon mètode.

Aquest es basa en el principi de que, si les coordenades estan centrades a l'origen i distribuïdes en forma d'esfera, és a dir, que la seva matriu de covariàncies és la matriu identitat, la direcció en la qual la descomposició del T_C^2 d'un atípic dóna zero en els termes condicionals, és aquella que va des de l'origen fins a l'atípic.

Tal i com passava amb el mètode gràfic per $p = 3$, és molt probable que la direcció que va des de l'origen fins a l'atípic no sigui interpretable en termes balanç de components. És per això que aquesta direcció s'aproxima a la ràtio més propera. Aquesta aproximació es fa utilitzant un algorisme de cerca anomenat NN (de l'anglès *Nearest Neighbour*), que troba el punt

més proper, dins d'una llista prèvia de direccions logràtio.

Aquest és el procediment en el cas de que la distribució de les coordenades sigui esfèrica, cosa poc probable en els problemes reals. Es proposa per tant, esferificar les coordenades ($\boldsymbol{y}_s$) aplicant la transformació $\boldsymbol{y}_s = (\boldsymbol{y} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1/2}$ on $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ són la mitjana i la matriu de covariàncies de les coordenades ilr. La llista de direccions logràtio també s'ha de transformar de forma que el valor de la T^2 univariant en una direcció de les coordenades originals sigui la mateixa en la direcció de l'espai esferificat.

Proposem utilitzar el primer mètode per a valors de $p < 11$ ja que el temps de càlcul és pràcticament negligible. En canvi per valors $p \geq 11$ recomanem l'ús del segon mètode, que escurça el temps de càlcul, però té l'inconvenient que requereix actualitzar la llista de direccions logràtio cada cop que canvia la relació de covariàncies entre les coordenades.

Els dos programes computacionals per identificar la causa de les observacions fora de control del gràfic T^2_C es descriuen a l'article Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. **Signal Interpretation in Hotelling's T^2 Control Chart for Compositional Data** que ha estat enviat a la revista IIE Transactions.

5.2 Gràfic p

Detallem a continuació el treball desenvolupat pel que fa referència al gràfic de control per a una proporció (gràfic p).

Des d'un punt de vista composicional, la proporció d'unitats defectuoses que es controla clàssicament amb el gràfic de Shewhart p, no és una variable univariant sinó que representa el vector $\boldsymbol{p} = (p, 1-p)$ que es troba al símplex $\mathcal{S}^2$. L'enfoc proposat consisteix en fer un control de la coordenada logràtio d'aquest vector.

Suposem que x és una variable aleatòria distribuïda segons $\text{Bin}(n, p)$ que representa el nombre d'unitats defectuoses en una mostra de mida n amb probabilitat d'èxit p. La coordenada ilr del vector $\boldsymbol{p} = (x/n, (n-x)/n)$ és proporcional a la funció logit, és a dir:

$$\text{ilr}(\boldsymbol{p}) = \frac{1}{\sqrt{2}}\text{logit}(p) = \frac{1}{\sqrt{2}}\log(\frac{p}{1-p})$$

Per a construir el CC, és necessari disposar del valor de la mitjana i de la desviació de la coordenada ilr. Es dona que, quan $p \to p_0$

$$E[\text{ilr}(\boldsymbol{p})] \approx \frac{1}{\sqrt{2}}\text{logit}(p_0)$$

$$Var[\text{ilr}(\boldsymbol{p})] \approx \frac{1}{2np_0(1-p_0)} \tag{5.3}$$

Demostració 5.1 Sabem que la derivada de la coordenada ilr és

$$D(\text{ilr}(\boldsymbol{p})) = \frac{1}{\sqrt{2}}\frac{1}{p(1-p)}$$

La sèrie de Taylor de segon ordre de la funció ilr és

$$\text{ilr}(\boldsymbol{p}) = \frac{1}{\sqrt{2}}\text{logit}(p_0) + \frac{1}{\sqrt{2}}\frac{1}{p_0(1-p_0)}(p-p_0) + \delta_2(p,p_0)$$

$$\lim_{p\to p_0}\frac{\delta_2(p,p_0)}{p-p_0} = 0$$

I

$$\text{ilr}(p_0+\delta) = \frac{1}{\sqrt{2}}\text{logit}(p_0) + \frac{1}{\sqrt{2}}\frac{1}{p_0(1-p_0)}\delta + \delta_2(p_0,\delta)$$

$$\lim_{\delta\to 0}\frac{\delta_2(p_0,\delta)}{\delta} = 0$$

Per tant, una aproximació de ilr($\boldsymbol{p}$) és

$$\text{ilr}(\boldsymbol{p}) = \frac{1}{\sqrt{2}}\text{logit}(p) \approx \frac{1}{\sqrt{2}}\text{logit}(p_0) + \frac{1}{\sqrt{2}}\frac{1}{p_0(1-p_0)}(p-p_0)$$

De la distribució binomial sabem que $E[x] = np$ i $Var[x] = np(1-p)$. Com que $p = x/n$, la mitjana i la variància de la coordenada ilr es pot expressar com

$$E[\text{ilr}(\boldsymbol{p})] \approx \frac{1}{\sqrt{2}}\text{logit}(p_0) + \frac{1}{\sqrt{2}}\frac{1}{p_0(1-p_0)}\,(E[p]-p_0) = \frac{1}{\sqrt{2}}\text{logit}(p_0) + \frac{1}{\sqrt{2}}\frac{1}{p_0(1-p_0)}\,(p-p_0)$$

$$Var[\text{ilr}(\boldsymbol{p})] \approx \frac{1}{2[p_0(1-p_0)]^2}Var(p-p_0) = \frac{1}{2[p_0(1-p_0)]^2}Var(p) = = \frac{1}{2[p_0(1-p_0)]^2}\frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n[p_0(1-p_0)]^2}$$

Podem veure com quan $p \to p_0$

$$E[\text{ilr}(\boldsymbol{p})] \approx \frac{1}{\sqrt{2}}\text{logit}(p_0)$$

$$Var[\text{ilr}(\boldsymbol{p})] \approx \frac{1}{2np_0(1-p_0)}$$

Quan n i p són petites, hi ha moltes mostres en les que el nombre de no conformitats és 0 i per tant la proporció de defectuoses també és zero, és a dir, $\boldsymbol{p} = (0, 1)$. De la mateixa manera, quan n és petita i p és gran (propera a 1) hi ha moltes mostres en les que el nombre de no conformitats és n i per tant la proporció de defectuoses és 1, és a dir, $\boldsymbol{p} = (1, 0)$. En tots dos casos, no és possible calcular la coordenada ilr del vector $\boldsymbol{p}$.

Per evitar aquest problema proposem utilitzar una tècnica de reemplaçament de zeros proposada per Martín-Fernández et al. (2011), concretament la tècnica baiesiana multiplicativa utilitzant el prior de Perks, que suggereix reemplaçar el vector $\mathbf{x} = (0, n)$ per

$$\mathbf{x}_r = \left(\frac{1/2}{n+1}, \frac{n+1/2}{n+1} \right) \tag{5.4}$$

Per tant el vector $\boldsymbol{p} = \mathbf{x}/n$ serà reemplaçat per $\boldsymbol{p}_r = \mathbf{x}_r/n$.

Procedim a fer una simulació per verificar que efectivament la mitjana i la variància de la coordenada ilr s'ajusten a l'aproximació proposada. Comparem el mètode sense reemplaçament (obviant els valors 0 i 1) amb el mètode amb reemplaçament segons l'Equació 5.4. Per això simulem deu milions de variables x procedents d'una distribució $\text{Bin}(n, p)$ i en calculem la coordenada ilr d'una banda fent reemplaçament de zeros (Equació 5.4) i de l'altra ometent-los. Llavors calculem la mitjana i la desviació dels dos conjunts resultants.

En la simulació considerem valors de $p = 0.001, 0.01, 0.05, 0.25, 0.45$ i $n = 10, 50, 250, 1000$ combinats de forma que resulten en 20 casos per simular. Els resultats es mostren a la Taula 5.7 on la part esquerra mostra els resultats referents a la mitjana i la part dreta els referents a la variància.

La Taula 5.7 mostra que s'obtenen millors resultats amb el reemplaçament de zeros que sense, és a dir, el valor resultat de la simulació s'acosta més al valor teòric de l'Equació 5.2. També observem com els valors de la simulació (ESim pel càlcul sense reemplaçament i ESim0Repl amb reemplaçament) tendeixen a tenir valors inferiors al valor teòric (ETeo) perquè el terme de primer ordre de la sèrie de Taylor que es deprecia a l'aproximació és positiu.

A partir de la Taula 5.7 s'aprecia com, per valors petits de n com ara 0.01 o bé 0.05, calen mides de mostra n grans per tal d'aconseguir una bona aproximació tant de la mitjana com de la variància. Procedim a fer més anàlisis per a mides de mostra més grans per veure com tendeixen els resultats de les simulacions als valors teòrics.

La Figura 5.8 mostra un gràfic fruit d'aquestes anàlisis on s'observa l'evolució de la mitjana a mida que augmenta la mida de la mostra n per valors de $p = 0.001, 0.01$. Per aconseguir una bona aproximació, quan $p = 0.01$ calen mides de mostra superiors a 2000. Quan la $p = 0.001$, aquest valor s'incrementa fins a més de 10000. Això suposa un greu inconvenient perquè sovint no és factible disposar de mides de mostra tant grans.

p	n	ETeo	ESim	Esim 0Repl	VarTeo	VarSim	VarSim 0Repl	%zerosTeo
0.001	10	-4.88381	-1.55117	-2.14681	50.05005	0.00143	0.00358	**99%**
0.001	50	-4.88381	-2.73941	-3.23781	10.01001	0.00627	0.01305	**95%**
0.001	250	-4.88381	-3.83957	-4.27274	2.00200	0.02902	0.05973	**78%**
0.001	1000	-4.88381	-4.63500	-4.90706	0.50050	0.09821	0.18942	**37%**
0.01	10	-3.24924	-1.52783	-2.09315	5.05051	0.01459	0.03512	**90%**
0.01	50	-3.24924	-2.62603	-3.01157	1.01010	0.05750	0.11980	**61%**
0.01	250	-3.24924	3.29294	-3.38228	0.20202	0.15929	0.23694	**8%**
0.01	1000	-3.24924	-3.28783	-3.28792	0.05051	0.06055	0.06072	**0%**
0.05	10	-2.08203	-1.41889	-1.85833	1.05263	0.07096	0.15788	**60%**
0.05	50	-2.08203	-2.12378	-2.21148	0.21053	0.16921	0.24844	**8%**
0.05	250	-2.08203	-2.11092	-2.11092	0.04211	0.04794	0.04795	**0%**
0.05	1000	-2.08203	-2.08885	-2.08885	0.01053	0.01084	0.01084	**0%**
0.25	10	-0.77684	-0.79746	-0.87379	0.26667	0.24180	0.32582	**6%**
0.25	50	-0.77684	-0.79724	-0.79724	0.05333	0.05847	0.05847	**0%**
0.25	250	-0.77684	-0.78064	-0.78064	0.01067	0.01085	0.01085	**0%**
0.25	1000	-0.77684	-0.77775	-0.77775	0.00267	0.00268	0.00268	**0%**
0.45	10	-0.14190	-0.15597	-0.16026	0.20202	0.25116	0.26238	**0%**
0.45	50	-0.14190	-0.14493	-0.14493	0.04040	0.04220	0.04220	**0%**
0.45	250	-0.14190	-0.14241	-0.14241	0.00808	0.00815	0.00815	**0%**
0.45	1000	-0.14190	-0.14205	-0.14205	0.00202	0.00202	0.00202	**0%**

Figura 5.7: Resultats de la simulació per les 20 combinacions de n i p. ETeo i VarTeo representen l'esperança i variància teòriques donats els valors de n i p calculats utilitzant la Equació 5.2. ESim i VarSim són l'esperança i la variància resultants de la simulació un cop omesos tots els valors de $p = 0$. Esim 0Repl i VarSim 0Repl són l'esperança i la variància resultants de la simulació fent el reemplaçament dels zeros.

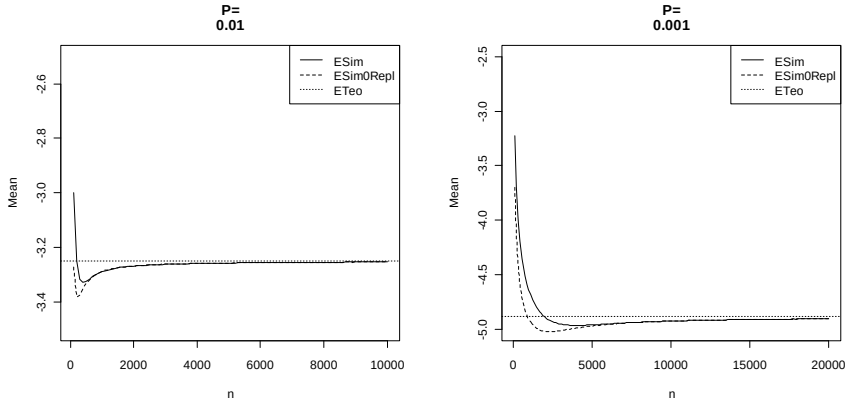


Figura 5.8: Evolució de la mitjana en funció de n segons el mètode amb i sense reemplaçament de zeros per valors de $p = 0.01, 0.001$.

Amb aquests resultats proposem un nou gràfic de control per a la proporció d'unitats defectuoses basat en la transformació ilr. El CC queda definit de la següent manera.

$$
\begin{aligned}
UCL &= E[ilr(\boldsymbol{p})] + 3\sqrt{Var[ilr(\boldsymbol{p})]} \\
CCL &= E[ilr(\boldsymbol{p})] \\
LCL &= E[ilr(\boldsymbol{p})] - 3\sqrt{Var[ilr(\boldsymbol{p})]}
\end{aligned}
$$

Els límits anteriors es poden expressar en termes de la variable original

$$
\begin{aligned}
UCL &= \frac{\sum \text{logit}(p)}{m\sqrt{2}} + 3\sqrt{\frac{1}{2np(1-p)}} \\
CCL &= \frac{\sum \text{logit}(p)}{m\sqrt{2}} \\
LCL &= \frac{\sum \text{logit}(p)}{m\sqrt{2}} - 3\sqrt{\frac{1}{2np(1-p)}}
\end{aligned}
$$

On p és el valor conegut de la proporció a controlar o bé és estimat a partir d'una mostra de mida m d'observacions que se sap estan sota control (fase I): $p = \sqrt[m]{\prod p_i}$ on $i = 1, \ldots, m$.

Hem comparat aquest gràfic de control amb el CC basat en la transformació arcsin mitjançant l'ARL. La probabilitat de cometre un error tipus II (β) del gràfic proposat és

$$
\begin{aligned}
\beta =& P(LCL \leq x \leq UCL) = \\
& P\left(\frac{n e^{\text{logit}(p)}}{e^{3\sqrt{\frac{1}{np(1-p)}}} + e^{\text{logit}(p)}} \leq x \leq \frac{n}{(e^{3\sqrt{\frac{1}{np(1-p)}}} e^{\text{logit}(p)})^{-1} + 1} \right)
\end{aligned}
$$

A partir d'aquí podem calcular l'ARL mitjançant l'Equació 3.1 per diferents valors de desplaçament de la mitjana i de p. A la Figura 5.9 es mostren les corbes de l'ARL per al CC de Shewhart amb $n = 1000$ i $p = 0.01$, el gràfic arcsin i el gràfic composicional. El desplaçament està expressat com a ràtio del valor de p desplaçat respecte al valor de sota control p_0.

De la Figura 5.9 se'n desprèn que el gràfic composicional té un comportament molt semblant al gràfic amb la transformació arcsinus; el gràfic composicional té un comportament lleugerament superior a l'arcsinus quan disminueix el valor de p mentre que succeeix a la inversa quan augmenta la p. Ambdós gràfics tenen comportaments similars ja que les seves funcions (arcsinus i logit) s'assemblen molt.

El gràfic p composicional proposat té l'inconvenient de que no és aplicable a processos en els que la proporció a controlar és molt petita, cosa que també passa en els gràfics clàssics. En aquests casos es proposa canviar l'estadístic de control, i.e. controlar el temps entre fallades.

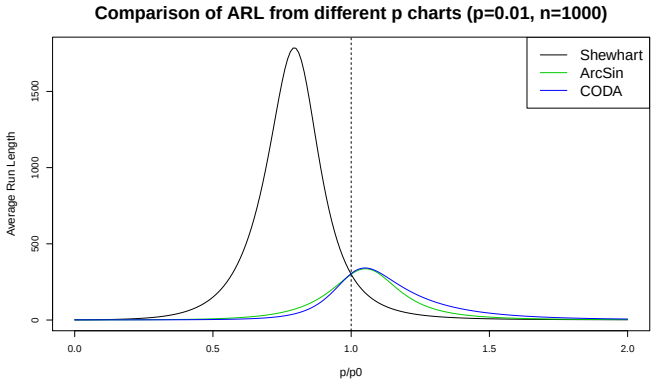


Figura 5.9: Corbes ARL de tres gràfics de control de proporcions: Shewhart, arcsin i composicional.

5.3 Gràfic CUSUM

Hem detectat un increment el l'ús d'aquest tipus de gràfics en aplicacions ambientals on són molt habituals les dades composicionals. Aquest ús ve motivat per la necessitat de detectar canvis petits però constants en el temps que es detecten amb eficiència amb el CC CUSUM.

Hem començat a treballar amb les dades de l'article Barratt et al. (2007) on es controla l'emissió de CO (mesurat en ppm) en un carrer de Londres abans i després de la implantació d'un carril bus. Els primers treballs indiquen que l'aplicació de la metodologia composicional al CC CUSUM, que consisteix en controlar el logit de les emissions, pot aportar un nou enfoc en aquest camp.

5.4 Conclusions

Tal i com hem vist al Capítol 1, hi ha un buit a la literatura pel que fa a mètodes per controlar processos on les característiques de qualitat són dades composicionals. En l'anàlisi de l'estat de l'art no hem trobat referències que tractin aquest tipus de dades utilitzant la metodologia composicional inicialment proposada Aitchison (1986) i que ha donat lloc a posteriori al desenvolupament de tot un seguit de tècniques que s'han demostrat adequades per tractar dades en espais restringits.

Hem iniciat la línia d'investigació tot adaptant l'esquema de control multivariant més antic i conegut: el gràfic T^2 de Hotelling. Som conscients que en els darrers anys s'ha avançat molt en les tècniques de control de processos, que ara tendeixen cap a mètodes més complexos i més adaptats als requeriments de la indústria. Són un exemple d'aquests els mètodes no paramètrics, que no requereixen cap coneixement ni assumpció sobre la distribució inicial

de les dades, o bé els mètodes *change-point*.

L'esquema proposat no aporta una innovació pel que fa al gràfic T^2 i de fet hereta els punts febles de l'esquema clàssic, que són la dependència de l'assumpció de normalitat de les dades, la necessitat d'una gran quantitat de dades a la fase I per estimar els paràmetres de la distribució o bé la baixa capacitat de detectar canvis petits en la mitjana del procés, entre d'altres.

No obstant, veiem el present treball com un primer pas que servirà per introduir els conceptes de la metodologia composicional en el camp del control estadístic de processos on, des del nostre punt de vista, no hi ha massa consciència sobre les característiques d'aquestes dades. Observem un cert interès en difondre aquests conceptes, cosa que es veu reflectida per la publicació del primer article en una revista d'alt impacte.

5.5 Futures línies d'investigació

Queda pendent proposar un gràfic de control T_C^2 per al control de mitjanes. En la mateixa línia, també es poden proposar altres estimadors per a la matriu de covariàncies o bé per a la mitjana de les coordenades, per exemple per fer que l'estadístic T^2 sigui més robust.

En l'àmbit de la quimiometria es detecta el creixement de propostes de control utilitzant el three-way data analysis, és a dir, representant les dades en matrius de tres dimensions que tenen en compte, per exemple, la localització, el temps i la mostra. En aquest sentit volem destacar que el treball de tesi vol continuar treballant no només en els mètodes clàssics sinó en les noves propostes per al control de processos.

D'altres línies encara per abordar són el disseny d'experiments, principalment el disseny de mixtures que per la seva natura són composicions. Estem segurs que la metodologia composicional aplicada en aquest camp pot aportar solucions innovadores. També queda pendent treballar més en profunditat el gràfic de control CUSUM i mostrar el seu potencial en aplicacions ambientals, així com les versions multivariants. D'altres gràfics per explorar són els EWMA univariants i multivariants.

Bibliografia

Acosta-Mejia, C. A. (1999). Improved p charts to monitor process quality. *IIE Transactions 31*(6), 509–516.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability (Reprinted 2003 with additional material by The Blackburn Press). London, UK: Chapman and Hall Ltd., 416 p.

Barceló-Vidal, C. (2011). De les dades composicionals a una geometria euclidiana sobre el símplex. *Butlletí de la Societat Catalana de Matemàtiques 26*(1), 5–28.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In *in Proceedings of IAMG'01. The sixth annual conference of the International Association for Mathematical Geology*, Cancun (México). CD-ROM, pp. 20.

Barratt, B., R. Atkinson, H. Ross Anderson, S. Beevers, F. Kelly, I. Mudway, and P. Wilkinson (2007). Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme. *Atmospheric Environment 41*(8), 1784–1791.

Bersimis, S., S. Psarakis, and J. Panaretos (2007). Multivariate statistical process control charts: an overview. *Quality and Reliability Engineering International 23*(5), 517–543.

Blyth, C. R. and H. A. Still (1983). Binomial Confidence Intervals. *Journal of the American Statistical Association 78*(381), 108–116.

Boyles, R. (1997). Using the chi-square statistic to monitor compositional process data. *Journal of Applied Statistics 24*(5), 589–602.

Chakraborti, S. and S. W. Human (2006). Parameter Estimation and Performance of the p -Chart for Attributes Data. *IEEE Transaction on Reliability 55*(3), 559–566.

Chakraborti, S., P. Van Der Laan, and S. T. Bakir (2001). Nonparametric control charts. *Journal of Quality Technology 33*(3), 304–315.

Champ, C. W., L. A. Jones-Farmer, and S. E. Rigdon (2005). Properties of the T^2 control chart when parameters are estimated. *Technometrics 47*(4), 437–445.

Chang, T. C. and F. F. Gan (2007). Modified Shewhart Charts for High Yield Processes. *Journal of Applied Statistics 34*(7), 857–877.

Chen, S. C. and J. N. Pan (2011). Determining Optimal Number of Samples for Constructing Multivariate Control Charts. *Communications in Statistics - Simulation and Computation 40*(2), 216–228.

Das, N. and V. Prakash (2007). Interpreting the out-of-control signal in multivariate control chart — a comparative study. *The International Journal of Advanced Manufacturing Technology 37*, 966–979.

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Simplicial geometry for compositional data. *Geological Society, London, Special Publications 264*, 145–159.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Fuchs, C. and R. S. Kenett (1998). *Multivariate quality control: theory and applications*. New York: Marcel Dekker INC.

Gonzalez-de la Parra, M. and P. Rodriguez-Loaiza (2003). Application of the Multivariate T^2 Control Chart and the Mason–Tracy–Young Decomposition Procedure to the Study of the Consistency of Impurity Profiles of Drug Substances. *Quality Engineering 16*(1), 127–142.

Hawkins, D. M. (1993). Regression Adjustment for Variables in Multivariate Quality Control. *Journal of Quality Technology 25*(3), 170–182.

Holmes, D. S. and A. E. Mergen (1993). Improving the performance of the T^2 control chart. *Quality Engineering 5*(4), 619–625.

Hotelling, H. (1947). Multivariate quality control. In C. Eisenhart, H. Hastay, and W. A. Wallis (Eds.), *Techniques of Statistical Analysis*, pp. 111–184. New York: McGraw-Hill.

Jackson, J. (1985). Multivariate quality control. *Communications in Statistics - Theory and Methods1 14*, 2657–2688.

Jensen, W. A., L. A. Jones-Farmer, C. W. Champ, and W. H. Woodall (2006). Effects of parameter estimation on control chart properties: a literature review. *Journal of Quality Technology 38*(4), 349–364.

Kenett, R. S. and M. Pollak (2011). On Assessing the Performance of Sequential Procedures for Detecting a Change. In *The 11th International ENBIS Conference*, Coimbra (Portugal).

Kenett, R. S. and S. Zacks (1998). *Modern industrial statistics: design and control of quality and reliability*. Duxbury Press Pacific Grove.

Kenett, R. S., S. Zacks, and D. Amberti (2014). *Modern Industrial Statistics: with applications in R, MINITAB and JMP*. Chichester, West Sussex: John Wiley & Sons.

Lowry, C. A. and D. Montgomery (1995). A review of multivariate control charts. *IIE Transactions 27*(6), 800–810.

MacGregor, J. and T. Kourti (1995). Statistical process control of multivariate processes. *Control Engineering Practice 3*(3), 403–414.

Martín-Fernández, J. A. (2001). *Medidas de diferencia y classificación automática no paramétrica de datos composicionales*. Ph. D. thesis, Universitat Politècnica de Catalunya.

Martín-Fernández, J. A., J. Palarea-Albaladejo, and R. A. Olea (2011). Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 47–62. Chichester (UK): John Wiley.

Mason, R., Y. M. Chou, and J. Young (2001). Applying hotelling ' s T^2 statistic to batch processes. *Journal of Quality Technology 33*(4), 466–479.

Mason, R., Y. M. Chou, and J. C. Young (2003). Ch. 15. Effective sample sizes for T^2 control charts. In R. Khattree and C. Rao (Eds.), *Handbook of Statistics 22: Statistics in Industry*, Volume 22, pp. 595 – 607. Elsevier.

Mason, R., N. Tracy, and C. Young (1995). Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology 27*(2), 99–108.

Mason, R., N. Tracy, and J. Young (1997). A practical approach for interpreting multivariate T^2 control chart signals. *Journal of Quality Technology 29*(4), 396–406.

Mason, R. and J. Young (2002). *Multivariate statistical process control with industrial applications* (1st ed.). Alexandria, Virginia: American Statistical Association and Society for Industrial and Applied Mathematics.

Mason, R. and J. Young (2006). Interpretation of multivariate control charts. *Enciclopedia of statistics in quality and reliability*, 1201–1206.

Mateu-Figueras, G. (2003). *Models de distribució sobre el símplex*. Ph. D. thesis, Universitat Politècnica de Catalunya.

McCool, J. I. and T. Joyner-Motley (1998). Control charts applicable when the fraction nonconforming is small. *Journal of Quality Technology 30*(3), 240–247.

Montgomery, D. (2009). *Statistical quality control: a modern introduction* (Sixth ed.). Missouri, USA: John Wiley & Sons.

Nelson, L. (1994). A control chart for parts-per-million nonconforming items. *Journal of Quality Technology 26*(3), 239–239.

Newcombe, R. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine 17*(8), 857–72.

Newcombe, R. (2001). Logit confidence intervals and the inverse sinh transformation. *The American Statistician 55*(3), 200–202.

Ortiz-Estarelles, O., Y. Martín-Biosca, M. J. Medina-Hernández, S. Sagrado, and E. Bonet-Domingo (2001). Multivariate data analysis of quality parameters in drinking water. *The Analyst 126*(1), 91–96.

Palarea-albaladejo, J., J. A. Martín-Fernández, and R. A. Olea (2014). Bootstrap estimation of distributional statistics from compositional data with nondetects: a case study on coal ashes. *Journal of Chemometrics (in press)*.

Pawlowsky-Glahn, V. and A. Buccianti (2011). *Compositional Data Analysis: Theory and Applications* (1st ed.). Chichester, UK: John Wiley.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment 15*(5), 384–398.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2010). Lecture Notes on Compositional Data Analysis.

Pearson, K. (1897). Mathematical contributions to the theory of evoluton. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London 60*, 489–502.

Pires, A. (1998). Confidence intervals for a binomial proportion: comparison of methods and software evaluation. In *In Klinke, S., Ahrend, P. and*

Richter L.(editors), Proceedings of the Conference CompStat 2002, pp. 24–28.

Quesenberry, C. (1991). SPC Q charts for a binomial parameter p: short or long runs. *Journal of Quality Technology 23*(3), 239–246.

Reynolds, M. R. and G. Y. Cho (2006). Multivariate control charts for monitoring the mean vector and covariance matrix. *Journal of Quality Technology 38*(3), 230–253.

Ryan, T. (2011). *Statistical methods for quality improvement* (3rd ed.). Hoboken, New Jersey.

Schader, M. and F. Schmid (1989). Two Rules of Thumb for the Approximation of the Binomial Distribution by the Normal Distribution. *The American Statistician 43*(1), 23.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Produc*. New York: Van Nostrand.

Stoumbos, Z., M. R. Reynolds, T. P. Ryan, and W. Woodall (2000). The State of Statistical Process Control as We Proceed into the 21st Century. *Journal of the American Statistical Association 95*(451), 992.

Sullivan, J. H. and W. H. Woodall (1996). A Comparison of Multivariate Control Charts for Individual Observations. *Journal of Quality Technology 28*(4), 398–408.

Topalidou, E. and S. Psarakis (2009). Review of multinomial and multiattribute quality control charts. *Quality and Reliability Engineering International 25*(7), 773–804.

Tracy, N., J. Young, and R. Mason (1992). Multivariate control charts for individual observations. *Journal of Quality Technology 24*(2), 88–95.

Vives-Mestres, M., J. Daunis-i Estadella, and J. A. Martín-Fernández (2014a). Individual T^2 Control Chart for Compositional Data. *Journal of Quality Technology 46*(2), 127–139.

Vives-Mestres, M., J. Daunis-i Estadella, and J. A. Martín-Fernández (2014b). Out-of-Control Signals in Three-Part Compositional T^2 Control Chart. *Quality and Reliability Engineering International 30*(3), 337–346.

Wang, H. (2009). Comparison of p control charts for low defective rate. *Computational Statistics and Data Analysis 53*(12), 4210–4220.

Woodall, W. (1997). Control charts based on attribute data : bibliography and review. *Journal of Quality Technology 29*(2), 172–183.

Woodall, W. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology 38*(2), 89–104.

Woodall, W. and D. Montgomery (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology 31*(4), 376–386.

Woodall WH. (2000). Controversies and contradictions in statistical process control. *Journal of Quality Technology 32*, 341–350.

Xie, L. and U. Kruger (2012). *Statistical Monitoring of Complex Multivariate Processes : With Applications in Industrial Process Control*. Singapore: John Wiley & Sons, Ltd.

Yang, G., D. Cline, R. Lytton, and D. Little (2004). Ternary and multivariate quality control charts of aggregate gradation for hot mix asphalt. *Journal of materials in civil engineering* (10), 28–34.