Handledare: Jonas Wallin
STAH11 Ht 2020 Kandidatuppsats 15hp

LUNDS UNIVERSITET

Klusteranalys och bortfall: en studie i hur klusteranalys påverkas av imputation för variabelbortfall

Filip Berndtsson, Viktor Segersäll

Lunds universitet

Statistiska institutionen

2021

Abstract

The combination of item non response and cluster analysis is a field which has not been explored to its full extent. This study aims to investigate which consequences a data set with missing values has on the algorithm of cluster analysis and which model for imputation should be suggested in order to minimize standard error and the optimal number of clusters. Basing our research on principal component analysis, k-mean clustering and the average silhouette model, we intend to examine the depth of cluster analysis and how it handles missing values.

Firstly, we have performed cluster analysis on three different datasets, applying the average silhouette model to estimate the number of clusters. Secondly, we have used three different techniques for imputation of missing values: imputation with the sample mean, imputation with the sample median and imputation with MICE.

Thirdly, we have estimated the standard error for each imputation technique as a unit of measurement in order to draw conclusions about which imputation technique minimizes the errors and hence could be the optimal imputation technique for a cluster analysis.

Our results, based on the least estimate of the standard error, is that the MICE imputation tends to generate the best estimate compared to a dataset without missing values. A more simple imputation technique, such as the sample mean or median, could nonetheless be considered if the assumptions for the MICE-technique are not fulfilled or the data set is small.

1. Inledning

Kategorisering av data är ett ämne inom statistik som ständigt undersöks och nya sätt att kategorisera dyker ständigt upp. Klusteranalys är ett tillvägagångssätt för kategorisering och dataanalys och metoden används i allt större utsträckning. Statistiska undersökningar utmanas alltmer av svarsbortfall och välbyggda modeller för imputering efterfrågas till en högre grad.

Ett område som inte har utforskats i lika stor utsträckning är kombinationen av klusteranalys och ett datamaterial med variabelbortfall. Denna studie ämnar undersöka vilka konsekvenser ett datamaterial med bortfall har på en klusteranalys och vilken metod för imputering som är mest lämplig när man ska hitta rätt antal kluster och efterlikna ett datamaterial utan bortfall. Utifrån teorier om skattning av antal kluster och imputering kommer vi att undersöka tre datamaterial och försöka dra slutsatser om vilken eller vilka metoder som är lämpligast att använda sig av när man applicerar klusteranalys på ett datamaterial innehållande bortfall.

2. Frågeställning och avgränsning

Den så kallade klusteranalysen är ett verktyg hämtat ur den multivariata analysen, som ett sätt att gruppera data baserat på algoritmer för distanser mellan datapunkter och dimensioner. Denna metod kategoriserar data baserade på variablers värden och variation. Klusteranalysen har potential att utmana etablerad kategorisering av data och vi ämnar därmed att undersöka metodens potential genom att analysera tre datamaterial med hjälp av algoritmer för att hitta optimalt antal kluster och sedan klustra materialet utifrån det euklidanska distansmåttet.

Vi tar en teoretisk ansats när vi undersöker klusteranalysen för att undersöka styrkor och svagheter hos metoden och hur känslig metoden är för hantering av bortfall.

En utmaning för all modern dataanalys är hantering av bortfall, det vill säga en undersökning där individer, variabelvärden eller enbart vissa svar har fallit bort av olika anledningar. Vi undersöker olika metoder för att hantera dessa bortfall och vilka konsekvenser det får gentemot en klusteranalys på samma datamaterial utan bortfall. Påverkas algoritmens generering optimalt antal kluster för vilken imputeringsteknik som appliceras? Kommer detta att ge en inverkan för kategoriseringen och kan en imputationsmetod vara mer lämpad än andra vid bortfall? Vilken metod för imputation är bäst lämpad vid klusteranalys?

Sammanfattningsvis är målet med vårt arbete att undersöka hur känslig klusteranalysen är för bortfall och hur den reagerar på imputation.

3. Metod

För att undersöka hur bortfall och olika metoder för imputation påverkar en klusteranalys har vi använt oss av tre olika datamaterial: ett som innehåller information om viner, ett som undersöker brottslighet i USA och ett som studerar olika släkter av grodor. Då vårt syfte inte är analys av data, utan en undersökning av olika metoder för hantering av bortfall är inte de resultat dataanalysen kommer fram till intressanta i sig. Det vi ämnar undersöka är vilka kluster som algoritmerna skapar inledningsvis och hur olika metoder för imputation som efterliknar de ursprungliga kluster vi skapat. Vi har valt våra dataset utifrån principen att de bör ha olika stickprovsstorlekar och att de ska ha tillräckligt många observationer för att vi ska kunna anta approximativ normalfördelning. Tillvägagångssättet för varje dataset finns beskrivet under

“Resultat”, där varje dataset undersöks var för sig under “Experiment 1”, “Experiment 2” och “Experiment 3”.

I varje experiment delar vi upp respektive dataset i två delar: ett träningsdataset och ett test-dataset. Vår uppdelning är att 70 procent av observationerna ligger i träningsdatan och 30 procent i test-datan. Detta är en standarduppdelning i tränings-och test-dataset som används för att reducera risken för överanpassning av parametrar i träningsdatasetet. (Liu, H. & Cocea, M.,2017, 358)

I träningsdatasetet har vi inget bortfall, utan klustrar alla observationer utifrån euklidanska distansmått. Ett av de grundläggande problemen med klusteranalys är hur man väljer antal kluster i sin analys. Vi har åtgärdat denna problematik genom först testa hur datamaterialet beter sig med 2, 3 och 4 kluster och sedan skatta antalet kluster med hjälp av medelsilhuett-metoden. På så sätt har vi testat olika antal kluster för att göra en bedömning vilket som är det optimala i vårt fall. Tibshirani et al. utförde en studie där de undersökte en global metod för uppskattning av kluster och begränsade den teoretiska diskussionen till k-mean-klustring för att diskutera utfallet och lämpligheten i metoden. (Tibshirani et. al, 2000, 411) Vi valde att använda deras begränsning för att kunna utnyttja deras resultat men även dra generella slutsatser om val av antal kluster. Detta är en möjlig felkälla som vi återkommer till i vår diskussion.

I det här steget är träningsdatasetet uppdelad i kluster och varje kluster har ett medelvärde för varje variabel den är klustrad utifrån. Det är dessa medelvärden vi utgår ifrån i det senare skedet när vi jämför resultaten från de olika imputationsteknikerna i test-dataseten.

I test-datasetet generar vi tio procent bortfall genom att slumpmässigt radera tio procent av värdena för olika variabler. Vi valde tio procent bortfall då det finns ett mönster att större undersökningar behöver räkna med ett allt större bortfall, till exempel SCB:s Arbetskraftsundersökningar har i genomsnitt gått från under fem procent bortfall till över 16 procent. (Dahmström, 2011, 323)

När bortfallet genererats imputerar vi det i test-datasetet tre gånger med tre olika metoder: imputation med stickprovsmedelvärdet, med stickprovsmedianen och med MICE. MICE är en imputeringsteknik som skattar variabelbortfall baserat på cykler av predikterande regressionsskattningar utifrån gemensamma datavärden och kommer att diskuteras mer utförligt i sektion 4.3. Efter imputationerna testar vi att klustra utifrån samma principer som vi klustrat träningsdatan. Vi har därmed tre test-dataset, ett för varje imputationsmetod. Då vi är intresserade att ta reda på vilken metod för imputation som är bäst lämpad för klusteranalys jämför vi hur väl test-datasetens klustring efterliknar träningsdatasetets klustring genom att jämföra medelkvadratfelet utifrån de medelvärden som test-datan fått fram mot de medelvärden som träningsdatan genererar. Medelkvadratfelet används som mått därför att det är ett mått på det totala felet i skattningen och att det konventionellt går att välja den skattning med lägst medelfel (Körner & Wahlgren, 2015, 154). Det ska tilläggas att slumpen kan medföra att den teoretiskt bästa skattningen inte är den optimala, så att medelkvadratfelet inte är ett lämpligt mått. Vi har därmed gjort om generering av bortfall och imputering efter klustringen, för att undvika systematiska fel(bias) i skattningen av medelkvadratfelet. Baserat på resultatet av den första imputeringens medelkvadratfel har vi använt den metod som gav bäst resultat för att göra ytterligare en imputering och jämfört denna med originalmaterialet. Vi har därmed försökt åtgärda problemet med systematiskt fel genom att undersöka flera datamaterial av olika storlekar, men vi tar hänsyn till att medelkvadratfelen kan vara missvisande speciellt när de är snarlika. Vi drar därefter slutsatser om vilken imputationmetod som är bäst lämpad för klusteranalys baserat på hur stort medelkvadratfelet och hur denna metod reagerar på en

jämförelse med originaldatamaterialet.

Denna procedur återupprepas för varje imputationsmetod och vi avslutar med en diskussion om vilka slutsatser vi kan dra om vilken imputationsmetod som är bäst lämpad samt möjliga felkällor i vår studie.

4. Teori

Nedan presenteras de teorier som är relevanta för vår studie. Då de huvudsakliga modellerna vi intresserar oss av är klusteranalys och hantering av bortfall är dessa huvudrubriker med underrubriker för specifika grenar av teorierna.

För att beskriva klusteranalys har vi delat upp arbetet i en avdelning där vi beskriver vilken notation vi konsekvent kommer använda oss av under uppsatsen. Vi har sedan beskrivit det distansmått vi använder oss av, principiell komponentanalys samt en förklaring av k-mean-klustring. Vi avslutar delkapitlet med en föklaring av medelsilhuettmetoden, den vi utgått ifrån när vi skattat antal kluster.

När vi sedan beskriver teorier om bortfall börjar vi med översiktlig beskrivning av hur bortfall hanteras samt en beskrivning av de metoder för imputation vi använder oss av i våra experiment. Slutligen beskriver vi medelkvadratfelet och motiverar varför det är lämpligt som jämförelsemått.

4.1. Notation och matematisk bakgrund

d mått på distansers längd
p mått på antal variabeldimensioner
n mått på antal stickprovsobservationer
k mått på antal kluster
$r\ (i), q\ (i)$mått för hur väl en observation passar in i ett kluster utifrån medelsilhuettmetoden

$\underline{X_i}$ notation för vektorobservation nummer *i,* med p dimensioner, i ett stickprov
Z notation för principalkomponenter
a, b notation för konstanter
λ notation för egenvärden
c notation för kovarianser
S notation för grupper i klusteranalys
$s(i)$ notation för silhuettvärde

Kovariansmatrisen för variablerna $\underline{X} = (x_1, x_2, \ldots, x_p)$ betecknas

$$\boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & .. & c_{1p} \\ c_{21} & c_{22} & .. & c_{2p} \\ . & . & . & . \\ c_{p1} & c_{p2} & .. & c_{pp} \end{bmatrix}$$

4.2. Klusteranalys

För att få utföra en klusteranalys krävs det att samtliga datanivåer är numeriska och grunden bygger på att det finns *n* observationer med värden på *p* variabler. Vi inför därmed notationen $\underline{X_i} = (x_{i1}, \ldots, x_{ip})$ som en vektor för samtliga observationer i ett stickprov i *p* dimensioner. Vi antar att antalet grupper vi ämnar att skapa är okänt och att algoritmerna i sig själva skapar det antal som är optimalt enligt modellen. (Manly & Navarro Alberto, 2017, s.163)

Euclides distansmått

Ett centralt begrepp för klusteranalys är multivariata distanser och hur de mäts. Vi kommer nedan att presentera det euklidanska distansmåttet som vi använt oss av och hur det bör tolkas.

Grunden för idén om multivariata distansmått är att på något sätt mäta distanser mellan observationer i ett stickprov, mellan hela stickprov i sig eller mellan populationerna från vilken ett stickprov är draget från. Det enklaste fallet när två observationer, $\underline{X_j} = (x_{j1}, x_{j2})$ och $\underline{X_k} = (x_{k1}, x_{k2})$, befinner sig i en tvådimensionell rymd där antalet variabler, *p,* är två. Dessa observationers värden går därmed att skriva som $\underline{X_j}$ samt $\underline{X_k}$, där *j* och *k* är index som definierar de olika observationerna. Sedan mäter man enligt pythagoras sats enbart avståndet mellan observationerna på följande sätt:

$$d_{ij} = \{(x_{j1} - x_{k1})^2 + (x_{j2} - x_{k2})^2\}^{1/2}.$$

När vi istället har tre variabler, det vill säga när *p=3,* mäts distanserna på följande vis:

$$d_{ij} = \{(x_{j1} - x_{k1})^2 + (x_{j2} - x_{k2})^2 + (x_{j3} - x_{k3})^2\}^{1/2}.$$

När sedan *p>3* upprepas proceduren, och den matematiskt generella formeln för *p* ser ut på följande vis:

$$d_{jk} = \sum_{k=1}^{p} (x_{jl} - x_{kl})^2 \quad (4.2.1)$$

(Manly & Navarro Alberto, 2017, s. 84)

När en klusteranalys bygger på detta distansmått är variablerna ofta standardiserade så att de *p* variablerna blir lika viktiga när klustrena skapas. Ett problem som uppstår när man standardiserar värdena är att man riskerar att minska variationen mellan grupper vilket i sig är en möjlig grund för ett kluster. (Manly & Navarro Alberto, 2017, 168) Det är därmed lämpligt att utföra en principalkomponentanalys för att undersöka variationen i datamaterialet.

Principalkomponentanalys

Principalkomponentanalys är ett verktyg inom den multivariata analysen som bygger på ett datamaterial med *p* variabler. Vi introducerar vektorn $\underline{X} = (x_1, x_2, \ldots, x_p)$. Det gäller sedan att hitta icke-korrelerade kombinationer av dessa som vi kallar $Z_1, Z_2, \ldots, Z_p$ som är ordnade efter dess relevans för variationen i datamaterialet. Detta innebär att $var[Z_1] \geq var[Z_2] \ldots \geq$

$var[Z_p]$utefter deras varians. Dessa värden Z_i är därmed våra principalkomponenter. (Manly & Navarro Alberto, 2017, 103)

Det som avgör vilken ordning komponenterna får beror på vilken linjärkombination av värdena på $(X_1, X_2, \ldots, X_p)$ som ger en kombination av med högst varians. Den första principalkomponenten kommer därmed få värdet

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \qquad (4.2.2)$$

där värdena på konstanterna *a* måste uppfylla restriktionen:

$$a^2_{11} + a^2_{12} + \ldots + a^2_{1p} = 1.$$

Denna restriktion krävs för att maximera variansen inom Z_1så att den uppfyller olikheten $var[Z_1] \geq var[Z_2] \ldots \geq var[Z_p]$.

Den andra principalkomponenten kommer att skattas genom

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

där värdena på konstanterna *a* uppfyller restriktionen

$$a^2_{21} + a^2_{22} + \ldots + a^2_{2p} = 1.$$

Samma process fortsätter för $Z_3, Z_4, \ldots$ ända upp till Z_p. Vi kan därmed definiera en ny vektor för principalkomponenterna, $\underline{Z} = (Z_1, \ldots, Z_p)$. (Manly & Navarro Alberto, 2017, 105)

För att förstå på vilket sätt PCA är relevant för klusteranalys och i vår studie specifikt är det viktigt att beskriva dessa funktioners grundläggande egenskaper utifrån egenvärdena från stickprovets kovariansmatris. Utifrån kovariansmatrisen kan vi se att diagonalerna är respektive varians för variablerna $X_1, X_2, \ldots, X_p$. Termerna som motsvarar $c_{ij} = c_{ji}$ är istället kovarianserna mellan variablerna $X_i$och X_j. Egenvärdena, λ_i, i kovariansmatrisen **C** kommer att motsvara varianserna för $Z_1, Z_2, \ldots, Z_p$ så att

$$\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0.$$

Egenvärdena kommer alltså att motsvara varianserna i principalkomponenterna och konstanterna $a_{i1}, a_{i2}, \ldots, a_{ip}$ är i sin tur de motsvarande egenvektorerna med restriktionen

$$a^2_{i1} + a^2_{i2} + \ldots + a^2_{ip} = 1.$$

En egenskap hos egenvärdena är att summan av dem motsvarar summan av varianserna för$X_1, X_2, \ldots, X_p$, alltså summan av diagonalelementen i kovariansmatrisen **C**. Detta kan skrivas som

$$\lambda_1 + \lambda_2 + \ldots + \lambda_p = c_{11} + c_{22} + \ldots + c_{pp} .$$

(Manly & Navarro Alberto, 2017, 106)

Då summan av egenvärdena, λ_i, är detsamma som summan av varianserna i principalkomponenterna, $var[Z_i]$, argumenterar Manly och Navarro Alberto för att principalkomponenterna kan innehålla all information om variationen i ursprungsmaterialet. (Manly & Navarro Alberto, 2017, 106)

Många algoritmer för klusteranalys möjliggör en nedskalning av variabler till färre principalkomponenter. Detta påverkar analysen och de kluster som skapas och kan vara en nackdel såväl som en fördel. Nackdelen är att principalkomponenterna riskerar att ta över och radera ut mindre starka samband som kan vara lika viktiga som principalkomponenterna i skapandet av kluster. Fördelen är ifall principalkomponenterna står för en majoritet av variationen i datamaterialet kan de användas som grund för att jämföra individer mot den och därmed upptäcka kluster med hjälp av det. (Manly & Navarro Alberto, 2017, 168)

K-means-klustring

K-means-klustring matematiska definition beskrivs på följande vis:

Givet ett datamaterial med *n* observationer$\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ där varje observation är en vektor med *p* dimensioner ämnar metoden skapa *k* stycken grupper där $k \leq n$, det vill säga att klustrena ska vara färre än eller lika med antalet observationer. Dessa grupper(“kluster”) betecknas med S och de *k* grupperna $S_1, S_2, \ldots, S_k$ är sådana att varje observationsvektor tillhör en, och endast en, grupp. Målet är att dela upp dessa delmängder för att minimera kvadratsummorna inom klustrena. Denna process kan beskrivas som :

$$argmin_{S_1,\ldots,S_k}(\sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2) \qquad (4.2.3)$$

där μ_i definieras som

$$\mu_i = \sum_{x \in S_i}^{k} \frac{x}{|S_i|}$$

där$|S_i|$betecknas antal element (observationer) i grupp S_i. Detta sätt att uttrycka medelvärdena i de olika klustren belyser komplexiteten i att hitta k stycken grupper, $S_1, \ldots, S_k$, bestående av observationer bland totalt *n* observationer, och sådana att de minimerar den totala (inom-grupps)variansen. Antal möjliga kombinationer för dessa grupper är $(n-1)! = (k-1)!\,(n-k)!$. Redan vid ett datamaterial med över *n*=51 och *k*=5 finns det 230 300 möjliga gruppbildningar, och i realiteten använder man en sökfunktion som estimerar vilka av dessa gruppbildningar som minimerar ekvationen (4.2.3). Ett alternativt sätt att uttrycka det är att minimera den parvisa kvadratiska avvikelsen mellan observationer i samma kluster. Då den totala variationen är konstant så innebär att minimera kvadratsummorna inom klustrena samma sak som att maximera variansen mellan klustrena. (Kriegel et al., 2016, 341-350)

En alternativ beskrivning av k-means klustring är att man minimerar variationen inom grupperna via skattningar av kvadratsummor. Prefixet *k* innebär att man på förhand bestämmer hur många kluster man vill att algoritmerna ska skapa. Anta att vi har ett stickprov med *n* observationer. Algoritmen kommer därefter skapa *k* stycken kluster och mäta distanserna mellan de verkliga observationerna, x_i, och våra kluster. Observationerna kommer att allokeras till det kluster som ligger närmast utifrån valt distansmått. Vi beräknar sedan medelvärdet i de olika klustren och utifrån medelvärdena mäter vi på nytt distansen mellan dessa och de *k* klustren. Då denna metod inte tar hänsyn till variationen i stickprovet kommer den totala variationen mätas för de *k* klustren som skapats. Algoritmen kommer därefter skapa *k* nya kluster som är placerade på andra ställen jämfört med den första klustringen och genomgå samma process. Vi kommer därmed att upprepa proceduren tills alla möjliga kombinationer av observationerna ingått i samtliga kluster och använda oss av den där variansen är jämnt fördelad. (Fraley, 1998)

Detta väcker frågan, hur stort bör *k* vara? Det är uppe för diskussion, men en metod för att undersöka detta är att testa olika värden på *k* och därefter se var den totala variationen minimeras. Vi kommer i vårt resultat undersöka vårt material genom att testa olika värden på *k* och se vilka utfall vi får beroende på det värdet. Vi kommer även att skatta antalet kluster med hjälp av medelsilhuett-metoden som är beskriven nedan.

Medelsilhuett-metoden

Den matematiska definitionen för medelsilhuett-metoden lyder på följande sätt:

Om vi antar att vi klustrat ett datamaterial med hjälp av k-means klustring i *k* stycken kluster och observation *i* tillhör grupp S_l så låter vi

$$r(i) = \frac{1}{|S_l| - 1} \sum_{j \in S_l, j \neq i} d\,(i,j)$$

vara den genomsnittliga distansen mellan observation *i* och övriga datapunkter i gruppen. Här är $d\,(i,j)$ distansen mellan observation *i* och *j* i kluster S_l. Därmed kan $r(i)$ ses som ett mått på hur väl observation *i* tillhör sitt kluster S_l. Här är *l* en funktion av *i, l(i)*, och ett resultat av klustringen. Vi definierar sedan den genomsnittliga olikheten i mellan observation *i* och ett annat kluster, S_g, som distansen mellan observation *i* och alla andra observationer i kluster S_g. Vi kan nu införa ett nytt mått på hur väl observation *i* tillhör ett annat kluster, S_g på följande vis: för en observation *i* som tillhör kluster S_l är

$$q(i) = min_{k \neq i} \frac{1}{|S_g|} \sum_{j \in S_g} d\,(i,j)$$

det vill säga den minsta distansen mellan *i* och alla närliggande kluster som *i* ej tillhör. Detta ses därmed som det "nästbästa" klustret som *i* skulle kunna tillhöra. Vi kan utifrån detta definiera ett *silhuettvärde*, *s(i),* på följande vis:

$$s\,(i) = \frac{q\,(i) - r\,(i)}{max\{r\,(i), q\,(i)\}} \text{ om } |S_l| > 1 \text{ och } s\,(i) = 0 \text{ om } |S_l| = 1. \qquad (4.2.4)$$

Silhuettvärdet kommer i och med denna definition att ligga i ett intervall mellan $-1 \leq s\,(i) \leq 1$. Låga värden på distansen $r(i)$ implicerar att observation *i* ej passar bra in i sitt kluster S_i medan ett stort värde på $q(i)$ implicerar att observation *i* ej är lämpligt i sitt "nästbästa" kluster.

Medelsilhuett-metoden är ett mått som hjälper oss att bestämma vilket det bästa antalet kluster är för en ett visst dataset. Silhuett-spridningsdiagrammet visar hur nära en punkt i ett visst kluster förhåller sig till andra punkter inom samma kluster. Ett högt värde indikerar att en observation är långt ifrån andra kluster och ett lågt värde att observationen ligger nära brytvärdet för att hamna i ett annat kluster. En silhuett är alltså ett mått på avståndet mellan övriga observationer i ett hypotetiskt kluster. Algoritmen testar olika antal kluster och om silhuetterna är smala kan det indikera på ett dåligt val av antal kluster. Ett spridningsdiagram av silhuetterna för olika antal kluster kan därmed vara ett sätt att skatta ett optimalt antal utifrån detta distansmått. (Rousseeuw,1987, 55-56)

Problem med klusteranalys

Ett grundläggande problem med klusteranalys är att det inte finns en väldefinierad definition av vad ett kluster är. Då kategorisering av data traditionellt sett inte nödvändigtvis bygger på relationer mellan variablers värden, utan snarare är en uppdelning av praktiska skäl, kan klusteranalysens uppdelning av kluster vara upplagda för diskussion. Denna problematik är inte unik för k-means-klustring utan är något som samtliga metoder utmanas av. (Tibshirani *et al*., 2000, 419-420)

Tibshirani diskuterar olika metoder för estimering av antal kluster och refererar till Gordons uppdelning i globala och lokala metoder, där globala metoder innebär att utvärdera vissa mått på hela datamaterialet och sedan optimera det som en funktion av antal kluster. De lokala metoderna kollar istället på individuella par av kluster och testar huruvida de borde slås samman. Ett problem med de globala metoderna är att de inte säger något om huruvida datamaterialet ens är lämpligt för klusteranalys. (Tibshirani *et al.*, 2000, 416)

Är de kluster som algoritmerna skapar en bra kategorisering eller riskerar de att bygga på slump? Denna frågeställning återkommer vi till i en diskussion utifrån våra egna test.

4.3. Bortfall

Nedan följer en beskrivning av imputationstekniker, dess för-och nackdelar och lite mer utförlig diskussion om de tekniker vi valt att använda oss av. Avsnittet är mindre utförligt eftersom undersökning av imputationsteknik inte är vårt huvudsakliga mål, utan ett hjälpmedel för att undersöka hur klusteranalys påverkas av bortfall och imputation.

Sharon Lohr beskriver i sin bok “Sampling, design and analysis” från 2010 fyra sätt för hur man kan hantera bortfall:

1. Undvika det i skapandet av en undersökning
2. Ta ett stickprov av de bortfallna svaren och använd det för att via inferens skapa en ny skattning av parametrarna i bortfallet
3. Hitta en modell för att skatta bortfallet, imputation
4. Ignorera bortfallet

(Lohr, 2010, 330).

Då vårt syfte är att undersöka hur en klusteranalys påverkas av bortfall och hur olika metoder för att hantera det påverkar analysen kommer vi främst att undersöka modeller för att skatta bortfall; så kallad imputation.

Enligt Lohr finns det olika metoder för imputation, det vill säga metoder för att skatta bortfall. Vi förklarar här de metoder som vi kommer att använda oss av när vi gör imputationer i vårt arbete.

Imputation med stickprovsmedelvärdet och stickprovsmedianen
I detta fallet ersätter vi bortfallen med stickprovets medelvärde. Denna imputationsmetod är effektiv då då stickprovets medelvärde i ett tillräckligt stort stickprov är en väntevärdesriktig skattning av populationsmedelvärdet enligt "Stora talens lag". (Gut, 2009, 161) Sådana skattningar kan därmed innehålla tillräckligt med information för att imputera bortfallen. Problemet med denna metod är att när man ersätter bortfall med det övriga stickprovets medelvärde är att man går miste om den spridning från medelvärdet som bortfallet skulle ha medfört.(Lohr, 2010, 348) Samma princip gäller för imputation med stickprovsmedianen, som kan vara ett komplement då den är mindre känslig för spridning.

Imputation med MICE
Mice är ett R-programpaket som används för att imputera saknade data i dataset. Tillvägagångssättet ser ut som följande: först imputeras ett temporärt värde för det saknade värdet för en observation med en enkel metod såsom att ersätta dessa med medelvärdet. Dessa temporära värden för en variabel skattas sedan tillbaka till saknade värden, varpå en regressionsskattning görs på denna variabel med hjälp av de andra variablernas värden i datasetet. De saknade värdena i denna variabel ersätts sen med de värdena som imputeras från regressionen. Denna process upprepas sen för de andra variabler där vi har saknade värden tills allt bortfall är imputerat. Denna process är en form av cykel där antalet cykler kan bestämmas i paketet och vi använder oss utav fem cykler vilket är standard. Efter dessa cyklar körts har vi våra slutgiltliga imputationer. (van Buuren & Groothuis-Oudshoorn, 2011)

4.4 Jämförelsemått

Vi har valt att jämföra utfallen från de olika imputeringsmetoderna med hjälp av medelkvadratfelet som skattningarna ger upphov till utifrån träningsdatamaterialets klustermedelvärden. Medelkvadratfelet är ett mått på det totala felet i en skattning, och vi använder oss av Körners och Wahlgrens formulering: "en generell regel kan nu formuleras så här. Av alla möjliga skattningar ska vi välja den skattning vars medelkvadratfel är så litet som möjligt." De diskuterar vidare att den regeln inte alltid är tillämpbar, utan att det måttet kan vara beroende till viss del av slump så att en teoretiskt sämre skattning kan generera ett bättre resultat. (Körner & Wahlgren, 2015, 154)

Vår tillämpning av medelkvadratfelet som jämförelsemått innebär att för den klustrade träningsdatamaterialet hittar vi de olika medelvärden i alla kluster för samtliga variabler. Samma sak görs sedan för testdatamaterialet, för varje imputationsteknik. Differenserna mellan träningsmaterialets medelvärde och testmaterialens medelvärde används sedan för att beräkna medelkvadratfelet. Detta använder vi i sin tur som ett mått för hur bra de olika imputeringsmetoderna är, det vill säga den imputeringsmetod där det skattade medelkvadratfelet är lägst.

5. Resultat

Alla dataset går att hitta i uppsatsens appendix. Vi använder oss av programmet R och paketen Tidyverse , MICE, VIM, cluster, Factorextra, NBclust och gridExtra för att utföra klustringen. Koden som används finns också tillgänglig för den intresserade läsaren. Samtliga tabeller för medelvärden i test-datamasetet, samt test-dataseten med de olika imputationsteknikerna är tillgängliga i uppsatsens appendix.

5.1. Experiment 1: Vin

Beskrivning av datamaterialet och uppdelning i träning- och test-set

Detta dataset behandlar olika viner och består av 178 observationer med 13 variabler där samtliga variabler är numeriska och är mått på bland annat alkoholhalt och hur vinet lagrats. Vi delar sedan upp datamaterialet där träningsdatasetet består av 124 observationer och testdatasetet av 54 observationer. Vi skalar sedan värdena utifrån en principalkomponent-analys.

Klustring av träningsdatasetet

Vi använder oss av K-means-klustring för att placera ut observationerna i rätt kluster. Vi väljer först antal kluster genom att testa olika värden på *k:* 2, 3, 4 och 5. Vi använder sedan medelsilhuett-metoden för att hitta vårt antal kluster för att se vilket antal som är optimalt från 1 till 15 kluster. Algoritmen skattar att 3 kluster är optimalt för detta dataset. Vi utför klusteranalysen och får fram medelvärden för varje variabel för varje kluster. Dessa går att finna i appendix A.1.1. och används som utgångspunkt när vi jämför medelkvadratfelen i test-datan för de olika imputeringsteknikerna.

Imputering: MICE

Vi arbetar nu med test-datan och genererar slumpmässigt tio procent bortfall och imputerar detta bortfall med hjälp av MICE. Vi återupprepar sedan den tidigare processen och enligt medelsilhuett-metoden finner vi även här 3 kluster som det optimala. Vi beräknar sedan medelvärdena för respektive kluster och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel: 24.4919

Imputering: stickprovsmedelvärde

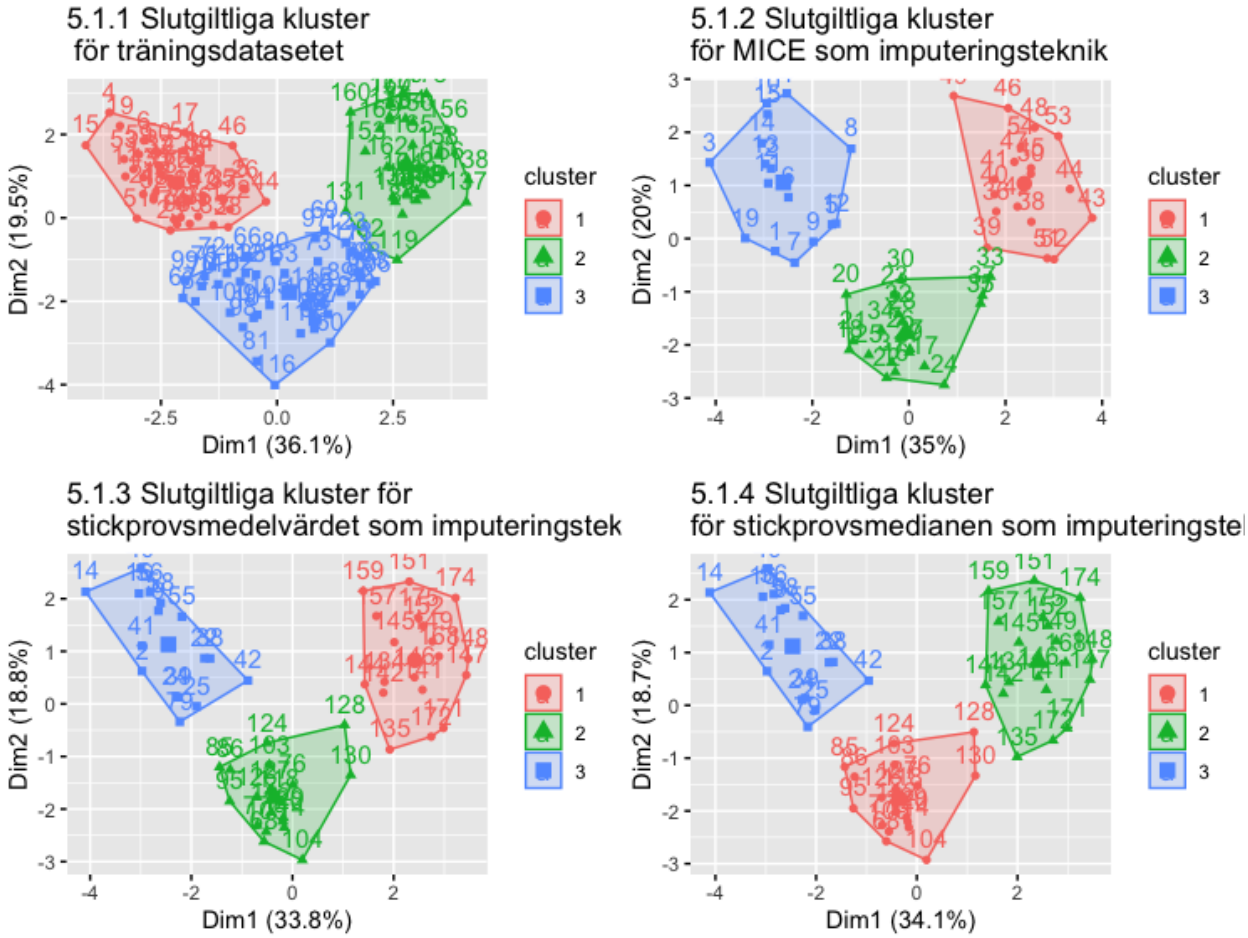
Vi återanvänder test-datan och genererar bort samma bortfall som tidigare för att inte genereringen i sig ska vara en felkälla. Vi utför samma process som tidigare och finner att 3 kluster är optimalt. Vi imputerar sedan det bortfall varje observation har med medelvärdet för det kluster som observationen befinner sig i. Vi beräknar sedan medelvärdena för respektive kluster

och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel: 26.01713

Imputering: stickprovsmedian

Vi återanvänder testdatan, och genererar samma bortfall som vi fick i föregående test. Vi får 3 som det optimala antalet kluster utifrån medelsilhuett-metoden. Vi imputerar därefter bortfallen med medianen för varje variabel i det kluster som observationen befinner sig i. Vi beräknar på nytt medelvärdena för varje kluster och skattar utifrån dessa medelkvadratfelet gentemot träningsdatans medelvärden och får följande medelkvadratfel: 25.92102

Slutgiltiga kluster för Vin-datamaterialet



Figur 5.1. Slutgiltiga kluster för träningsdata och testdata i vin-datamaterialet

I figur 5.1 presenteras de slutgiltiga klustren för träningsdatamaterialet, samt kluster för hur väl imputeringsteknikerna i de olika test-dataseten lyckades replikera träningsdatasetets klustring.

Jämförelse av medelkvadratfelet

	Medelkvadratfel
Imputering med MICE	*24.4919*
Imputering med stickprovsmedelvärdet	*26.01713*
Imputering med stickprovsmedianen	*25.92102*

Tabell 5.1. Medelkvadratfel skattade utifrån träningsdatamaterialet för varje imputeringsteknik i datamaterialet Vin

Som kan utläsas ur tabell 5.1 är medelkvadratfelet lägst för imputering med MICE. Utifrån detta resultat imputerar vi testdatamaterialet för MICE en gång till och jämför det med originaldatamaterialet vilket genererar ett medelkvadratfel på 0.0009205545.

5.2. Experiment 2: CDI

Beskrivning av dataset och uppdelning i träningsdata och testdata

Detta dataset är en undersökning av brottsstatistik i olika områden i USA där man tagit hänsyn till bland annat inkomst, utbildning och arbetslöshet. Datamaterialet består av 440 observationer med 17 variabler. Vi delar upp datamaterialet i ett träningsdataset som sedan består av 308 observationer och testdatasetet består av 132 observationer. Vi skalar sedan värdena och utför en klusteranalys.

Klustring av träningsdatasetet

Vi testar först att klustra med 2, 3 och 4 kluster för att se vilka utslag vi får. Vi använder sedan medelsilhuett-metoden som ger $k = 2$ som det optimala antalet kluster. Medelvärden för respektive kluster beräknas och finns tillgängliga i appendix, tabell A.2.1. Det är medelvärdena från detta träningsdataset som kommer att användas som utgångspunkt när vi jämför medelkvadratfelen från imputationsteknikerna.

Imputering: MICE

Vi använder oss nu av test-datasetet och genererar tio procents bortfall som vi sedan imputerar med hjälp av MICE. Sedan upprepar vi processen för test-datasetet och får ut $k=2$ som optimalt antal kluster. Vi beräknar sedan medelvärdena för respektive kluster och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel: 56.09514

Imputering: stickprovsmedelvärde

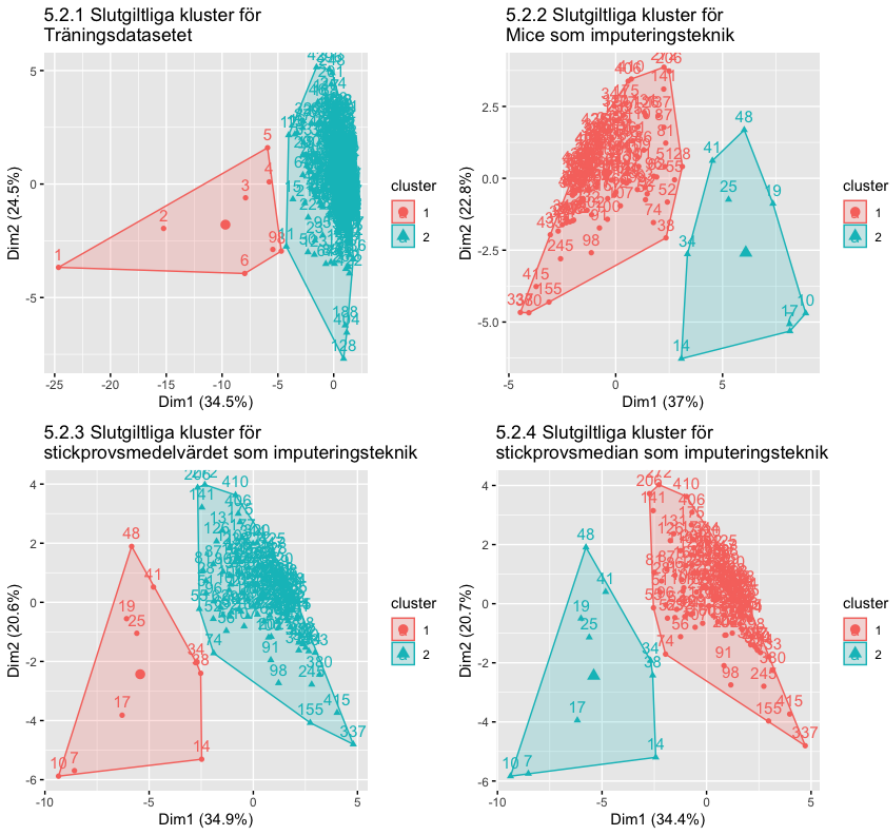
Utifrån test-datasetet imputerar vi det variabelbortfall vi genererat med hjälp av stickprovsmedelvärdet för det kluster som observationen placerats i och testar med hjälp av average-silhouett-metoden att 2 kluster är det optimala. Vi beräknar sedan medelvärdena för

respektive kluster och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel: 139.1653

Imputering: median

Vi imputerar sedan de variabelbortfallen i test-datasetet med medianen för det kluster som observationerna placerats i. Vi skattar med hjälp av medelsilhuett-metoden även här 2 kluster som det optimala. Vi utför sedan vår k-meansklustring och beräknar medelvärdena för respektive kluster och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel: 139.2185

Slutgiltiga kluster för CDI-datamaterialet



Figur 5.2. Slutgiltiga kluster för träningsdata och testdata i CDI-datamaterialet

I figur 5.2 presenteras de slutgiltiga klustren för träningsdatamaterialet, samt kluster för hur väl imputeringsteknikerna i de olika test-dataseten lyckades replikera träningsdatasetets klustring.

Jämförelse av medelkvadratfelet

	Medelkvadratfel
Imputering med MICE	*56.09514*
Imputering med stickprovsmedelvärdet	*139.1653*
Imputering med stickprovsmedianen	*139.2185*

Tabell 5.2. Medelkvadratfel skattade utifrån träningsdatamaterialet för varje imputeringsteknik i datamaterialet CDI

Som går att utläsa från tabell 5.2 är MICE den imputeringsteknik som genererar lägst medelkvadratfel. Utifrån detta resultat imputerar vi testdatamaterialet för MICE en gång till och jämför det med originaldatamaterialet vilket genererar ett medelkvadratfel på 6.735366.

5.3. Experiment 3: Grodor

Beskrivning av dataset och uppdelning i träningsdata och testdata

Detta datamaterial är en undersökning av grodor som delats upp efter 4 familjer, 8 släkten och 10 arter. Materialet består av 7195 observationer i 26 variabler. Då 4 av dessa variabler ej är numeriska valde vi att bortse från dessa. Träningsdatamaterialet består efter uppdelningen av 5037 observationer med 22 variabler och test-datasetet av 2158 observationer i 22 variabler. Vi skalar sedan värdena för att göra dem lämpliga för klusteranalys.

Klustring av träningsdatan

Utifrån medelsilhuett-metoden får vi fram att det optimala antalet kluster är *k=2.* Medelvärden för respektive kluster beräknas och finns tillgängliga i appendix, tabell A.3.1. Det är medelvärdena från detta träningsdataset som kommer att användas som utgångspunkt när vi jämför medelkvadratfelen från imputationsteknikerna.

Imputering: MICE

Vi använder oss nu av test-datasetet och genererar tio procents bortfall som vi sedan imputerar med hjälp av MICE. Sedan upprepar vi processen för test-datasetet och får ut *k=2* som optimalt antal kluster. Vi beräknar sedan medelvärdena för respektive kluster och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel: 0.01825954

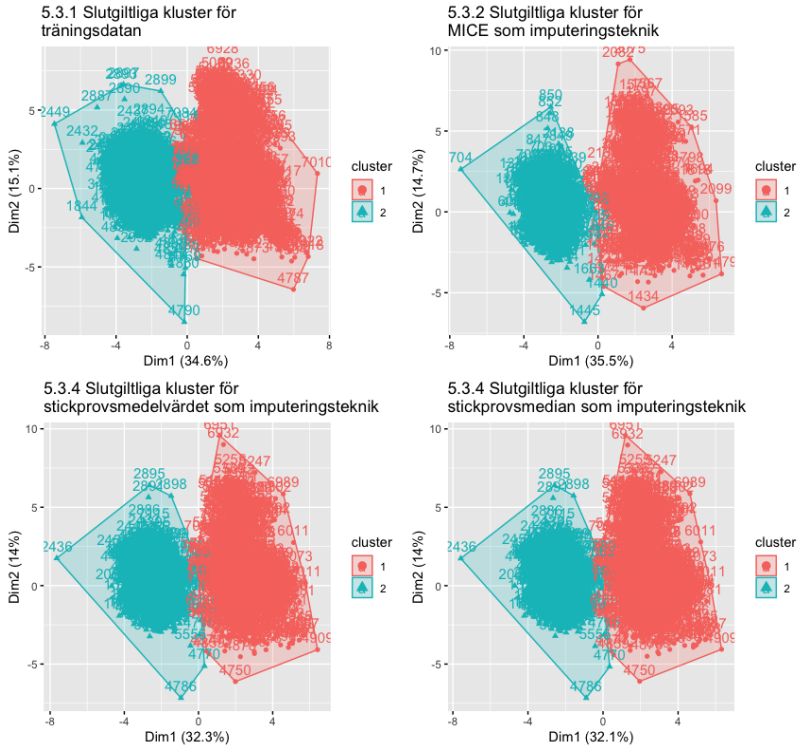
Imputering: stickprovsmedelvärdet

Vi använder oss nu av test-datasetet och genererar tio procents bortfall som vi sedan imputerar med hjälp av stickprovsmedelvärdet. Sedan upprepar vi processen för test-datasetet och får ut *k=2* som optimalt antal kluster. Vi beräknar sedan medelvärdena för respektive kluster och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel: 0.03400661

Imputering: stickprovsmedian

Vi använder oss nu av test-datasetet och genererar tio procents bortfall som vi sedan imputerar med hjälp av stickprovsmedianen. Sedan upprepar vi processen för test-datasetet och får ut *k=2* som optimalt antal kluster. Vi beräknar sedan medelvärdena för respektive kluster och skattar medelkvadratfelet utifrån träningsdatans medelvärden vilket ger följande medelkvadratfel:0.03529599

Slutgiltiga kluster för Grod-datamaterialet



Figur 5.3. Slutgiltiga kluster för träningsdata och testdata i Grod-datamaterialet

I figur 5.3 presenteras de slutgiltiga klustren för träningsdatamaterialet, samt kluster för hur väl imputeringsteknikerna i de olika test-dataseten lyckades replikera träningsdatasetets klustring.

Jämförelse av medelkvadratfelet

	Medelkvadratfel
Imputering med MICE	*0.01825954*
Imputering med stickprovsmedelvärdet	*0.03400661*
Imputering med stickprovsmedianen	*0.03529599*

Tabell 5.3. Medelkvadratfel skattade utifrån träningsdatamaterialet för varje imputeringsteknik i datamaterialet Grodor

Som går att utläsa från tabell 5.3 är MICE den imputeringsteknik som genererar lägst medelkvadratfel. Utifrån detta resultat imputerar vi testdatamaterialet för MICE en gång till och jämför det med originaldatamaterialet vilket genererar ett medelkvadratfel på 0.07096026.

6. Diskussion och slutsats

Våra slutsatser går att dela upp i en som behandlar imputationsteknikernas skattning av antal kluster och en som ämnar hitta den imputationsteknik som ger bäst utfall för att skapa de kluster som hade skapats i ett datamaterial utan bortfall.

Syftet med denna studie har varit att undersöka vilken metod för imputation som är bäst lämpad för ett datamaterial med bortfall och det går att dra flera slutsatser utifrån vårt resultat. Den frågeställning vi har ämnat undersöka har varit huruvida valet av imputationsteknik påverkar hur många kluster som är optimalt. Vi kan utröna att imputation med medelvärde, median samt MICE gav samma resultat i valet av antal kluster. Samtliga metoder replikerade det antal som skattades i träningsdatasetet. En slutsats utifrån detta resultat är att valet av imputationsteknik inte påverkar algoritmen för optimalt antal kluster. Detta påstående går till viss del emot tidigare forskning, då forskare som såväl Manly som Fraley beskriver hur valet av antal kluster är en av de allra största utmaningarna för klusteranalysen, där definitionen av kluster är under ständig diskussion. För att nyansera denna slutsats har vi valt att använda oss av medelsilhuett-metoden för att skatta antalet kluster och det är den metod som vi har baserat våra resultat på. Detta är en potentiell felkälla i vår studie och vi har försökt utveckla vår skattning av antal kluster med hjälp av andra metoder, såsom “gap statistics”-algoritmen. Metoder för att skatta antal kluster är ett stort område och för att avgränsa oss har vi valt att stanna vid denna jämförelse och sedan undersöka skillnader i medelvärden. Då det inte finns ett konventionellt accepterat tillvägagångssätt att hitta rätt antal kluster har vi utgått ifrån en underbyggd metod för att stärka våra resultat. Vår slutsats att imputationsteknikerna inte påverkar valet av antal kluster skulle behöva utsättas för vidare forskning, men kan vara en intressant utgångspunkt i framtida studier för hantering av bortfall och val av antal kluster.

Utifrån våra resultat är en möjlig slutsats att MICE är den imputationsteknik som genererar lägst medelkvadratfel i förhållande till originaldatamaterialet. Då syftet med denna studie är att undersöka hur bortfall bör hanteras inför en klusteranalys kan denna slutsats vara en indikation

på att MICE är den bäst lämpade metoden för att återskapa de kluster som ett datamaterial utan bortfall hade skapat. Då MICE är en form av regressionsskattning för imputering är det viktigt att ha i åtanke de grundläggande antaganden man gör för regressionsanalys, där såväl normalfördelning som residualanalysen är viktiga verktyg för att avgöra huruvida en MICE-skattning ens är en rimlig metod att använda sig av vid olika datamaterial. Vissa av våra test-dataset är mycket små och antaganden om normalfördelning, såväl i datamaterialet som i residualerna, riskerar därmed att brista. Denna problematik har vi försökt åtgärda genom att undersöka olika datamaterial med större och mindre antal observationer för att se om detta påverkade våra resultat. Vid mindre datamaterial där alla antaganden för regressionsanalys inte är uppfyllda kan det vara bättre lämpat att använda sig av medelvärde eller median som imputationsteknik, även om de i snitt utifrån vårt resultat genererar högre medelkvadratfel. Dessa tekniker är mindre komplexa och som diskuterades innan raderar de all potentiell variation i bortfallet. Därmed kan man testa med både stickprovsmedelvärdet och stickprovsmedianen, och få en bild av spridningen i materialet och var variabelbortfallet borde ha hamnat. Summerat har vi kommit fram till att MICE har en mer precis skattning utifrån träningsdatamaterialet, men i fall där materialet är litet eller ej uppfyller antaganden kan det vara mer lämpligt att använda sig av stickprovsmedelvärdet eller stickprovsmedianen som imputeringsteknik.

Vi ser denna studie som en ansats inför kommande undersökningar av såväl imputationsteknik vid bortfall som skattning av optimala kluster. Om en undersökning gjorts och bortfall uppstått kan klusteranalysen fortfarande vara ett relevant verktyg om det hanteras rätt. Det gäller att väga fördelarna med ett lägre medelkvadratfel som MICE-skattningen ger upphov till mot att anpassa sin imputation för det datamaterial man har tillgängligt. Uppfyller vårt datamaterial de antaganden som är grundläggande för regressionsskattningar? Är en mindre komplex modell som stickprovsmedelvärdet som imputation mer tillämpbar på bekostnad av större medelfel och förlorad variation? Dessa är frågeställningar som vår studie väcker och vi uppmuntrar till framtida forskning inom ämnet.

7. Källförteckning

Dahmström, K. (2011) *Från datainsamling till rapport: att göra en statistisk undersökning.* (5. uppl.). Studentlitteratur AB, Lund

Fraley, C. & Raftery A. E. (1998) How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, *The Computer Journal*, vol 41(8), s.578-588

Gut, A. (2009) *An Intermediate Course in Probability Theory,* 2:a upplagan, Springer, Berlin

Kriegel, H. & Schubert, E. & Zimek, A. (2017) The (black) art of runtime evaluation: are we comparing algorithms or implementation? *Knowledge and information Systems,* vol. 52 s. 341-378

Körner, S. & Wahlgren, L. (2015) *Statistisk dataanalys,* 5:e upplagan, Studentlitteratur AB, Lund

Liu, H. & Cocea, M.(2017) Semi-random partitioning of data into training and test sets in granular computing context. *Granul. Comput*. vol 2, s. 357–386.

Lohr, S. (2010) *Sampling, design and analysis,* 2:a upplagan, Brooks/Cole, Cengage Learning

Manly, B.F.J. & Navarro Alberto, J. A. (2017) *Multivariate Statistical Methods, A primer*, 4:e upplagan, CRC Press

Rousseeuw, P. (1987) Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, vol. 20, s. 53–65

Tibshirani, R. & Walther, G. & Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic*, Journal of the Royal Statistical Society,* vol 63, s. 411-423

van Buuren S. & Groothuis-Oudshoorn K. (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, vol. 45(3), s. 1-67 https://www.jstatsoft.org/v45/i03/ (hämtad 2021-01-09)

Appendix

A.1 VIN-datamaterialet

A.1.1 **Medelvärden för träningsdatamaterialets kluster**

	Kluster 1	*Kluster 2*	*Kluster 3*
Alcohol	*0.81760082*	*-0.92668390*	*0.10229803*
Malic_Acid	*-0.28679140*	*-0.38283030*	*0.95957998*
Ash	*0.34225898*	*-0.47326051*	*0.16282971*
Ash_Alcanity	*-0.51862180*	*0.13513384*	*0.57169380*
Magnesium	*0.55679479*	*-0.54264792*	*-0.05469371*
Total_Phenols	*0.83357141*	*-0.17795879*	*-0.97405345*
Flavanoids	*0.92408125*	*-0.08867420*	*-1.23254624*
Nonflavanoid_Phenols	*-0.53447256*	*-0.01842279*	*0.81091363*
Proanthocyains	*0.56049487*	*-0.01691127*	*-0.79944536*
Color_intensity	*0.23957427*	*-0.91940603*	*0.94104001*
Hue	*0.33804311*	*0.51325010*	*-1.21825877*
OD280	*0.73568177*	*0.20284399*	*-1.36578197*
Proline	*1.04801554*	*-0.73380885*	*-0.50735414*

Tabell A.1.1. Vin-träningskluster med medelvärden för samtliga kluster

A.1.2.Medelvärden för imputering med MICE

	Kluster 1	*Kluster 2*	*Kluster 3*
Alcohol	*-0.90430065*	*0.69573586*	*0.31841782*
Malic_Acid	*-0.39947447*	*-0.38270884*	*0.72175560*
Ash	*-0.43709602*	*0.25626439*	*0.22129442*
Ash_Alcanity	*0.33879604*	*-0.92779552*	*0.44250544*
Magnesium	*-0.58418538*	*0.81451524*	*-0.10172219*
Total_Phenols	*0.29009092*	*0.86331451*	*-1.01709262*
Flavanoids	*0.30038742*	*1.04042490*	*-1.17653471*
Nonflavanoid_Phenols	*-0.04456878*	*-0.60574975*	*0.55467383*
Proanthocyains	*0.22305668*	*0.62557537*	*-0.74985699*
Color_intensity	*-0.87543135*	*-0.04880104*	*0.91652696*
Hue	*0.25640007*	*0.86619911*	*-0.98583090*
OD280	*0.49239771*	*0.77426296*	*-1.14440862*
Proline	*-0.89422340*	*1.18715907*	*-0.10548950*

Tabell A.1.2. Vin-testkluster med Imputering med MICE

A.1.3.Medelvärden för imputering med stickprovsmedelvärdet

Alcohol	*Kluster 1 -0.885254483*	*Kluster 2 0.716398429*	*Kluster 3 0.281971595*
Malic_Acid	*-0.363027206*	*-0.323636248*	*0.635562994*
Ash	*-0.415934341*	*0.251122802*	*0.204462507*
Ash_Alcanity	*0.266749868*	*-0.871943205*	*0.467518094*
Magnesium	*-0.513002455*	*0.771171682*	*-0.136405277*
Total_Phenols	*0.190642226*	*0.893462483*	*-0.943031685*
Flavanoids	*0.252976052*	*1.014196258*	*-1.107036059*
Nonflavanoid_Phenols	*-0.162491795*	*-0.538890033*	*0.616293928*
Proanthocyains	*0.208182073*	*0.584287097*	*-0.700213313*
Color_intensity	*-0.873270648*	*-0.006937212*	*0.879112511*
Hue	*0.300937447*	*0.837065916*	*-1.005835060*
OD280	*0.428135846*	*0.777579288*	*-1.082939456*
Proline	*-0.916360021*	*1.188310944*	*-0.084322879*

Tabell A.1.3. Vin-testkluster med Imputering med stickprovsmedelvärdet.

A.1.4.Medelvärden för imputering med stickprovsmedianen

	Kluster 1	*Kluster 2*	*Kluster 3*
Alcohol	*-0.9224275*	*0.7347931*	*0.3036544*
Malic_Acid	*-0.3836107*	*-0.3665197*	*0.6922589*
Ash	*-0.5038580*	*0.2490664*	*0.2941178*
Ash_Alcanity	*0.3416840*	*-0.9461353*	*0.4550614*
Magnesium	*-0.54636775*	*0.72509282*	*-0.06423673*
Total_Phenols	*0.2365344*	*0.8868953*	*-0.9833935*
Flavanoids	*0.3140581*	*1.0112449*	*-1.1656328*
Nonflavanoid_Phenols	*-0.0780522*	*-0.6379497*	*0.6152730*
Proanthocyains	*0.1377186*	*0.6865482*	*-0.7158645*
Color_intensity	*-0.8647629*	*-0.0528612*	*0.9092776*
Hue	*0.3591921*	*0.8169736*	*-1.0471698*
OD280	*0.4654430*	*0.7896899*	*-1.1304450*
Proline	*-0.9014136*	*1.2551138*	*-0.1555244*

Tabell A.1.4. Vin-testkluster med Imputering med stickprovsmedianen.

A.2. CDI-datamaterialet

A.2.1 Medelvärden för träningsdatamaterialets kluster

	Kluster 1	*Kluster 2*
Land area	*-0.02016865*	*0.7563245*
Population	*-0.1202572*	*4.5096439*
Young	*-0.008692868*	*0.325982555*
Old	*0.007583419*	*-0.284378221*
Physicians	*-0.1069641*	*4.0111524*
Hospital beds	*-0.113844*	*4.269149*
Crimes	*-0.1280552*	*4.8020699*
High school graduates	*0.01794397*	*-0.67289902*
Bachelor degrees	*-0.002841598*	*0.106559941*
Poverty	*-0.01972314*	*0.73961783*
Unemployment rate	*-0.009418665*	*0.353199927*
Income per capita	*-0.00934656*	*0.35049599*
Tot pers income	*-0.1149198*	*4.3094906*
Region	*-0.01146641*	*0.42999024*

Tabell A.2.1. CDI-träningskluster med medelvärden för samtliga variabler

A.2.2.Medelvärden för imputering med MICE

	Kluster 1	*Kluster 2*
Land area	*0.82454952*	*-0.08245495*
Population	*2.4883667*	*-0.2488367*
Young	*-0.12365929*	*0.01236593*
Old	*0.26118545*	*-0.02611854*
Physicians	*2.548630*	*-0.254863*
Hospital beds	*2.2171779*	*-0.2217178*
Crimes	*2.1194815*	*-0.2119482*
High school graduates	*0.31165547*	*-0.03116555*
Bachelor degrees	*0.74761653*	*-0.07476165*
Poverty	*-0.29007355*	*0.02900735*
Unemployment rate	*-0.12671337*	*0.01267134*
Income per capita	*1.116170*	*-0.111617*
Tot pers income	*2.5842285*	*-0.2584229*
Region	*-0.26654379*	*0.02665438*

Tabell A.2.2. CDI-testkluster med Imputering med MICE.

A.2.3.Medelvärden för imputering med stickprovsmedelvärdet

	Kluster 1	*Kluster 2*
Land area	*0.77572239*	*-0.05676018*
Population	*2.9072088*	*-0.2127226*
Young	*0.109116006*	*-0.007984098*
Old	*-0.04272460*	*0.00312619*
Physicians	*2.9070559*	*-0.2127114*
Hospital beds	*2.6331342*	*-0.1926684*
Crimes	*2.4388765*	*-0.1784544*
High school graduates	*0.24507622*	*-0.01793241*
Bachelor degrees	*0.59865442*	*-0.04380398*
Poverty	*-0.094128130*	*0.006887424*
Unemployment rate	*-0.29680265*	*0.02171727*
Income per capita	*1.08792630*	*-0.07960436*
Tot pers income	*2.6316242*	*-0.1925579*
Region	*-0.082294663*	*0.006021561*

Tabell A.2.3. CDI-testkluster med Imputering med Stickprovsmedelvärdet.

A.2.4.Medelvärden för imputering med stickprovsmedianen

	Kluster 1	*Kluster 2*
Land area	*0.65898595*	*0.05401524*
Population	*2.8379496*	*-0.2326188*
Young	*0.045245080*	*-0.003708613*
Old	*-0.0105136038*	*0.0008617708*
Physicians	*2.6424413*	*-0.2165936*
Hospital beds	*2.3474492*	*-0.1924139*
Crimes	*2.2982877*	*-0.1883842*
High school graduates	*0.23323888*	*-0.01911794*
Bachelor degrees	*0.63032739*	*-0.05166618*
Poverty	*-0.20180373*	*0.01654129*
Unemployment rate	*-0.33558995*	*0.02750737*
Income per capita	*1.12163040*	*-0.09193692*
Tot pers income	*2.5851705*	*-0.2118992*
Region	*-0.16866895*	*0.01382532*

Tabell A.2.4. CDI-testkluster med Imputering med Stickprovsmedianen.

A.3. Grodmaterialet

A.3.1 Medelvärden för träningsdatamaterialets kluster

MFCCs_.1	*Kluster 1* *0.1187631*	*Kluster 2* *-0.1152318*
MFCCs_.2	*-0.2087043*	*0.2024987*
MFCCs_.3	*-0.3676248*	*0.3566938*
MFCCs_.4	*0.6752456*	*-0.6551678*
MFCCs_.5	*0.3836148*	*-0.3722084*
MFCCs_.6	*-0.5777272*	*0.5605491*
MFCCs_.7	*-0.6678728*	*0.6480143*
MFCCs_.8	*0.3990785*	*-0.3872123*
MFCCs_.9	*0.7410353*	*-0.7190014*
MFCCs_.10	*-0.1808508*	*0.1754734*
MFCCs_.11	*-0.7995704*	*0.7757960*
MFCCs_.12	*0.3520187*	*-0.3415518*
MFCCs_.13	*0.8006219*	*-0.7768162*
MFCCs_.14	*-0.4944072*	*0.4797065*
MFCCs_.15	*-0.7926684*	*0.7690993*
MFCCs_.16	*0.4230902*	*-0.4105100*
MFCCs_.17	*0.7937662*	*-0.7701644*
MFCCs_.18	*0.01965107*	*-0.01906677*
MFCCs_.19	*-0.6227080*	*0.6041924*
MFCCs_.20	*-0.7013518*	*0.6804979*
MFCCs_.21	*0.2897786*	*-0.2811623*
MFCCs_.22	*0.8244989*	*-0.7999833*

Tabell A.3.1. Grodor-träningskluster med medelvärden för samtliga variabler

A.3.2.Medelvärden för imputering med MICE

MFCCs_.1	*Kluster 1* *0.1038582*	*Kluster 2* *-0.1060949*
MFCCs_.2	*-0.1849566*	*0.1889398*
MFCCs_.3	*-0.3559808*	*0.3636471*
MFCCs_.4	*0.6775413*	*-0.6921325*
MFCCs_.5	*0.3981540*	*-0.4067285*
MFCCs_.6	*-0.5743351*	*0.5867038*
MFCCs_.7	*-0.6576133*	*0.6717754*
MFCCs_.8	*0.3984359*	*-0.4070165*
MFCCs_.9	*0.7296193*	*-0.7453321*
MFCCs_.10	*-0.1793356*	*0.1831977*
MFCCs_.11	*-0.7814034*	*0.7982313*
MFCCs_.12	*0.3676782*	*-0.3755963*
MFCCs_.13	*0.7861668*	*-0.8030974*
MFCCs_.14	*-0.5120868*	*0.5231149*
MFCCs_.15	*-0.7730157*	*0.7896630*
MFCCs_.16	*0.4388117*	*-0.4482618*
MFCCs_.17	*0.7754729*	*-0.7921732*
MFCCs_.18	*0.02048080*	*-0.02092186*
MFCCs_.19	*-0.5956171*	*0.6084441*
MFCCs_.20	*-0.6837336*	*0.6984582*
MFCCs_.21	*0.2754047*	*-0.2813358*
MFCCs_.22	*0.8189956*	*-0.8366332*

Tabell A.3.2. Grodor-testkluster med Imputering med MICE.

A.3.3.Medelvärden för imputering med stickprovsmedelvärdet

MFCCs_.1	*Kluster 1* *0.1048952*	*Kluster 2* *-0.1067578*
MFCCs_.2	*-0.1823045*	*0.1855416*
MFCCs_.3	*-0.3406152*	*0.3466635*
MFCCs_.4	*0.6410316*	*-0.6524145*
MFCCs_.5	*0.3721240*	*-0.3787318*
MFCCs_.6	*-0.5384635*	*0.5480250*
MFCCs_.7	*-0.6283011*	*0.6394578*
MFCCs_.8	*0.3859828*	*-0.3928367*
MFCCs_.9	*0.6935858*	*-0.7059018*
MFCCs_.10	*-0.1643129*	*0.1672306*
MFCCs_.11	*-0.7418160*	*0.7549885*
MFCCs_.12	*0.3519373*	*-0.3581867*
MFCCs_.13	*0.7465114*	*-0.7597672*
MFCCs_.14	*-0.4927044*	*0.5014534*
MFCCs_.15	*-0.7377879*	*0.7508888*
MFCCs_.16	*0.4156811*	*-0.4230623*
MFCCs_.17	*0.7347974*	*-0.7478452*
MFCCs_.18	*0.02354462*	*-0.02396270*
MFCCs_.19	*-0.5674388*	*0.5775148*
MFCCs_.20	*-0.6513006*	*0.6628658*
MFCCs_.21	*0.2713646*	*-0.2761833*
MFCCs_.22	*0.7800437*	*-0.7938950*

Tabell A.3.3. Grodor-testkluster med Imputering med stickprovsmedelvärdet.

A.3.4.Medelvärden för imputering med stickprovsmedianen

MFCCs_.1	*Kluster 1* *0.1052249*	*Kluster 2* *-0.1066973*
MFCCs_.2	*-0.1816989*	*0.1842413*
MFCCs_.3	*-0.3448148*	*0.3496396*
MFCCs_.4	*0.6396141*	*-0.6485639*
MFCCs_.5	*0.3721865*	*-0.3773943*
MFCCs_.6	*-0.5358682*	*0.5433664*
MFCCs_.7	*-0.6248324*	*0.6335754*
MFCCs_.8	*0.3825830*	*-0.3879363*
MFCCs_.9	*0.6849608*	*-0.6945451*
MFCCs_.10	*-0.1644715*	*0.1667729*
MFCCs_.11	*-0.7428672*	*0.7532618*
MFCCs_.12	*0.3501682*	*-0.3550680*
MFCCs_.13	*0.7498408*	*-0.7603330*
MFCCs_.14	*-0.4901546*	*0.4970131*
MFCCs_.15	*-0.7383443*	*0.7486757*
MFCCs_.16	*0.4149588*	*-0.4207651*
MFCCs_.17	*0.7369693*	*-0.7472814*
MFCCs_.18	*0.02560214*	*-0.02596038*
MFCCs_.19	*-0.5683568*	*0.5763096*
MFCCs_.20	*-0.6525475*	*0.6616783*
MFCCs_.21	*0.2714676*	*-0.2752661*
MFCCs_.22	*0.7789064*	*-0.7898052*

Tabell A.3.4. Grodor-testkluster med Imputering med stickprovsmedianen